Old age mortality in South Africa

Takwanisa Machemedze

A dissertation submitted to the faculty of Commerce of the University of Cape Town in partial fulfillment of the requirements for the Degree of Master of Philosophy in Demography

Centre for Actuarial Research

July 09

Plagiarism declaration form

This research is my original work, produced with normal supervisory assistance from my supervisor. All the relevant sources of knowledge that I have used during the course of writing this dissertation have been fully credited using the Harvard convention for citation and referencing. Also, this dissertation has not been submitted for any academic or examination purpose at any other university.

---------- ----------

Takwanisa Machemedze Date

Abstract

This study estimates the mortality of the South African oldest old age population (in five year age groups from age 75 up to the open age interval 100 and above) and in the process re-estimates the numbers of people in the population at these ages at the time of the 1996 and 2001 censuses, and the 2007 Community Survey.

In countries where the data on the old age population have been verified, it has been observed that the data are marred by errors in the form of age exaggeration, age digit preference, relative under/over count of the population and under-registration of deaths. These errors have been observed to have the net effect of underestimating mortality of the oldest old age groups.

The current research applies the method of extinct generations to estimate indirectly the population numbers at the oldest old age groups (75 up to 100 and above) using data on reported deaths alone.

Age heaping and year of birth preference in the reported deaths are assessed using ratios of the probability of death estimated from the data. Age exaggeration in the data on reported deaths is assessed using ratios of deaths compared with same ratios from a standard population. Age heaping and year of birth preference in the census/survey population is assessed using the modified Whipple's Index of age accuracy.

The Generalized Growth Balance (GGB) and Synthetic Extinct Generations (SEG+delta) methods are applied to adjust for under reporting of deaths and to assess patterns of age exaggeration in the census/survey population. The difference between the estimates of the completeness of reporting of deaths from the two methods is small (less than 1 per cent) and has been observed to have little impact on the mortality estimates. Final estimates of the completeness of reporting of deaths used are those derived using the SEG+delta method.

After re-estimating the population numbers and adjusting for completeness of reporting of deaths, mortality rates were then estimated. Results obtained from the method of extinct generations suggest that there is no systematic difference between the census/survey population and the population numbers estimated from deaths except at ages 95 and above.

Measures of age accuracy show that there are patterns of preferring 1910, 1914, 1918, 1920 and 1930 as the years of birth in the census/survey population and these patterns are also found in the registered deaths. The impact of these errors was investigated and the results show that preference of certain years of birth cause fluctuations in the mortality rates.

Patterns observed after applying the SEG+delta method suggest that the completeness of reporting of deaths falls with age at the advanced ages (from age 90 and above) and as a result, the estimated mortality rates above this age are lower than those estimated from the United Nations Population Division (UNPD) and US Census Bureau (USCB) population projections, and Dorrington, Moultrie and Timaeus (2004).

Conclusions reached are that the mortality rates for the age groups 75 to 89 derived after re-estimating the population numbers and after allowing for the fall in the completeness of reporting of deaths are lower but not significantly different from those inferred from the UNPD and USCB population projections, and estimates derived by Dorrington, Moultrie and Timaeus (2004). The research recommends mortality estimates from the UNPD since they are the closest to the estimates derived using the published census population numbers for the whole period between the nights of 9-10 October 1996 and 9-10 October 2001. However, the research produced better estimates of the oldest old age population numbers relative to the census/survey numbers.

Acknowledgements

I would like to acknowledge the tireless efforts of my supervisor: Professor R. Dorrington for his contribution to this study. I would like to thank him for his understanding and patience shown towards me.

I also would like to thank the department (CARe) for a detailed programme that I found challenging, stimulating and enriching to the profession of demography.

I thank family members and friends for their encouragement and support.

Last but not least, I thank the Andrew W Mellon and Hewlett Foundation, and the University of Cape Town, Postgraduate Funding Office for their financial assistance which facilitated my studies and stay in Cape Town.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

It has been projected that the world population aged 80 and above is going to grow faster than the other age groups and as a result, the absolute size and proportion of the population in this age group will increase (National Research Council, 2006; United Nations Department of Economic and Social Affairs, 2007). This shift in population structure has drawn attention from various policy makers because of its long term implications. This current research seeks to estimate the South African oldest old age mortality which is important for understanding the impacts of this aging on, for example, the burden of disease, provision for retirement, and relative size of the older population among others.

Mortality estimates for the oldest age groups are important for improving estimates of population size at these ages. Researchers who are interested in the study of mortality have faced challenges in estimating oldest old age mortality in developing countries. Among the challenges they face are availability of reliable data of the population count and statistics on reported deaths. The data that are available have errors in the form of: age exaggeration, age heaping, relative under count/over count of people and relative completeness of reporting of deaths by age.

The population size at old age is small and any misstatement of age (usually age exaggeration and age digit preference) causes a significant difference in the estimates of mortality rates from their true values. Research by Preston, Elo and Stewart (1999) investigating the effect of age misreporting on the mortality rates up to age 100 observed that age exaggeration leads to underestimating old age mortality. Bennett and Horiuchi (1981) highlighted that reported deaths that are required to estimate mortality rates have been observed to be under reported, especially in developing countries. They derived a method for adjusting for the under registration of deaths which they observed to fail for old age groups because of age exaggeration.

The method of extinct generations first proposed by Vincent (1951) has been used to produce more accurate estimates of mortality at old age. The method uses information on reported deaths by age to re-estimate the population at some point in time in the past by single ages. The recording of age at death on the death certificate is assumed to be more accurate than the recording of age in a census enumeration. As a result, the population

numbers at the old ages re-estimated from the information on reported deaths have been observed to be more accurate than the numbers from a census enumeration.

The current study estimates the South African oldest old age mortality rates and in the process re-estimates the population numbers at these ages using data on reported deaths alone. The reported deaths to be used are going to be thoroughly investigated for possible bias in the form of age heaping, age exaggeration and preference of certain years of birth. A comparison of the re-estimated population numbers to the census/survey population numbers is done to explain any observable discrepancies between the numbers. Mortality estimates are determined from the re-estimated population numbers and the reported deaths, and then compared with the mortality estimates derived from other sources.

1.2 Aims and Objectives of the research

The research seeks to produce estimates of mortality of the South African oldest old age population (in five year age groups from age 75 up to the open age interval 100 and above) and in the process to re-estimate the numbers of people in the population at these ages at the time of the 1996 and 2001 censuses, and the 2007 Community Survey. The method of extinct generations is usually used to estimate the population aged 80 and above and this research investigates extending this method to cater for people from age 75. The research also interrogates and investigates the common data errors that are known to be associated with the demographic data of the old age population and among them are age exaggeration, age heaping and birth year preference.

1.3 Statement of the problem

This current research seeks to investigate the accuracy of the South African oldest old age mortality estimates and as part of that process needs to consider the accuracy of the population estimates. It has been observed that estimates of the South African population at old age (75 up to say 100 and above) from different bodies do not match the census/survey population numbers. The sources of estimates that have been used to compare with the census/survey population numbers are the online databases from the US Census Bureau (2005), United Nations (2007) and estimates from the demographic model designed by the Actuarial Society of South Africa (2005). This difference in population numbers would suggest that the population size at old age is not well known and hence the mortality at these ages is also not known. Associated with this difference in numbers at old

age are errors in the form of age misreporting (usually age exaggeration and age digit preference), population under/over count and preference of certain years of birth. The extent of these errors needs to be investigated as well.

1.4 Significance of the research

This research is relevant to the ongoing debate on the global burden of aging and population aging. It has been projected that the oldest old age world population (80 years and above) will grow faster than the other age groups in the next two decades (National Research Council, 2006; United Nations Department of Economic and Social Affairs, 2007). While this increase in human life span has been celebrated as an achievement of the last millennium, it is also a burden to families, governments and private sectors.

Aging refers to the increase in the absolute numbers of old people and population aging refers to the increase in the proportion of the old age population (National Research Council, 2006; United Nations Department of Economic and Social Affairs, 2007). Both changes in the population structure are a result of reductions in fertility and to some extent the mortality at older ages. A reduction in fertility reduces the proportion of young people, whereas a reduction in mortality increases the proportion of people who will survive to old age (Preston, Heuveline and Guillot, 2001; United Nations Department of Economic and Social Affairs, 2007). It may be useful for research and policy purposes to distinguish between old age population (aged 60 and above) and the oldest old age population (aged 80 and above) (National Research Council, 2006; United Nations Department of Economic and Social Affairs, 2007).

The challenges that arise from aging include health and living conditions of older people, financial support, challenges to existing models of social support, strains on social insurance and pension systems, and changes in disease patterns and prevalence (National Research Council, 2006; United Nations Department of Economic and Social Affairs, 2007).

The demographics of aging are not well known in most developing countries. As a result, the resources available to address health and demographic changes have focused on issues that are assumed to be of immediate concern to the young population (infant, child and maternal health, nutrition and HIV/AIDS) at the expense of the needs of the aged population (National Research Council, 2006).

Research into aging is important in countries where old age programs account for a considerable proportion of public expenditure budget. South Africa is one of the few countries in the Sub-Saharan region that has a significant security programme for the aged (Makino, 2004; National Research Council, 2006).

The ravages of the HIV/AIDS epidemic have affected the wage-earning population and these are the people who are expected to provide financial support for their aged relatives and children (National Research Council, 2006). To make matters worse, the impact of HIV/AIDS epidemic leaves the aged population with the burden of looking after the orphans as well as themselves. Looking at ages, say above 65, these people will have reached retirement age and the income they might have at that age may not be enough to cater for their responsibilities. This calls for governments and other organizations to assist with resources (United Nations Department of Economic and Social Affairs, 2007). More accurate estimates of the population numbers at the oldest old age groups in South Africa assist policy makers and planners on the allocation of resources to these people.

The changes in population structure means that there is going to be a shift from infectious diseases that are common in children to degenerative diseases that are common in adults and the elderly (Omran, 2005). Knowing the aging patterns and mortality rates helps to prepare for the resources to counter the health effects of aging.

The insurance industry is another beneficiary of such research. It would help them in pricing annuities and life insurance premiums in relation to the pattern of old age mortality (Himes, Preston and Condran, 1994).

In order to plan for the implications of aging highlighted above, there is the need to know the numbers of the old age population and their dynamics (mortality patterns). Sources of data of population numbers in general and the old age population numbers in particular are either censuses which may not be reliable and to some extent are not detailed, or small scale surveys, which may not be generalized to the whole population. This research uses indirect methods to re-estimate the oldest old age population numbers and the mortality rates at these ages after first having re-estimated the population numbers to remove various age errors. The results produced in this research can only be derived because South Africa has a relatively complete vital registration system.

1.5 Chapter outline

This research is organized in five chapters. The next chapter (2) gives a thorough review of the literature behind old age mortality. This will be followed by chapter (3) on the methods used to answer the research aims and objectives. The following chapter (4) presents the results and analysis. Chapters 5 will then discuss the results, summarize the conclusions reached and provide suggestions for further research.

2 Literature Review

2.1 Introduction

This chapter reviews previous work on old age mortality in the context of what has been done to improve the estimates and problems encountered in working with data at the advanced ages. This gives a picture of how to improve the methods used in previous research and areas that needs careful consideration. The method of extinct generations of re-estimating the old age population using data on reported deaths alone is reviewed next. The quality of the data on deaths used to estimate the population also needs to be assessed and the literature on the assessment methods is discussed next. Following this is a review of the death distribution methods. The review of these methods describes how data from two successive censuses and the data on reported deaths over the period between the censuses can be used to estimate completeness of the reporting of deaths and relative coverage of the censuses.

2.2 Old age mortality estimates

Mortality rates at old age are generally difficult to estimate. The problems arise as a result of irregular patterns in the data on reported deaths and the population at old age. The most common data problems that have been documented are age overstatement, age digit preference and under count of the population or under reporting of deaths (Wilmoth, 1995; Jdanov, Jasilionis *et al.*, 2008). Other research observes that these problems in the population data at old age generally biases mortality estimates downwards (Horiuchi and Coale, 1982; Preston, Elo and Stewart, 1999). This is because there is a tendency by individuals to overstate their ages and as a result increasing the number of people who survive to old age. Due to these implications, the mortality estimates are usually aggregated at high enough ages where the errors are assumed to be minimal and the errors in the aggregated ages nearly cancel out to give better estimates of mortality (Horiuchi and Coale, 1982).

Taking South Africa in particular, and according to Dorrington, Moultrie and Timaeus (2004:iii), the national estimates of mortality in general, "… have always been fairly approximate". Part of the explanation is that reporting of deaths was very incomplete until after 1994 (Anderson and Phillips, 2006). According to Anderson and Phillips (2006),

sufficiently complete data on registered deaths could be obtained as from 1997. They used registered deaths and compare with those from Spectrum (United Nations, 2005) with default (UNPD) assumptions to estimate the completeness of reporting of deaths. They observed that the completeness of registration of deaths for the population aged 65 and above over the period 1997 to 2004 was implausibly high. This finding is the reason why Anderson and Phillips (2006) estimated mortality rates up to age 64 only. Other efforts have been made to estimate mortality at high ages. Research that has estimated mortality rates for the period under consideration include efforts by Dorrington, Moultrie and Timaeus (2004) who used the 1996 and 2001 censuses to estimate the South African mortality over the period between the two censuses and their life tables are aggregated at age 90. Statistics South Africa (2008b) published death rates for the year 2006 and the rates are aggregated at age 80. Death rates extending beyond age 90 can be inferred from the USCB and UNPD population projections.

The following sections discuss methods that have been employed in trying to produce better estimates of old age mortality and in the process re-estimating the population numbers at these ages.

2.3 Extinct Generations Method

The method of extinct generations, first proposed by Vincent (1951), is used to estimate the numbers of people at old ages from data on reported deaths alone. The method is based on the fact that cumulating deaths of a given age cohort from a point in time to its extinction gives, if all deaths are reported accurately, the size of the cohort at that point in time. One of the motivations for use of the method is that the death rates seem to be more accurate if the errors present in the reported deaths (numerator) and the populations (denominator) are of the same magnitude (Rosenwaike, 1981). The method of extinct generations can be applied in countries where there are accurate statistics by age and which are closed to migration at least at the advanced ages. The migration assumption has been accepted to hold at old age because this group of people is less likely to be mobile. (Bourbeau and Lebel, 2000; Mesle, Vallin and Andreyev, 2002; Gomes and Turra, 2009). The method is usually applied to estimate the population aged 80 and above (Rosenwaike, 1968, 1979; Thatcher, 1992; Bourbeau and Lebel, 2000; Gomes and Turra, 2009). If the census enumeration is accurate, then any difference between the enumeration and the estimate

derived from the deaths is due to unreported deaths or the inclusion/exclusion of migrants (Rosenwaike, 1968).

The original derivation of the method of extinct generations required that the cohort to be estimated become extinct for an estimate of the population to be derived. While this is theoretically true, it is usually inconvenient to have to wait for the extinction before being able to apply the method. As a result, assumptions are made in order to estimate the small surviving population. This modification to the method is known as the "almost extinct generations method" (Andreev, 2004; Jdanov, Scholz and Shkolnikov, 2005; Jdanov, Jasilionis *et al.*, 2008). Thatcher, Kannisto and Andreev (2002) noted that even Vincent and his assistant sometimes had to estimate future deaths at some of the younger ages.

One of the approaches to estimating the surviving population is the use of survival ratios (Rosenwaike, 1981; Jdanov, Scholz and Shkolnikov, 2005). The surviving population is then estimated by extrapolating the survival ratios observed for earlier cohorts to deaths recorded to date. Thatcher, Kannisto and Andreev(2002) described methods for estimating the surviving population and among them the survival ratio method which works as follows:

Let $D(c,y)$ be the number of deaths in the cohort born in calendar year c observed in the calendar year y. Let $S(c,y)$ be the survivors in this cohort born in calendar year c who are alive at the end of calendar year y. The survivor ratio in this cohort at the end of calendar year y is estimated using deaths in the last k years as follows:

$$SR(c,y) = \frac{S(c,y)}{D(c,y) + D(c,y-1) + D(c,y-2) + ... + D(c,y-k+1)}$$

Considering the cohort born in calendar year, $c-1$, one gets the following corresponding equation

$$SR(c-1,y-1) = \frac{S(c-1,y-1)}{D(c-1,y-1) + D(c-1,y-2) + D(c-1,y-3) + ... + D(c-1,y-k)}$$

In addition, the following identity holds

$$S(c-1,y-1) = S(c-1,y) + D(c-1,y)$$

All the $D(c,y)'s$ are known from reported data and if $S(c-1,y)$ is known, it can be substituted in the above equation to estimate $S(c-1,y-1)$. Once $S(c-1,y-1)$ is estimated, it is then used to estimate $SR(c-1,y-1)$. Assuming that the cohort born in calendar year c has the same survival ratio as cohort born in calendar year $c-1$, one gets,

$$SR(c-1,y-1) = SR(c,y)$$

The above equation is then used to estimate $S(c,y)$. The above derivation assumes that mortality rates are not changing and allows for the changes in successive cohort sizes. Since most countries have implemented policies that aim to reduce mortality, the above assumption leads to under estimates of survivors by age in situations where mortality in those age groups is falling. As a result, the method derives over estimates of mortality rates.

In countries where reported annual deaths are given by age in completed years only as opposed to the completed years and months or days, adjustments have been made in order to estimate the population with a particular age at a particular point in time. The adjustment is made because an individual can have two ages last birthday within a year. If one considers an individual at the beginning of year y aged x last birthday, this would generally mean the individual was born x years ago. At the beginning of year y again, there are some individuals who were born $x-1$ years ago and who are still aged x last birthday. Provided it can be assumed that birthdays occur uniformly over the calendar year, half of the individuals at the beginning of any year y who are aged x last birthday were born $x-1$ years ago and the other half were born x years ago. By the same argument, the deaths that are recorded at age x last birthday at death during say year y are an average of the deaths to population at the beginning of year y aged x and $x+1$ last birthday (Rosenwaike, 1968; Jdanov, Jasilionis *et al.*, 2008). Given the deaths reported over a calendar year at age x last birthday at death for a considerable period of time, the population at the beginning of some year in the past can then be re-estimated by adding averages of the deaths backwards (Rosenwaike, 1968; Mesle, Vallin and Andreyev, 2002; Jdanov, Jasilionis *et al.*, 2008). An application of the method by Rosenwaike (1968) estimated the US population at the beginning of 1951 aged 85 and above. The population at the beginning of year y aged x last birthday $P(x,y)$ can be estimated as follows:

$$P(x,y)=\frac{1}{2}[D(x,y)+D(x+1,y)]+\frac{1}{2}[D(x+1,y+1)+D(x+2,y+1)]+\ldots \text{ where } D(x,y)$$

are the reported deaths at age x last birthday throughout year y (Rosenwaike, 1968; Thatcher, 1992).

2.4 Data quality assessment

The method of extinct generations is built on the assumption that the impact of age misstatement (age exaggeration and age heaping) in data on reported deaths is less than that on the enumerated population, thus we need to assess the quality of reported deaths data. Methods have been derived to assess the quality of reported data on deaths and the major sources of bias are age overstatement at death, age heaping or digit preference and under reporting of deaths. Part of the reason for age overstatement could be that the individual was born before a proper registration system was in place and a higher age is estimated at some later date or due to age exaggeration common in old people when reporting age as opposed to date of birth (Thatcher, 1992). Age heaping arises as a result of people preferring certain ages or years of birth. The data will then show many people in that particular preferred age and a decrease in the number of people in the adjacent ages. Under reporting of deaths is either a result of people not reporting the deaths of their deceased relatives and some deaths may still not have been reported at the time of processing of the vital registration statistics. This section discusses methods for assessing age heaping and age exaggeration in reported deaths and the population data. Under reporting of deaths will be discussed in methods which adjusts for this in sections to follow.

Age overstatement of deaths is assessed by a comparison of ratios of the ratios of deaths arising from the population in question with those from a suitable standard population that has a similar mortality pattern. Jdanov, Jasilionis, Soroko *et al.* (2008) used the ratios, D_{105+}/D_{100+} and D_{110+}/D_{105+} where D_{x+} is the sum of deaths to people aged x and above occurring over a particular year. In the application of these ratios of deaths, the authors compared the ratios with those derived from their chosen 'standard', Sweden, and classified them into quality groups according to the average values of clusters of the countries they were analyzing.

The Whipple's Index of age accuracy is used to measure the degree of age heaping in a population. The Whipple's Index used to assess the general degree of age heaping at ages ending in 0 or 5 in a population is calculated as follows (United Nations, 1955):

$$WI = \frac{\sum_{x=25,5}^{60} N_{x,t}}{\sum_{x=23}^{62} N_{x,t}} * 5 * 100$$, where $N_{x,t}$ is the population aged x last birthday at time t.

The assumption behind the above formula is that the population has an equal number of people in each age group. The standard criteria in Table 2.1 below for measuring age heaping is recommended by the United Nations (Yi and Vaupel, 2003:234). The 'standard' in Table 2.1 is some chosen suitable population that has similar mortality to the population under study and with accurate data by age.

Table 2.1 Whipple's Index grading system recommended by the United Nations

Whipple's Index	Quality of data	Per cent deviation from standard
<105	Very accurate	<5
105-110	Relatively accurate	5-9.99
110-125	OK	10-24.9
125-175	Poor	25-74.99
>175	Very poor	>=75

Demographers who are interested in the study of old age mortality have modified the original Whipple's Index to assess the data quality at old age. Willcox, Willcox, He *et al.* (2008) modified the index to be the sum of numbers at ages 95, 100 and 105 divided by the total number of the population between ages 93 and 107 years and compared this with that from the standard population. Yi and Vaupel (2003) in their study of the oldest old mortality of the Chinese population modified the index to be the sum of numbers at ages 65, 70, 75, ... and up to age 95 divided by the total number of the population between ages 63 and 97 years and compared this with the same index from the standard population. In each case, the method is modified to accommodate the data being analyzed and to make conclusions relative to a chosen standard population. Most research carried out in developed countries on this topic has considered Sweden as the 'standard' because of its reliable demographic data.

In order to explain the method used to assess age heaping at death, some technical terms are reviewed first. If one considers a population aged between x and $x+n$ over a

time period ranging from 0 to T, the period age specific death rate, ${}_nM_x$ can be defined as the number of deaths in the age range x to $x+n$ in the time interval divided by the number of person-years lived in the same age range over the time interval (Preston, Heuveline and Guillot, 2001). Both the numerator and denominator are exact years including fractional years. When summarized in a life table, ${}_nM_x$ can be substituted by a lower case notation ${}_nm_x$. Preston, Heuveline and Guillot (2001) derived a formula that converts the observed age-specific death rates into age specific probabilities of dying, ${}_nq_x$ by the following relationship:

$${}_nq_x = \frac{n.{}_nm_x}{1+(n-{}_na_x){}_nm_x}$$

where ${}_na_x$ is the average number of person-years lived in the interval by the number of people dying in the interval and the people dying are in the cohort aged between x and $x+n$.

To estimate ${}_na_x$ we assume that deaths occur on average, halfway through the age interval and this means that ${}_na_x = n/2$. Assuming ${}_na_x$ for single years, the above equation for ${}_nq_x$ becomes;

$$q_x = \frac{m_x}{1+0.5m_x}$$

Although l_{x+t} (number of people alive at age $x+t$ in any calendar year) is not particularly linear with respect to t at the older ages the approximation for ${}_na_x$ may be considered reasonable since we are working in single ages and considering the ratio q_x/q_{x+1} and thus the impact of any errors will be minimal.

There ought to be a higher probability of dying at age $x+1$ (q_{x+1}) than at age x (q_x) and especially at the advanced ages. A ratio of the probabilities of dying, q_x/q_{x+1}, greater than one signals data problems at age x or age $x+1$. This means that either there are more deaths reported at age x than the actual number of people dying at this age or there are fewer deaths reported at age $x+1$. Kannisto (1988) used the ratio q_{100}/q_{101} to measure age heaping of deaths at age 100 and a value of the ratio that is greater than one by more than random fluctuation is considered a sign of age heaping at age 100. Jdanov, Jasilionis,

Soroko *et al.* (2008) used the same method to identify signs of age preference at death at age 80 and used the ratio q_{80}/q_{81}. A ratio less than 1.05 is considered to indicate good or acceptable quality, a ratio between 1.05 and 1.19 is considered to show moderate age heaping at death and a ratio greater or equal to 1.20 is considered a sign of significant age heaping at death (Jdanov, Jasilionis *et al.*, 2008).

2.5 Death distribution methods

It has been observed that not all people in a population are enumerated during a census and a certain proportion of deaths experienced by the population are also not reported. Methods have been created to estimate this under enumeration in population numbers and under reporting of deaths, relative to one another, in order to estimate mortality rates. The Generalized Growth Balance (GGB) method proposed by Hill (1987) and the Synthetic Extinct Generations (SEG) method (Bennett and Horiuchi, 1981, 1984) are two methods used to estimate the completeness of the reporting of deaths.

The GGB and SEG methods both assume that; the population is closed to migration, coverage of the censuses and the completeness of reporting of deaths are constant with age, and that the ages of population and deaths are accurately reported.

2.5.1 Generalized Growth Balance Method (GGB)

Brass (1975) developed a method for estimating completeness of death reporting relative to the coverage of enumeration of a census which is known as the Brass growth balance method. The method makes use of the demographic balance equation for a stable and closed population at or at and above any age. Completeness of death reporting is assumed to be constant at all ages above early childhood. The method assumes that coverage of census enumeration is constant for all ages. Hill (1987) generalized Brass' method for non-stable populations and the new method is known as the generalized growth balance method. In addition to estimating the completeness of death reporting, the generalized growth balance method estimates relative coverage of two successive censuses.

The demographic balance equation for a closed population is stated as follows:

$$P_2 = P_1 + B - D$$

where: P_1 is population at time 1

P_2 is population at time 2

B denotes births during the period

D denotes deaths during the period

The equation can also be applied over any age range and rewritten as (Hill, 1987);

$$P_2(x+) = P_1(x+) + N(x) - D(x+)$$

where: $N(x)$ is the population reaching exact age x during the period.

$P_1(x+)$ and $P_2(x+)$ are the population aged x and above at the beginning and end of the period respectively

$D(x+)$ are deaths during the period of the population aged x and above.

Rewriting the above equation gives

$$N(x) - [P_2(x+) - P_1(x+)] = D(x+).$$

Dividing by the true population reaching exact age x and above, $N(x+)$, during the intercensal period yields

$$\frac{N(x)}{N(x+)} - r(x+) = \frac{D(x+)}{N(x+)} \quad ---(1)$$

where $r(x+)$ is the growth rate for the population aged x and above.

The death rate for the population aged x and above is directly represented by the right hand side of equation (1) above. If deaths are under reported, the right hand side will be lower than the left hand side provided that the population is accurately reported. Therefore, the death rate can be estimated indirectly using the left hand side of the above equation.

If coverage of the enumeration of the first and second censuses are k_1 and k_2 respectively, and k_3 denotes the completeness of death reporting, then the true values of $P_1(x+)$, $P_2(x+)$ and $D(x+)$ can be expressed as a function of the observed quantities denoted by superscript o, as follows (Hill, 1987; United Nations, 2002):

$$P_1(x+) = P_1^o(x+)/k_1 \quad ---2(a)$$
$$P_2(x+) = P_2^o(x+)/k_2 \quad ---2(b)$$
$$D(x+) = D^o(x+)/k_3 \quad ---2(c)$$

For an intercensal period equal to t years:

$$r(x+) = \frac{1}{t} \ln \frac{P_2(x+)}{P_1(x+)} \text{ by definition}$$

Substituting $2(a)$ and $2(b)$ into the above equation gives

$$r(x+) = \frac{1}{t}\ln\frac{k_1}{k_2} + r^o(x+)$$

where $r^o(x+)$ is the growth rate derived from the observed population numbers.

The population aged x and above during the period can be estimated by the geometric mean of the population aged x and above in the two censuses multiplied by the intercensal period as follows:

$$N(x+) = t*[P_1(x+)*P_2(x+)]^{0.5}$$

Substituting $2(a)$ and $2(b)$ into the above equation gives

$$N(x+) = \frac{1}{(k_1k_2)^{0.5}} * N^o(x+)$$

where $N^o(x+)$ is the population aged x and above derived from the observed population numbers.

$N(x)$ from equation (1) can be estimated using the following approximation (United Nations, 2002):

$$N(x) = \frac{t}{5}[P_1(x-5,5)P_2(x,5)]^{0.5}.$$

The approximation is an interpolation between the population aged $x-5$ at the first census ($P_1(x-5,5)$), who can be assumed to reach exact age x after the first census, and the population aged x at the second census ($P_2(x,5)$), who can be assumed to have turned x prior to the second census. The estimate is then multiplied by the number of five-year intervals in the intercensal period in which these individuals could have turned x.

Substituting $2(a)$ and $2(b)$ into the above equation gives

$$N(x) = \frac{1}{(k_1k_2)^{0.5}} * N^o(x)$$

where $N^o(x)$ are the births into age x estimated from the enumerated censuses.

$D(x+)$ are the deaths to population aged x and above during the intercensal period and are adjusted for completeness as illustrated by equation $2(c)$.

Substituting the derived and adjusted quantities into equation (1) and rearranging yields:

$$\frac{N^o(x)}{N^o(x+)} - r^o(x+) = \frac{1}{t}\ln\frac{k_1}{k_2} + \frac{(k_1k_2)^{0.5}}{k_3}\frac{D^o(x+)}{N^o(x+)}$$

This new equation is a straight line equation. One problem is to obtain the three unknown estimates of completeness simultaneously from the intercept and the slope. Since we are only interested in the completeness of deaths relative to the coverage of the censuses, we set, arbitrarily, the larger of k_1 and k_2 to be one and the other remaining quantities are estimated relative to this. Defining the pairs of independent and dependent variables as x and y respectively, we can fit a straight line for each age to get the estimates of relative completeness. The line could be fitted using one of several techniques of fitting a robust line such as the method by the United Nations (2002).

Hill and Choi (2004) investigated how well the GGB method performs when the assumptions upon which the method is based are violated. In their research, they simulated a population whose mortality is similar to that of the West model life table (Level 15) and imposed a combination of errors on the simulated data. The errors applied in various combinations include age misreporting in both the population and deaths, differential in death and population coverage by age, decline in census coverage over time, omission of deaths and assuming that there is migration. It was observed that the GGB method is sensitive to, among other things, age misreporting and differential coverage of the census by age. Age exaggeration in the enumerated population and reported deaths leads to an over estimate of the reporting of deaths. Differential coverage by age distorts the estimates of census coverage and death reporting completeness. The estimates of k_1 and k_2 are no longer constant with age and as a result both the intercept and gradient vary for different sets of ages. The overall effect is over estimation of the completeness of reporting of deaths.

2.5.2 Synthetic Extinct Generations Method (SEG)

Bennett and Horiuchi (1981) developed a method of estimating completeness of death registration for a non-stable population. Their method is an extension of the method by Preston, Coale, Trussell *et al.* (1980) which estimated completeness of death registration for a stable population. The method simply estimates the population by age from the number of deaths by age. Completeness is then estimated by the ratio of the estimated population to the enumerated population.

The method was derived based on the method of extinct generations proposed by Vincent (1951), as discussed previously, which uses deaths alone to re-estimate the population size at some point in time. This relationship can be expressed mathematically as follows:

$$N(a,t) = \sum_{k=0}^{w-a-1} D(a+k, t+k)$$

where $N(a,t)$ represents the size of the cohort aged a at some point in time t in the past and $D(a+k,t+k)$ are the deaths observed in that cohort from age a up to the last age w, in which the last member of the cohort dies.

However, in order to apply Vincent's suggestion one needs to wait until the cohort is extinct (which could be many years into the future). Thus Preston, Coale, Trussell *et al.* (1980) pointed out that if it can be assumed that the population was stable, then the population at time t can be estimated as follows:

$$N(a,t) = \int_a^w D(x,t) \exp\left(\int_a^x r(z,t)\,dz \right) dx \quad - - - (3)$$

where $D(x,t)$ is the actual number of deaths experienced by people aged x in the cohort in question over time and r is the population growth rate of the population aged z.

$D(x,t)\exp(\int_a^x r(z,t)dz)$ estimates the number of deaths from the population aged a at time t who will die at age x at time $t+x-a$. The above equation is a period representation of Vincent's method of extinct generations. The time variable t in equation (3) may be dropped on the assumption that growth rate is constant at any time t and the equation simplified to:

$$N(a) = \int_a^w D(x) e^{r(x-a)} dx,$$

If completeness of death reporting at age a and above is a constant proportion, k, and the observed deaths in the cohort in question are denoted by a subscript o, one gets,

$D(x) = D^o(x)/k$ for all ages x greater or equal to age a.

By substituting the above equation into the stable population equation above, one gets,

$$N(a) = \frac{1}{k}\int_{a}^{\infty} D^{o}(x)e^{r(x-a)}dx$$

If one can define an estimate of $N(a)$ as:

$$\hat{N(a)} = \int_{a}^{\infty} D^{o}(x)e^{r(x-a)}dx$$

Completeness of death reporting is then estimated as $\hat{N(a)}/N(a)$.

In order to compute $\hat{N(a)}$, Bennett and Horiuchi (1981) derived the estimate as follows:

Taking equation (3) and dropping t, one gets,

$$N(a) = \int_{a}^{\infty} D(x)\exp(\int_{a}^{x} r(z)dz)dx$$

$$= \left\{ \int_{a+n}^{\infty} D(x)\exp[\int_{a+n}^{x} r(z)dz]dx \right\} * \exp[\int_{a}^{a+n} r(z)dz] + \int_{a}^{a+n} D(x)\exp[\int_{a}^{x} r(z)dz]dx$$

$$= N(a+n)\exp[\int_{a}^{a+n} r(z)dz] + \int_{a}^{a+n} D(x)\exp[\int_{a}^{x} r(z)dz]dx$$

If $r(u) = {}_{n}r_{a}$ where $a \le u \le a+n$ and ${}_{n}D_{a} = \int_{a}^{a+n} D(x)dx$, then

$$N(a) = N(a+n)\exp[n.{}_{n}r_{a}] + \int_{a}^{a+n} D(x)\exp[(x-a).{}_{n}r_{a}]dx$$

$$= N(a+n)\exp[n.{}_{n}r_{a}] + {}_{n}D_{a}\exp[y.{}_{n}r_{a}]$$

since there exists a y where $a \le y \le x$, such that

$$\int_{a}^{a+n} D(x)\exp[(x-a).{}_{n}r_{a}]dx = \exp[y.{}_{n}r_{a}]\int_{a}^{a+n} D(x)dx$$

For $n = 5$, and having that $y = 2.5$ and $r(u) = {}_{5}r_{a}$ where $a \le u \le a+5$ and hence

$$N(a) = N(a+5)\exp[5.{}_{5}r_{a}] + {}_{5}D_{a}\exp[2.5.{}_{5}r_{a}]$$

If $\hat{N}(a+5)$ is an estimate of the population aged $x+5$ and ${}_{5}r_{a}$ can be computed, then

$$\hat{N}(a) = \hat{N}(a+5)\exp[5.{}_{5}r_{a}] + {}_{5}D_{a}\exp[2.5.{}_{5}r_{a}] \quad ---(4)$$

where ${}_{5}D_{a}$ is the number of deaths occurring in the age group a to $a+5$.

In order to estimate $\hat{N}(x)$ for all age groups recursively, we need $\hat{N}(x)$ for the last age group, which Bennett and Horiuchi (1981) estimated as follows:

$$\hat{N}(a) = D^{o}(a+)[\exp(r(a+)e(a)) - (r(a+)e(a))^{2}/6] - - - (5),$$

where $r(a+)$ is the growth rate in the open interval and $e(a)$ is life expectancy at the beginning of the open interval (presumed to be available from elsewhere, such as from estimates of mortality produced at some earlier time point). Thus if $r(a+)$ and $e(a)$ are known, $\hat{N}(a)$ for the open interval can be estimated and the other $\hat{N}(x)'s$ can be estimated recursively using equation (4) stated previously. The use of age specific growth rates accommodates the differential in growth rates within a population by age.

If the growth rate, ${}_5r_a$, is estimated from two successive censuses, it may be distorted by the differential enumeration between the two censuses. If the first census was under enumerated relative to the second census, the age specific growth rates will be biased upwards and vice versa if the second census was under enumerated relative to the first census. If the proportional differences in census coverage do not vary with age, $\ln(k_2/k_1)/t$ can be used to adjust the age specific growth rates, where k_1 and k_2 are the estimates of completeness of the first and second census respectively, and t, is the length of time between the censuses (Bennett and Horiuchi, 1981). The adjusted age specific growth rates ${}_5r_a^*$ becomes,

${}_5r_a^* = {}_5r_a + \delta$ where $\delta = \ln(k_2/k_1)/t$ as defined above.

This research will refer to this method as the SEG+delta method since it is different from the common application of the SEG method which does not adjust for differential coverage by age between censuses.

The method was applied by Bennett and Horiuchi (1981) to data on the Swedish male population and they reproduced almost the same population numbers by age from the number of deaths by age. Application of the method by the same authors to Korean female population yielded an improved estimate of completeness from that produced by the Preston, Coale, Trussell *et al.* (1980) method. The use of age specific growth rates produced a smoother set of estimates of the completeness of reporting of deaths compared to that produced assuming stability of the population.

Hill and Choi (2004) investigated how well the SEG method performs when the assumptions upon which the method is based are violated. It was observed that the method performs perfectly in the absence of data errors. Age exaggeration in both the population numbers and the reported deaths leads to an over estimate of the completeness of the reporting of deaths. They concluded that differential coverage by age over time leads to an over estimate of death reporting completeness but they applied the method without adjusting for this differential completeness (the addition of delta). Dorrington and Timaeus (2008) challenged the conclusions reached by Hill and Choi (2004) and showed that the method with the correction suggested by Bennett and Horiuchi (1981) on the same scenario performs well in the presence of differential coverage by age between two censuses.

2.6 Mortality Estimation

After estimating the relative completeness between successive censuses, the observed population numbers are then adjusted accordingly to get consistent estimates of the population at the two time points. As defined previously that $P_1(x,5)$ and $P_2(x,5)$ are the adjusted populations aged between x and $x+5$ at the first and second censuses respectively, the number of person-years lived between age x and $x+n$ during the intercensal period (t), $PYL(x,5)$, is estimated, assuming that the population is changing linearly over the intercensal period, as follows, $PYL(x,5) = t[P_1(x,5)P_2(x,5)]^{0.5}$ (United Nations, 2002).

The deaths aged between x and $x+5$ are adjusted for incompleteness and can be denoted by $D(x,5)$. The death rate in the age range x to $x+n$ for the period, ${}_nm_x$, is then calculated as the sum of all adjusted deaths in that age range occurring over the intercensal period divided by the number of person years lived during the intercensal period in the same age range.

3 Methodology

3.1 Background on data

This current research uses data from two censuses that were carried out in South Africa in 1996 and 2001, both with the night of 9-10 October as reference date. In 2007, another nationally representative large scale household survey, the Community Survey, was carried out in February 2007. The survey had no reference date and for purposes of this research, it will be assumed to be the night of 14-15 February 2007.

In both the censuses and survey questionnaires, there were fields where an individual was supposed to provide his or her date of birth and age in completed years. In cases where the age and date of birth were inconsistent, the date of birth was preferred and used to calculate the age. When neither the age nor the dates of birth were given, the enumerators had to estimate the age as accurately as possible. The age in single years was accepted only up to 120 years.

Death registration in South Africa is done by the Department of Home Affairs (DHA) (Statistics South Africa, 2008b). After a death is registered, a death certificate is issued. The DHA then creates a death notification form that is eventually sent to Statistics South Africa (Stats SA) for processing. The death notification form has information on the date of death, date of birth and age at last birthday.

The death notification forms from the respective provinces are sometimes delayed in reaching the DHA and hence are not processed by Statistics South Africa as quickly as might be desired. This research used reported deaths obtained from Statistics South Africa and therefore there is the chance that some deaths may not have been included in this research as the death notification forms were still outstanding.

The data on deaths obtained from Statistics South Africa are given by sex and single ages for each calendar year of death from 1997 to 2006. As publicly available data aggregate data at advanced ages, details of deaths at individual ages were provided by Stats SA on special request. Recorded deaths for the year 1996 were obtained from Statistics South Africa (2001).

There are deaths with unspecified sex for particular ages within a year and these were proportionally reallocated to the male and female deaths in that age group. Deaths with

unspecified age for each sex are ignored since they will be adjusted for using death distribution methods.

The current research also uses mid-year population estimates of the South African population that were published by the UN Population Division (United Nations, 2007) and the US Census Bureau (US Census Bureau, 2005). Both use the cohort component projection method which tracks a group of people and exposes them to assumed age - specific fertility, and age and sex-specific mortality and migration rates to get an estimate of the population in future.

In applying the cohort component projection method, the UN Population Division assumes that total fertility will converge to a level of around 1.85 children per woman by the period 2045-2050. It is also assumed that fertility will follow a path derived from models of fertility decline established by the UN Population Division based on past experience. The projected fertility trends are then compared with recent trends of fertility derived using recently available data and adjusted accordingly for consistency. The age specific fertility rates are derived as a function of total fertility rate by interpolation. The mortality pattern for South Africa is projected on the basis of the UN East Asia model life table and the changes in life expectancy of the population. The mortality patterns of the model life table are also adjusted to include the impact of HIV/AIDS and the adjustments are based on recently available census/survey data. Future trends of the epidemic are projected yearly and ensure that the trends are also consistent with recent population data.

The USCB projects future mortality patterns by fitting a logistic curve on the estimates of life expectancy at birth ignoring the impact of HIV/AIDS and ensuring that the projected curve yields an acceptable projected level of the South African data. After assumptions about future level of life expectancy are made, age patterns of mortality are then developed using an iterative interpolation process. The age patterns of mortality are adjusted to incorporate the future patterns of HIV/AIDS. Future patterns of fertility are projected by fitting a logistic curve and ensure that the fitted curve is consistent with recent data from surveys.

Both the UN Population Division and the US Census Bureau use international migration to improve the consistency between the projected population and the most recently available census/survey data. International net migration is usually low and the estimation of international migration is speculative compared to that of fertility and

mortality. The base population in each case is derived from either census data or reliable mid-year population estimates.

3.2 Data Exploration

The first investigation was to determine the cut-off age above which the deviation between the population count and the numbers estimated from the deaths might be expected to be significant.

Census data by sex and single years for the years 1996 and 2001, and the Community Survey data for 2007 were downloaded from the Statistics South Africa website (Statistics South Africa, 2009). For both the 1996 and 2001 censuses, the Community profile descriptive data weighted by 'person weight' were downloaded.

Estimates of the population at each of the censuses and survey dates were interpolated from each of the midyear estimates by the US Census Bureau (US Census Bureau, 2005), UN Population Division (United Nations, 2007) and a demographic projection model (ASSA2003lite) produced by the Actuarial Society of South Africa (Actuarial Society of South Africa, 2005). The US Census Bureau (USCB) and the UN Population Division (UNPD) estimates are online databases. The USCB and ASSA provide annual estimates whereas the UNPD estimates are five-yearly.

The population count at each of the censuses and the Community Survey dates in quinqennial age groups up to age 100 and above are divided by the corresponding expected numbers from each of USCB, UNPD and ASSA estimates to determine the pattern of deviation. As expected, if the counted population numbers are matching the estimated population numbers, their ratios must be one or very close to one. The cut-off age was determined by inspection of the ratios and a decision made to choose the age where the excess of the enumerated population start to increase significantly by age relative to all three estimates.

3.3 Extinct generations method

After the cut-off age was determined, the method of extinct generations was then applied to estimate this population as at the 1996 and 2001 censuses. Recapping from Chapter 2 and letting $D(x, y)$ be the reported deaths to population aged x last birthday at death in year

y, one can then estimate the population at the beginning of year y aged x last birthday, $P(x, y)$ as follows:

$$P(x,y) = \frac{1}{2}[D(x,y) + D(x+1,y)] + \frac{1}{2}[D(x+1,y+1) + D(x+2,y+1)] + \ldots$$

The above formula assumes that deaths are uniformly distributed in the calendar year of death and birthdays are independently and uniformly distributed over the calendar year.

Estimates of the population are required at a time that is not at the beginning of the census year and the above derivation has to be adjusted to estimate the population at the census date. If one considers the population at the census date aged x last birthday and the census is at a time equal to f before the end of year y, one needs to adjust the above formula so that one can estimate this population using only deaths to the population aged x last birthday at death from time $1-f$ onwards. There is a need to adjust the above formula to include deaths that might have occurred at age x last birthday after the census date during year y but were aged $x-1$ last birthday at some point during the same census year y and then weight by the proportion of time to the census date.

In adjusting, each component to the right hand side of the above equation will include deaths to population aged $x+i-1$ in year $y+i$ and the deaths to this population can only occur at age $x+i$ last birthday from point $1-f$ onwards in year $y+i$ if they turned $x+i$ in the interval of time $(0,1-f)$. By the same argument, the population aged $x+i+1$ last birthday in year $y+i$ can only die at age $x+i$ last birthday in year $y+i$ in the interval of time $(1-f,1)$. After allowing for this, the above formula becomes;

$$\begin{aligned} P^f(x,y) = f * \frac{1}{2} * \{(1-f) * [D(x-1,y) + D(x,y)] + f * [D(x,y) + D(x+1,y)]\} + \\ \frac{1}{2} * \{(1-f) * [D(x,y+1) + D(x+1,y+1)] + \\ f * [D(x+1,y+1) + D(x+2,y+1)]\} + \ldots. \end{aligned}$$

where, $P^f(x,y)$ is the population at some point in time which is f before the end of year y aged x last birthday. This can be simplified to

$$\begin{aligned} P^f(x,y) = f * \frac{1}{2} * \{(1-f) * D(x-1,y) + D(x,y) + f * D(x+1,y)\} + \\ \frac{1}{2} * \{(1-f) * D(x,y+1) + D(x+1,y+1) + f * D(x+2,y+1)\} + \ldots \end{aligned}$$

In this research, f is the time between the census date and the end of the census year, and is equal to 0.227 for the two censuses.

In addition, we need to estimate the number of deaths by age beyond year 2006 so that all the deaths for a particular cohort until its extinction are available. It is assumed that the proportion of deaths within an age cohort by age, say age x in year y, to the observed deaths in the last, say 5 years, in that cohort is the same as the corresponding proportion of deaths in age cohort aged x in year $y+1$. The 5 years have been chosen arbitrarily and if one can compute the observed death ratio of a cohort aged x in year y, one can then use this ratio to extrapolate the unknown deaths in the younger and adjacent cohort which will be aged x in year $y+1$.

The death ratio derived using deaths in the last five years at age x, DR_x was computed as follows:

$$DR_x = \frac{D(x,y)}{D(x-1,y-1)+D(x-2,y-2)+D(x-3,y-3)+D(x-4,y-4)+D(x-5,y-5)}$$

The unknown deaths, $D'(x,y+1)$, expected in the year, following year y for the same age x are then estimated as:

$$D'(x,y+1) = DR_x * [D(x-1,y)+D(x-2,y-1)+D(x-3,y-2)+D(x-4,y-3)+ D(x-5,y-4)]$$

For example, to estimate the number of deaths at age 86 in year 2007, $D(86,2007)$, one need first to estimate the death ratio as follows:

$$DR_{86}^{06} = \frac{D(86,2006)}{D(85,2005)+D(84,2004)+D(83,2003)+D(82,2002)+D(81,2001)}$$

The deaths occurring at age 86 in year 2007 are then estimated as

$$D(86,2007) = DR_{86}^{06} * [D(85,2006)+D(84,2005)+D(83,2004)+D(82,2003)+ D(81,2002)]$$

The extrapolation of deaths is done until the estimate of deaths at age 110 and above for each cohort is obtained. The age 110 is chosen on the assumption that few people survive beyond 110 and deaths occurring after age 110 are aggregated at this age (Rosenwaike, 1968, 1979; Jdanov, Jasilionis *et al.*, 2008). Once the unknown deaths in the years beyond 2006 for each age cohort aged 75 and above at the 1996 census are estimated,

the population aged 75 and above by single ages as at the 1996 census date can then be estimated. This procedure is repeated to estimate the population aged 75 and above at the 2001 census.

However, these numbers can be expected to underestimate the population since the reported deaths used to estimate the numbers are incompletely reported, and so they need to be adjusted. The ratio of the estimated to the enumerated population aged 75 and above was assumed to estimate the completeness of reporting of deaths and on the assumption that completeness of registration was the same for each age, the ratio was used to adjust the other age groups. Thus let $\hat{P}(75+)$ be the estimate of the population aged 75 and above, and $P(75+)$ be the enumerated population aged 75 and above, the adjustment factor (k) is then estimated as follows:

$$k = \frac{\hat{P}(75+)}{P(75+)}$$

The estimate of the population aged x last birthday after adjusting for under reporting, $\hat{P}_{adj}(x)$, as at the date of the census is now estimated as follows:

$\hat{P}_{adj}(x) = \frac{\hat{P}(x)}{k}$, where $\hat{P}(x)$, is the initial estimate of the population aged x which is not adjusted for completeness. These adjusted estimates of the population are then compared with the enumerated population.

The population aged 75 and above at the 2007 survey is estimated differently since there are no deaths to which the method of extinct generations could be applied. After obtaining adjusted estimates of the population at each of the 1996 and 2001 censuses, five year survival ratios are computed as follows:

$$_5S_x = \frac{\hat{P}_{adj}(x+5,2001)}{\hat{P}_{adj}(x,1996)}$$

where $\hat{P}_{adj}(x,y)$ is the estimate of the population at the census date in year y aged x last birthday and x ranges from 70 to 100 and above.

Applying these survival ratios to the estimates of the 2001 census gives us estimates of the population aged 75 and above on the night of 9-10 October 2006. This is the

estimate of the population expected if a census was carried out exactly after five years from the 2001 census date and the population experiencing the same survival ratios estimated between the 1996 and 2001 census estimates.

Estimates of the population at the date of the Community Survey in 2007 are then obtained by extrapolating estimates of the population on the night of 9-10 October 2006 using the growth rates for ages 75 to 100 and above between the 2001 census estimates and the 2006 estimates. The estimates of the population aged 75 and above at the survey date can then be compared with the enumerated population numbers.

3.4 The Synthetic Extinct Generations and Generalized Growth Balance methods

The GGB and SEG+delta methods are applied to explore the patterns of completeness of the reported deaths and the relative completeness between the censuses and the survey for the ages 15 to 100 and above. To apply both the GGB and SEG+delta methods for the period between the 2001 census and the 2007 survey, deaths that occurred over the period between the beginning of 2007 and the survey date (night of 14-15 February 2007) have to be estimated. They are estimated by averaging deaths over the five year period from 2002 to 2006 and multiplying by 0.125 (the proportion of time of a year from the beginning of January to the middle of February in 2007).

A robust straight line fitting technique recommended by the United Nations (2002) was used in the application of the GGB method. The pairs of points (x, y) are first plotted and used to identify outlying points which will be excluded from the calculations of the slope and the intercept. The method then divides the remaining data into three groups according to the x values: lower third, middle third and upper third. The median of the pairs of points in the lower third (x_l, y_l) and the median of the pairs of points in the upper third (x_u, y_u) are used to estimate the slope of the line that passes through these median points. An intercept for each point is then calculated using the estimated slope and the intercept of the line is then estimated as the median of these intercept values (United Nations, 2002).

The SEG+delta method requires life expectancy at the age of the open age interval in order to estimate the populations at younger ages recursively from deaths. In order to estimate the life expectancy, the GGB method is first applied to adjust for relative coverage

of two censuses and completeness of reporting of deaths. Death rates over the interval of time can then be estimated and used to compute life expectancy. This research makes use of an idea from Dorrington and Timaeus (2008) of using the average of the life expectancy from the West family of the regional model life tables (Coale, Demeny and Vaughan, 1983), corresponding to the ${}_5m_{60}$, ${}_5m_{65}$ and ${}_5m_{70}$ estimated from the population. The mortality estimates used to estimate life expectancy in this way are derived from the population data after adjusting for completeness using the GGB method. The values are obtained by interpolation and the averaging based on old age mortality is done to avoid estimating life expectancy using measures that are affected by HIV/AIDS.

In applying the SEG+delta method, the median of the ratios of the estimated population to the observed population ($\hat{N}(x)/N(x)$) is used to estimate completeness of death reporting assuming that the completeness of death reporting relative to census coverage is constant at all ages. The ratios used exclude the young age groups (0 to 15 years) since it is assumed that the extent of their coverage is different from the rest of the population and the age of the open interval is 100. In order to estimate completeness that is constant for all ages, the δ (delta) defined previously is adjusted until a flat sequence of the ratios is obtained.

The population aged 75 and above is replaced by the estimates from the method of extinct generations. The GGB and SEG+delta methods are again applied to this adjusted population and the changes in interpretation and any improvements in distortions noted.

3.5 Data quality

Before getting into the detail of the methods applied to assess the quality of the data used in this research, it is helpful to reflect on the general quality of the South African data. Both the 1996 and 2001 censuses under-counted the national population and the published population numbers were obtained by adjusting each of the census enumeration using data from post enumeration surveys. The under count of the national population was 10.7 per cent (Statistics South Africa, 1998) and 17.6 per cent (Statistics South Africa, 2004), in the 1996 and 2001 censuses respectively. The 2007 survey did not enumerate the total national population and therefore the realized sample was adjusted to get the national population in such a way that the data are consistent both internally and with other censuses/surveys (Statistics South Africa, 2008a).

The registration of vital statistics in South Africa, as in most other African countries, is far from complete. Statistics show that the under-registration of deaths is common in rural areas and among children, but data collected in recent periods suggest that there has been an improvement in the completeness of registration of deaths (Statistics South Africa, 2008b). In addition, the data are also missing detail in some cases on variables such as population group, age last birthday at death and sex of the deceased.

3.5.1 Age heaping at death

Age heaping in the reported deaths is assessed using ratios of the probabilities of dying. The research uses the ratio q_x / q_{x+1} and a ratio greater than one is a sign of age digit preference at death at age x for x ranging from 75 to 98. The choice of age 98 is a result of the fact that probabilities of death for the open interval age (age 100) cannot be used to derive these ratios and the maximum age of the denominator is 99. Given the age at death, the average year of birth can be estimated by deducting the age from the census year. For this measure, the research assesses age heaping at death for the census years only. Age heaping as a result of preferring certain years of birth are assessed using these ratios of probability of dying as well. A ratio greater than one at age x is assessed to decide whether there is a pattern of preferring the terminal digit of that age or the heaping is generally observed within a cohort over time because of the preference of year of birth.

In order to estimate q_x, it is assumed that the estimated population at each of the census dates at the 1996 and the 2001 censuses do not differ much from the estimates as at

the middle of the respective years. There is a difference of slightly more than three months between the middle of each census year and the census date, and even for a population growing at an average rate of say 3 per cent per annum, the difference in population numbers between points in time three months apart will be less than one per cent. The deaths reported aged x last birthday at death over a year divided by the mid year estimates of the population gives the central death rate, m_x, for the population at the middle of the year aged x last birthday. It is assumed that for individuals who die within an age group, they survived half of the time in the interval of the age group and so q_x is estimated as follows:

$$q_x = \frac{m_x}{1 + 0.5m_x}$$

The assumption that deaths occur half way through the interval at old age may be considered reasonable since the research is working in single ages and considering the ratio q_x / q_{x+1}, the impact of any errors will be minimal.

3.5.2 Age heaping at enumeration

After observing the ages and years of birth at which the reported deaths are heaped, another independent assessment of the presence of age and birth year preference is done on the census/survey population numbers. The Whipple's Index of age accuracy discussed in Chapter 2 is modified to assess age heaping at the old ages only.

Age heaping which occurs as a result of age digit preference is assessed at ages which are multiples of 5, that is those ages ending with a 0 or 5. Since the research is more concerned with the quality of data of the old age population, the indices are derived for ages between 73 and 87 inclusive and this would mean that the old age population is defined from age 70 and above. Age 87 was chosen as the upper limit since data by single ages from the standard population that will be used for comparison is limited to age 89. The age groups being considered are no longer rectangular, but more pyramidal in structure and the research uses the criteria in the last column of Table 2.1 in Chapter 2 to identify age heaping.

Following the above discussion, the standard Whipple's Index is then modified as follows:

$$WI = \frac{\sum_{x=75,5}^{85} N_{x,t}}{\sum_{x=73}^{87} N_{x,t}} * 5 * 100$$

where $N_{x,t}$ is the population aged x last birthday at time t.

The year of birth heaping is assessed at years where the population in the age groups between age 70 and 89 could have been born. The research simply subtracted the age last birthday at census date from the census year to estimate the average year of birth. An example is that of people at the census date in 1996 aged 76 last birthday and these are assumed to have been born mostly in 1920 (1996 - 76). The 2007 survey data was handled slightly differently since the survey was carried out in February which is around the beginning of the year and hence the population was mostly born a year earlier than the difference suggests.

The Whipple's Index is modified to assess year of birth heaping at years where the population could most probably have made reference especially in periods of mass registration and the years assessed are 1910, 1920 and 1930. These are the years with terminal digit zero which could be easy to remember or given as the year of birth. The Index used to assess the 1996 census data for year of birth digit preference for a year with a terminal digit 0 is calculated as follows:

$$WI = \frac{\sum_{y=1910,10}^{1920} N_{y,1996}}{\sum_{y=1907}^{1926} N_{y,1996}} * 10 * 100$$

where $N_{y,1996}$ is the population born in year y and enumerated at the 1996 census.

People could have found it easier to remember the beginning and the end of World War I and make reference of their birth year to this period. The research investigated this by assessing year of birth preference at years 1914 and 1918. The Index used to assess the 1996 census data for year of birth preference at the year 1914 was calculated as follows:

$$WI = \frac{N_{1914,1996}}{\sum_{y=1913}^{1915} N_{y,1996}} * 3 * 100$$

where $N_{y,1996}$ is the population born in year y and enumerated at the 1996 census. The 3 years in the above equation are chosen instead of 5 or 10 so as to avoid the large variation in population numbers by age at old age which leads to the violation of the assumption of rectangularity.

The Index used to assess the 1996 census data for year of birth preference at the year 1918 is calculated as follows:

$$WI = \frac{N_{1918,1996}}{\sum_{y=1917}^{1919} N_{y,1996}} * 3 * 100$$

where $N_{y,1996}$ is the population born in year y and enumerated at the 1996 census.

The same formulas of indexes are applied to the 2001 census data and the 2007 survey data but the years of birth being considered have to be restricted in such a way that they derive the population aged between 70 and 89 at each of the 2001 census and the 2007 survey.

The above derivation will manage to assess heaping at years of birth in which the population aged between 70 and 89 could have been born. The research is not able to assess independently the heaping at other years where the populations born in those years were already aged above 89 years at the 1996 census.

3.5.3 Age overstatement at death

Age overstatement at death is assessed by death ratios. Deaths ratios used by Jdanov, Jasilionis, Soroko *et al.* (2008) are slightly modified to use with lower ages and the research is using the ratios, D_{80+}/D_{75+} and D_{85+}/D_{80+}. The same ratios are calculated using deaths from the ASSA2003lite model and a comparison done to determine whether there was a general age exaggeration at death and note any patterns of improvement over time relative to the ASSA estimates.

The methods which need to compare with a standard population are likely to face limitations, for example, the identified standard population has data by single ages which go up to age 89 and we would like to assess the quality of the data up to age 100. Therefore, for those particular measures of quality; conclusions reached are limited to the age groups assessed. Another limitation is the estimation of year of birth both for individuals who are

alive and those that are dead. It has been assumed that the year of birth is the difference between the census/survey date and age last birthday. Since the censuses were carried out in October and if it is assumed that deaths and birthdays are uniformly distributed over a year, this would imply that the majority of the population with a particular age in October is a composition of the population born between January and October of the birth year. Thus any conclusion reached may be more marked because of the mixture of individuals with different birth years but with same age last birthday in a particular year.

4 Data analysis and results

4.1 Exploration of the cut-off age

In determining the cut-off age above which it is necessary to re-estimate the old age population, ratios of the census/survey population to those estimated by the projection models were used and results of the ratios are as shown in Figures 4.1 to 4.3 below.

Figure 4.1 Ratio (Enumeration/USCB), by age, 1996 and 2001 censuses, and the 2007 survey

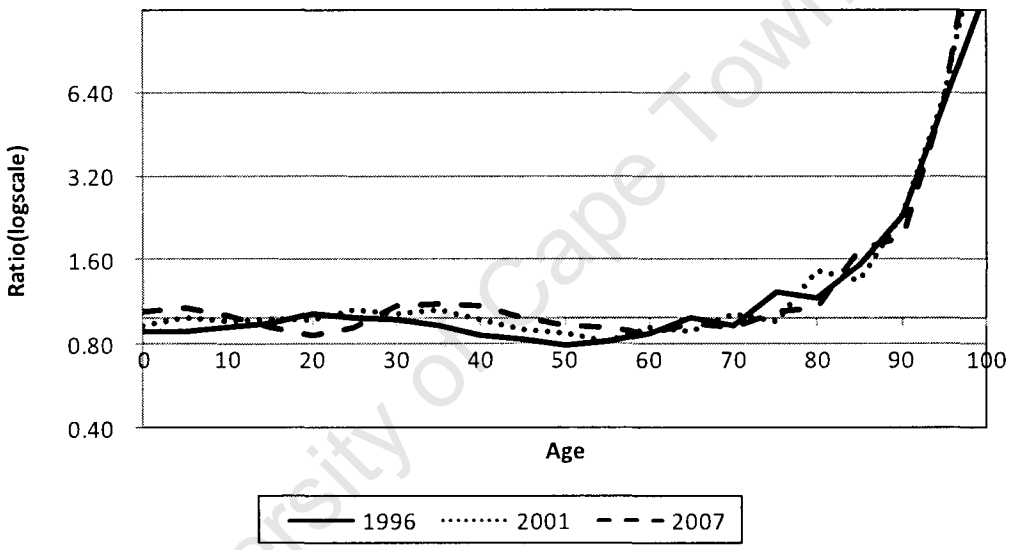


Figure 4.2 Ratio (Enumeration/UNPD), by age, 1996 and 2001 censuses, and the 2007 survey

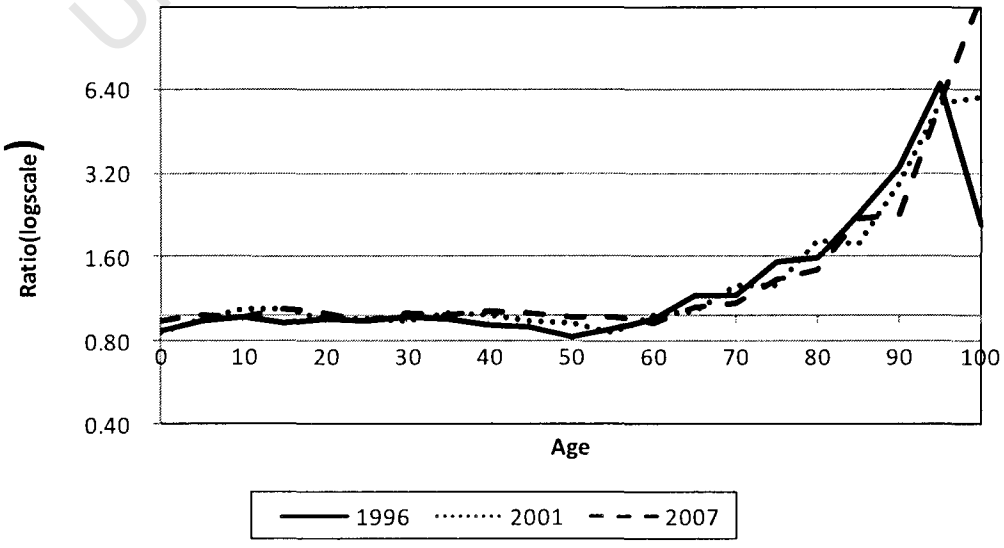
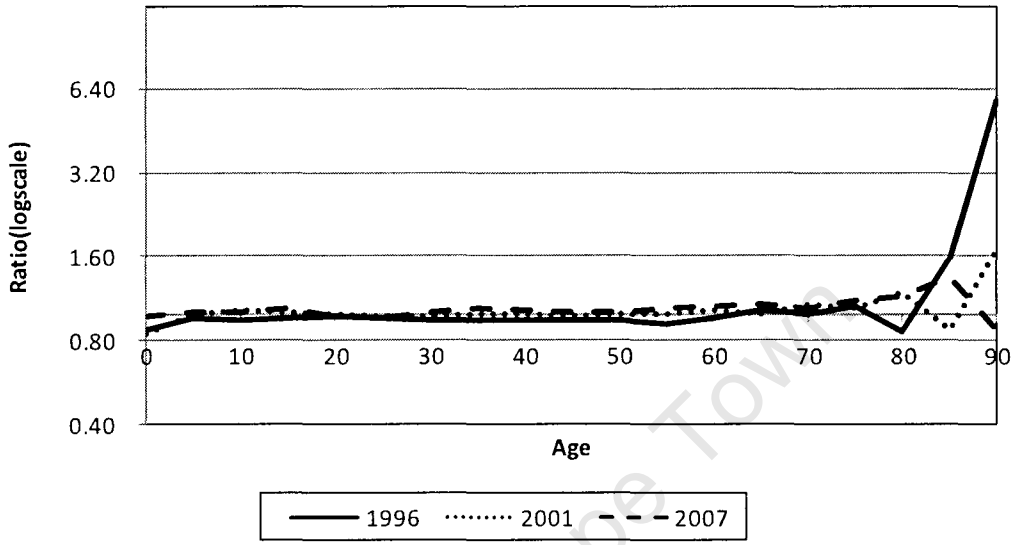


Figure 4.3 Ratio (Enumeration /ASSA), by age, 1996 and 2001 censuses, and the 2007 survey



One can see from the above figures that the excess of the census/survey population over the population estimated from deaths increases with age. By inspection, the survey population becomes excessively higher relative to the USCB and UNPD population estimates from age 75 and above. However, there are other inconsistencies observable in the comparison between the census/survey population and estimates by the USCB and UNPD for ages which are below the age of 60 above which age the population is defined as old age. This research is interested in the discrepancies observed at old age. A comparison with the ASSA population estimates shows that the discrepancies start from age 80 and above. The ratios suggest that ASSA estimates are more consistent with the census/survey population. The estimates from USCB and UNPD are lower than the counted population at ages as low as 75 which suggest that the two sources of estimates are probably over estimating the South African old age mortality. If the ASSA estimates of the population are correct, then either the difference between the ASSA estimates and the census/survey population is because the population was over counted from age 80 and above or because there is age exaggeration. Following these observations, it was decided to re-estimate the population numbers from age 75.

Since ASSA estimates of the population are consistent with the census/survey population, ASSA results are used as the standard against which to compare certain indices

of measure. The measures which are used to check for data quality in reported deaths compared with those derived from the ASSA estimates are the ratios of deaths for checking age overstatement and the Whipple's Index of age accuracy for checking age heaping.

4.2 Extinct generations method

The first results of interest derived from the method of extinct generations are the adjustment factors (in Table 4.1) used to adjust for completeness of the re-estimated population numbers at old age so that they equal the census/survey population aged 75 and above as shown in Table 4.2 below.

Table 4.1 Per cent estimates of the completeness of reporting of deaths

Sex	1996-	2001-
Male	91.2	91.9
Female	94.8	92.2

The population aged 75 and above at each of the 1996 and 2001 censuses, and the 2007 survey estimated using the method of extinct generations and adjusted using the factors in Table 4.1 are shown in Table 4.2 below.

Table 4.2 Census/survey and re-estimated old age population, by age and sex, 1996 and 2001 censuses, and the 2007 survey

	Census/survey		Re-estimated		Per cent deviation	
Age group	Male	Female	Male	Female	Male	Female
1996						
75-79	142 013	235 500	144 000	239 875	-1.4	-1.8
80-84	62 203	117 123	62 872	116 690	-1.1	0.4
85-89	28 985	62 313	27 801	60 616	4.3	2.8
90-94	9 944	23 204	9 268	21 394	7.3	8.5
95-99	3 949	7 359	3 358	7 585	17.6	-3.0
100+	496	1 584	290	919	71.0	72.4
2001						
75-79	136 353	231 190	137 639	232 828	-0.9	-0.7
80-84	90 845	180 097	87 000	175 035	4.4	2.9
85-89	28 928	65 303	31 759	72 478	-8.9	-9.9
90-94	11 265	30 448	11 676	30 198	-3.5	0.8
95-99	4 273	11 171	3 847	8 811	11.1	26.8
100+	1 636	4 367	1 378	2 974	18.7	46.9
2007						
75-79	163 118	316 968	170 001	328 496	-4.0	-3.5
80-84	87 683	176 113	86 353	173 408	1.5	1.6
85-89	47 121	113 097	45 706	110 659	3.1	2.2
90-94	12 564	32 602	13 709	36 300	-8.3	-10.2
95-99	6 652	15 590	4 868	12 495	36.6	24.8
100+	4 367	9 243	868	2 255	403.0	310.0

The distribution of the population estimated from deaths and census/survey population numbers aged 75 and above by single ages and sex are as shown in Figures 4.4 to 4.6 below.

Figure 4.4 Census and estimated population numbers, by age and sex, 1996 census

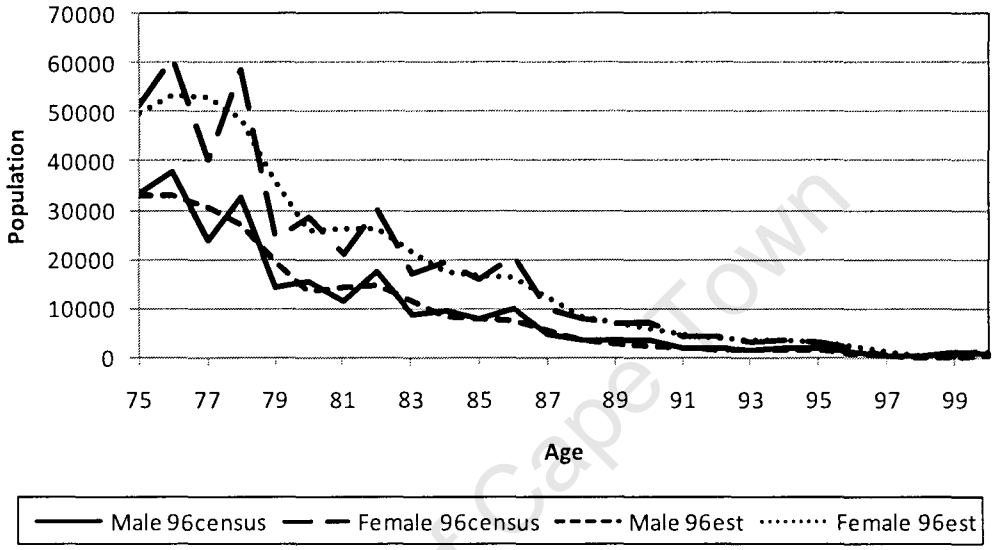


Figure 4.5 Census and estimated population numbers, by age and sex, 2001 census

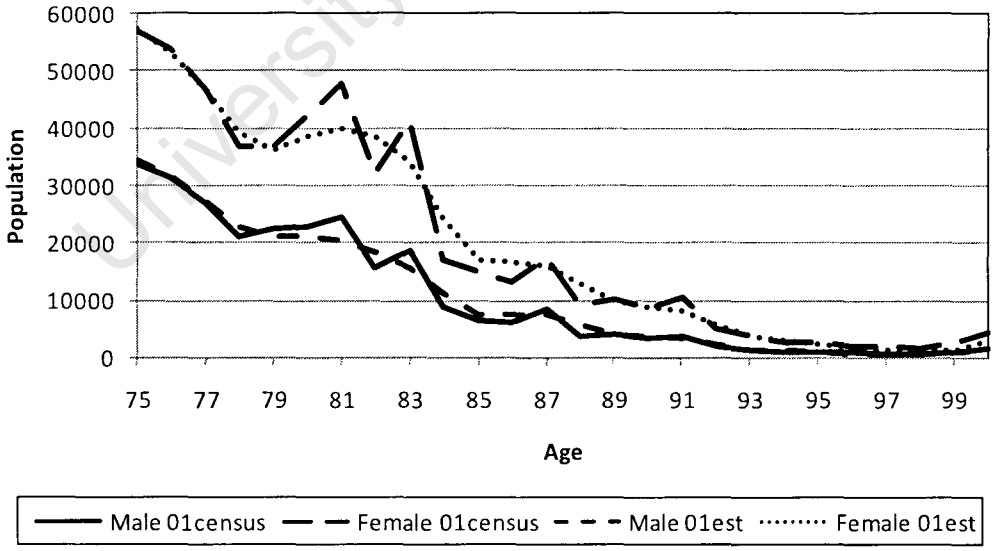
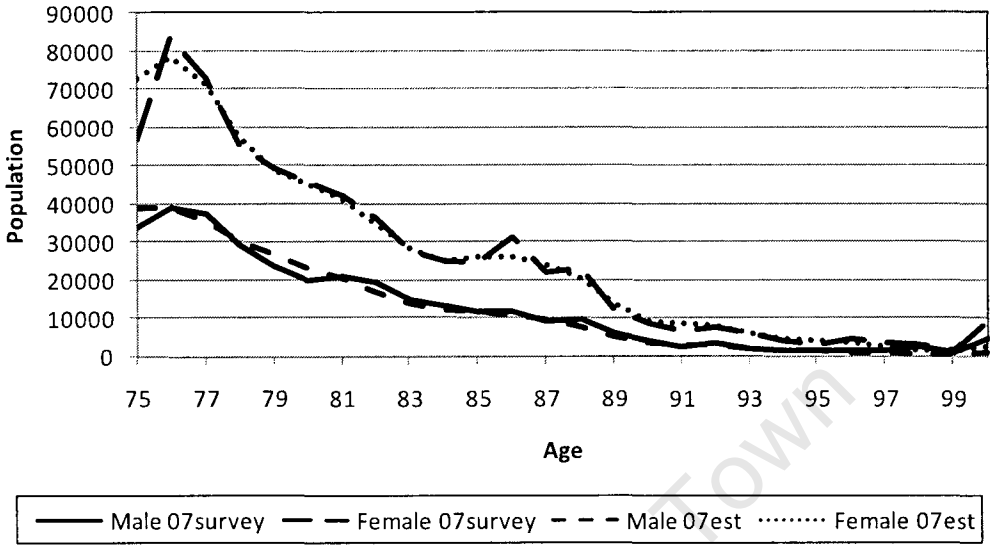


Figure 4.6 Survey and estimated population numbers, by age and sex, 2007 survey



The population numbers estimated from deaths in the above figures have a smooth but otherwise consistent distribution by age compared to the census/survey population numbers. These results suggest that re-estimation of the population numbers using reported deaths smoothens the age heaping observed in the census/survey population numbers.

The population numbers at the oldest age groups are small and the plots above may not show the differences between the census/survey and the re-estimated old age population numbers. For this reason, ratios of the census/survey population numbers to the numbers re-estimated from deaths are as shown in Figures 4.7 to 4.9 below.

The plot below (Figure 4.7) shows some noise around the ratio of one and the noise are fairly small from age 75 to 96. Large distortions are observed from age 97 to 100 and above, and one can see an excess in the census/ survey population from age 99 and above relative to the population estimated from the deaths. Given that the estimates are an aggregate of deaths for a range of ages which would smooth out any age specific fluctuations, the fluctuations in the ratio are a result of fluctuations in the census/ survey population. The deficit of people enumerated at age 97 relative to the estimates from deaths and the sudden excess at ages 99 and above would suggest some element of age exaggeration in the population at the younger ages. Most probably there are some people who should have been enumerated at age 97, but they were enumerated at a higher age. The distortions are in both sexes, with the age exaggeration generally more in males than in females.

Figure 4.7 Ratio (census/estimates), by age and sex, 1996 census

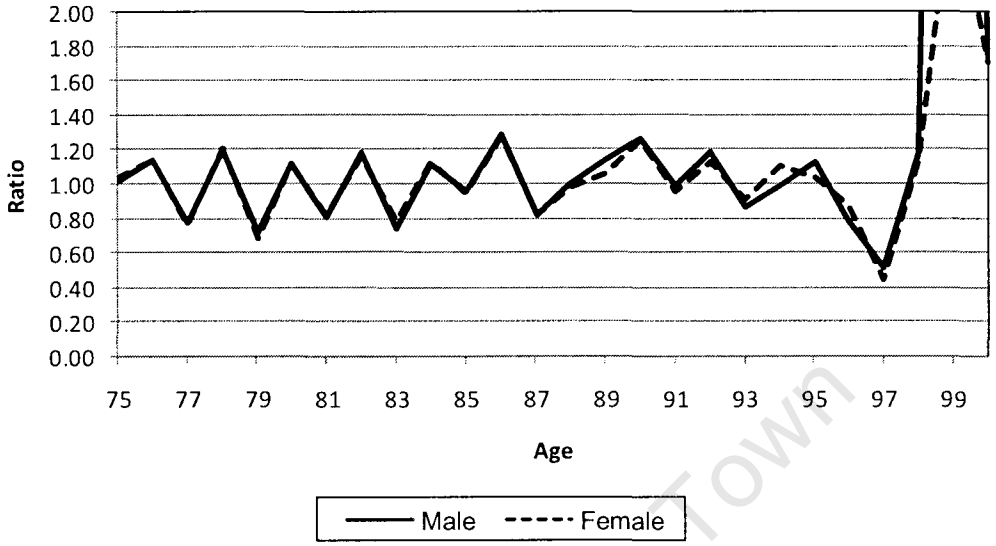


Figure 4.8 Ratio (census/estimates), by age and sex, 2001 census

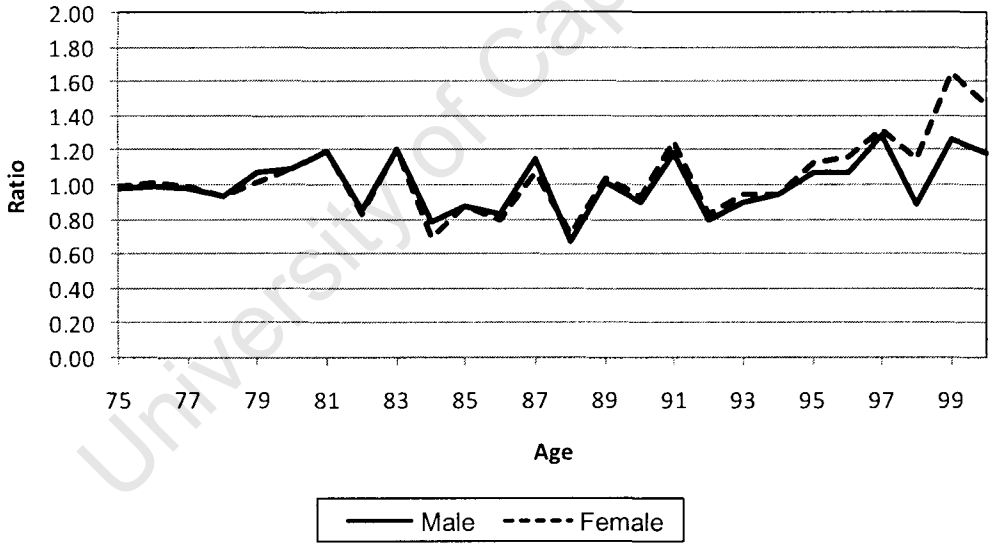


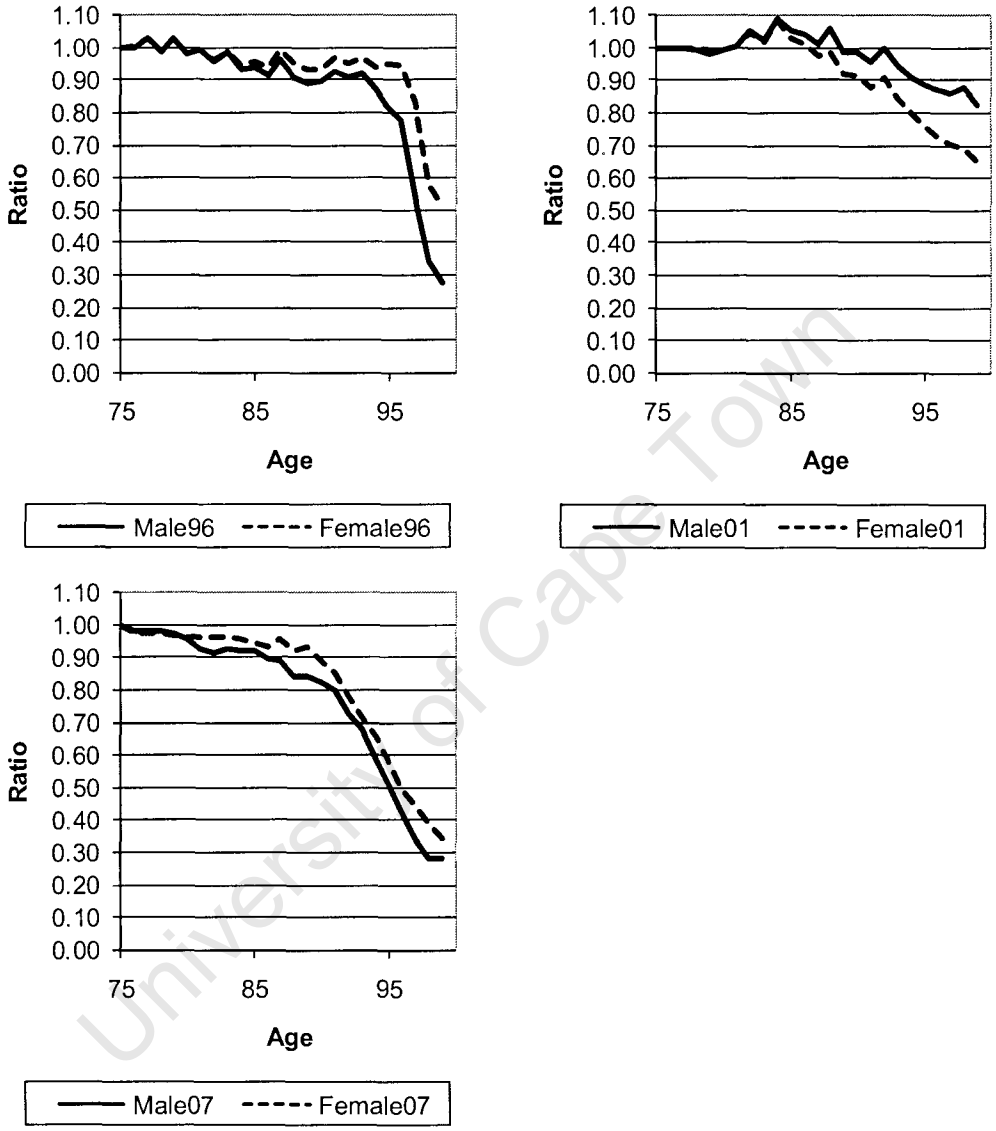
Figure 4.8 above shows some fluctuations in the ratios, but with a general reduction in the frequency of the fluctuation relative to that of 1996. Age reporting problems are visible at ages 97 and above with more people enumerated at age 99 relative to the estimates at this age. For this census the proportionate excess of females is higher than that of males.

Figure 4.9 Ratio (survey/estimates), by age and sex, 2007 survey

In Figure 4.9 above, one can see that the survey population and the estimates from death statistics are fairly close to each other from age 75 to 95. An excess of the survey population relative to the estimates is observed at ages 96, 97, 98 and, 100 and above (not shown on graph with ratios of 5.03 for males 4.10 for females). These results may suggest that there is age exaggeration at ages 96, 97 and 98, most probably from those aged between 91 and 96 in the case of males and from those aged between 89 and 95 in the case of females. The distortion at age 100 and above may be due to some people with unrealistically high ages above 110.

To be consistent with the synthetic extinct generations method, the series, $\hat{P}_{adj}(x+)/P(x+)$, was plotted against age as shown in Figure 4.10 below. Aggregation of the population above a particular age is done to eliminate some of the fluctuations observed in the analysis by single ages. A comparison of these ratios suggests that age exaggeration or over count of the population is increasing with age for both sexes as shown by a fall in the ratios.

Figure 4.10 Ratio ($\hat{P}_{adj}(x+)/P(x+)$), by age and sex, 1996 and 2001 censuses, and the 2007 survey



4.3 Data quality

The extent of bias in the data used is investigated in the sections to follow and used to draw conclusions on the implications of the bias to the final estimates of mortality.

4.3.1 Digit preference in reported deaths

The ratios q_x / q_{x+1} were computed for the years 1996 and 2001, and the results are shown in Table 4.3 below. One can see that in 1996, there was age heaping at death at ages 78, 82, 86, 92 and 94 for both sexes and also age heaping at death at ages 90 and 95 in males. Significant age heaping is observed in males but is less in females. There is no pattern in the ages at which heaping is observed and one can trace the years in which these individuals were born. Those who died at age 78 were probably born on average in 1918, those who died at age 82 were born on average in 1914, those who died at age 86 were born on average in 1910, those who died at age 92 were born on average in 1904 and those who died at age 94 were born on average in 1902. The same analysis can be done with the 2001 data and age heaping at the same years of birth as in 1996 were observed at ages 83 (1918), 87 (1914), 91 (1910) and 97 (1904). The arrows show the cohorts that were born in these years over time. This would suggest that people were probably making reference to the beginning and end of World War I (1914 and 1918) as their birth years. Bearing in mind that the majority of people in South Africa have identification documents and at the point that these were issued, the issuing offices would have needed to decide on a date of birth, such that a year of birth was allocated, and presumably this has determined the person's age since. Other years which show some year of birth preference are 1902, 1905, 1906 and 1910. The years 1905 and 1910 may have been preferred because they are easy to give and remembered as multiples of five. Similarly, the years 1902, 1904 and 1906 may have been preferred because they are even numbers. The effect of year of birth preference is probably more marked than shown, since as mentioned earlier that a person can have two ages last birthday within a year. The results presented are actually mixing two birth years and had it been possible to group by year of birth the change in pattern would have been more marked.

Table 4.3 Ratio of probability of death at age *x* to probability of death at age *x+1*, 1996 and 2001

	q_x/q_{x+1}			
	Male		Female	
Age (x)	1996	2001	1996	2001
75	0.92	0.96	0.94	0.88
76	0.94	0.93	0.93	0.99
77	0.92	0.96	0.88	0.95
78	**1.19**	0.93	**1.29**	0.90
79	0.74	0.91	0.70	0.83
80	0.87	0.92	0.90	0.98
81	1.00	1.02	0.91	0.93
82	**1.21**	0.83	**1.15**	0.83
83	0.82	**1.44**	0.77	**1.39**
84	0.87	0.73	0.87	0.70
85	0.92	0.79	0.94	0.88
86	**1.25**	0.93	**1.18**	0.94
87	0.76	**1.34**	0.80	**1.08**
88	0.96	0.77	0.87	0.82
89	0.85	0.90	0.95	0.95
90	**1.19**	0.92	0.96	0.84
91	0.91	**1.30**	0.98	**1.13**
92	**1.31**	0.89	**1.15**	0.96
93	0.86	0.93	0.89	0.88
94	**1.10**	1.00	**1.10**	0.94
95	**1.22**	**1.24**	0.90	1.00
96	0.87	0.84	0.96	0.92
97	0.50	**1.35**	0.72	**1.34**
98	0.81	0.74	0.66	0.91

Notes Normal font-acceptable quality (below 1.05)
Bold and Italics font-moderate age heaping (1.05-1.19)
Bold font-significant age heaping (1.20 and above)

4.3.2 Digit preference in census/survey population

An independent assessment of heaping at the same years of birth as observed in the reported deaths was done using the census/survey population and the results are shown in Table 4.4 below.

Table 4.4 Ratio of Whipple's Index derived from the census/survey population divided by the Whipple's Index derived from the ASSA population estimates.

	1996 census		2001 census		2007 survey	
Ratio of indexes	Male	Female	Male	Female	Male	Female
WI(0 or 5)	0.99	0.98	0.98	0.98	0.89	0.87
WI(1910, 1920, 1930)	1.36	1.38	1.20	1.33	1.14	1.27
WI(1914)	1.40	1.33	1.40	1.31	*	*
WI(1918)	1.38	1.43	1.31	1.37	1.17	1.21

Notes WI(0 or 5) is the Whipple's Index for an age with a terminal digit equal to 0 or 5.
WI(19xx) is the Whipple's Index for heaping at birth year 19xx.

* There are no data to compute the Whipple's Index for the standard population. This is because the population born in 1914 was on average aged 92 in 2007 and the age distribution of the chosen standard population for comparison ends at age 89.

One can see that none of the indexes derived to measure age digit preference are 5 per cent more than those derived from the standard population as reflected by the ratios of the Whipple's Index shown in Table 4.4. These results suggest that the census/survey population data by age show no signs of preferring the ages that end with a zero or a five. The ratios in Table 4.4 show that the index used to assess year of birth preference deviate significantly from the standard for the years of birth 1910 and 1920 in the case of the 1996 census data, and 1920 and 1930 in the case of the 2001 census and 2007 survey data. Using the recommended standard criteria of measuring age heaping from Table 2.1 in Chapter 2, one can see that the data for those born in 1910 and 1920 at the 1996 census were roughly accurate for both sexes, while at the 2001 census and 2007 survey; there was approximate age heaping for males and rough data for females who were born in 1920 and 1930. The Whipple's Index to assess year of birth preference at years 1914 and 1918 show that the data quality was generally rough as reflected by the ratios in Table 4.4 which deviate by more than 30 per cent in the case of the 1996 and 2001 census data.

4.3.3 Age overstatement at death

Table 4.5 below shows the results obtained in assessing age exaggeration in the reported deaths for males.

Table 4.5 Ratio of death ratios from reported deaths divided by death ratios from deaths in the ASSA model for males

	Year										
Ratios of ratios	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
D80+/D75+	1.13	1.05	1.05	1.03	1.05	1.05	1.04	1.04	1.03	1.03	1.06
D85+/D80+	1.93	1.68	1.41	1.32	1.20	1.11	1.04	1.07	1.05	1.06	1.08

The ratio of ratios in Table 4.5 shows that the reporting of age at old age is generally improving over time relative to the standard. The age-reporting of deaths up to age 80 is reasonably good as shown by an average deviation of 4 per cent excluding the 13 percent in 1996. One can see that in the years further in the past (1996 to 2000), the reported deaths at age 85 and above were proportionally more than what was expected from the ASSA model. The ratio of ratios for 1996 is almost double and from 2002 onwards, one can see that the ratio of ratios has almost stabilized at a deviation of around 6 per cent.

Table 4.6 below show the results obtained in assessing age exaggeration in the reported deaths for females.

Table 4.6 Ratio of death ratios from reported deaths divided by death ratios from deaths in the ASSA model for females

	Year										
Ratios of ratios	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
D80+/D75+	1.19	1.14	1.15	1.12	1.12	1.09	1.06	1.04	1.01	0.98	1.00
D85+/D80+	2.00	1.88	1.59	1.40	1.28	1.17	1.10	1.13	1.11	1.10	1.10

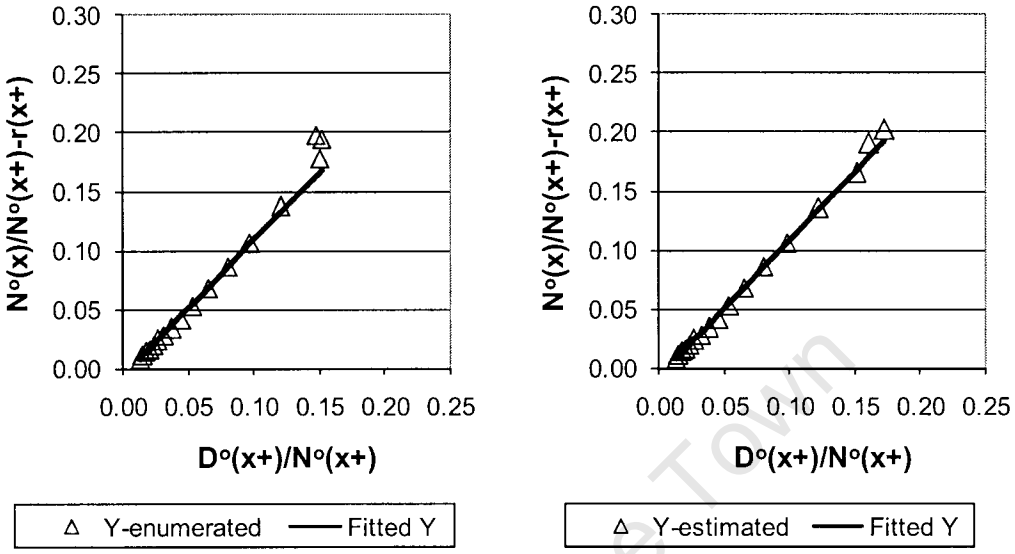
The ratio of ratios in Table 4.6 suggests the same conclusion as the results for males but with age exaggeration at death being more in the female population relative to the standard.

The next section analyzes the extent to which the deaths are reported and adjusts the deaths for completeness of reporting before estimating mortality. The methods used also helps to explain other quality characteristics of the data not analyzed so far.

4.4 The Generalized Growth Balance and Synthetic Extinct Generations methods

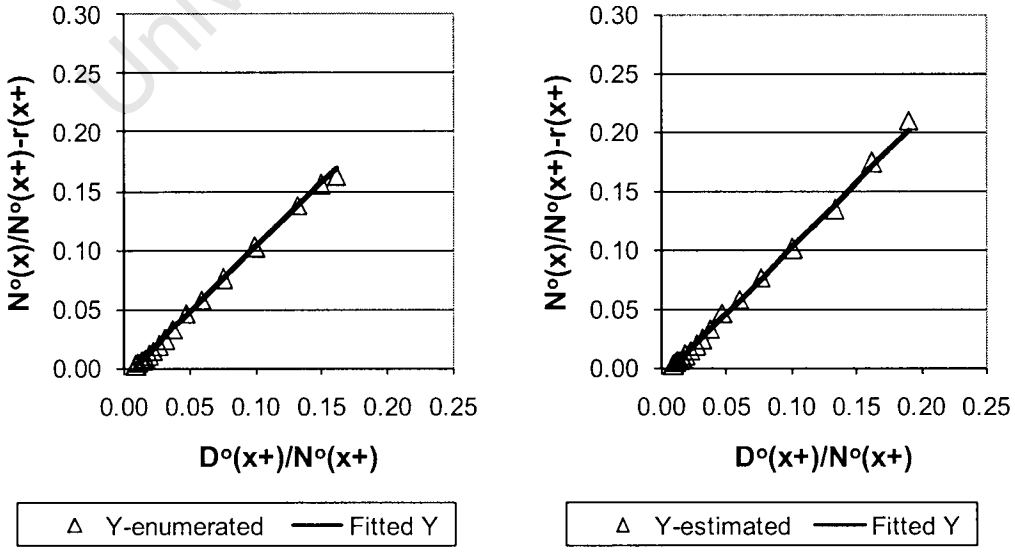
Figures 4.11 to 4.14 compare the application of the GGB method using the census/survey population with that using the estimated population derived from the registered deaths.

Figure 4.11 Application of the GGB method: Male population, 1996 and 2001 censuses



The completeness of reporting of male deaths estimated from the data in Figure 4.11 is 85.4 per cent relative to the census/survey population and 84.8 per cent relative to the population estimated from deaths. The old age population estimated from deaths produces points which are closer to a linear plot which suggests that the census/survey population data has errors such as age exaggeration in the population numbers or that the completeness of reporting of deaths by age is falling.

Figure 4.12 Application of the GGB method: Female population, 1996 and 2001 censuses



The completeness of reporting of female deaths estimated from the data in Figure 4.12 is 88.8 per cent relative to the census/survey population and 87.9 per cent relative to the population estimated from deaths. The data for females does not deviate much from linearity but the population numbers estimated from deaths produce a closer fit. Data for females for the period between these two censuses appears to be of better quality than that for males.

The completeness of reporting of male deaths estimated from the data in Figure 4.13 below is 88.6 per cent relative to the census/survey population and 87.0 per cent relative to the population estimated from deaths. Estimating the old age population from deaths gives a better fit except for that last point. Since the population numbers have been corrected, the deviating point at age 95 is most probably a result of a lower completeness in the reporting of deaths for ages 95 and above, and not an over correction of the population numbers.

Figure 4.13 Application of the GGB method: Male population, 2001 census and 2007 survey

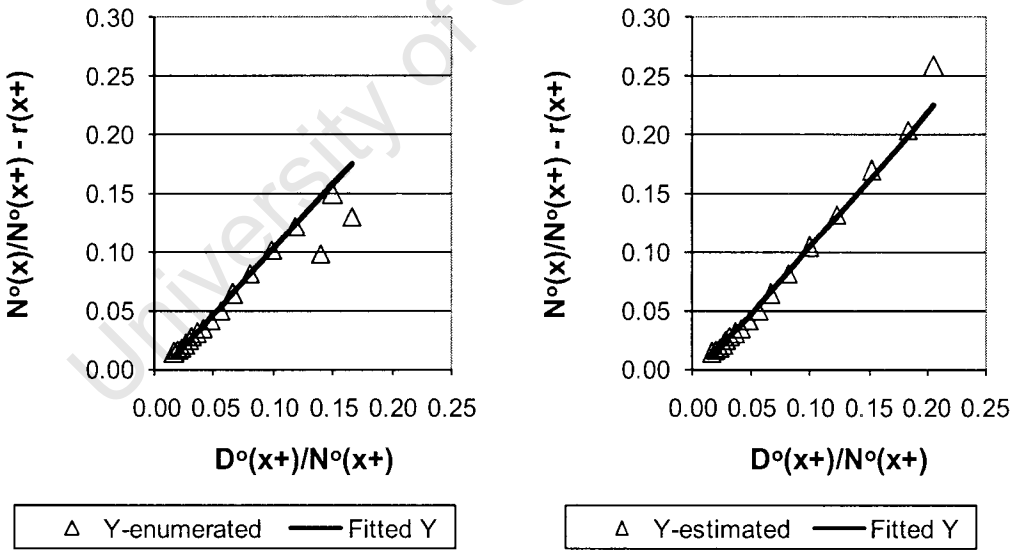
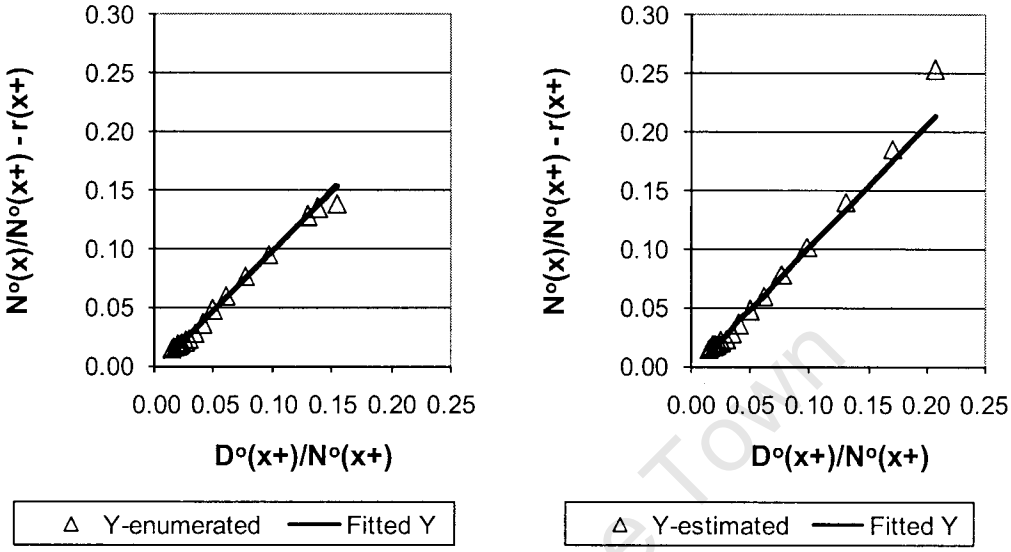


Figure 4.14 Application of the GGB method: Female population, 2001 census and 2007 survey



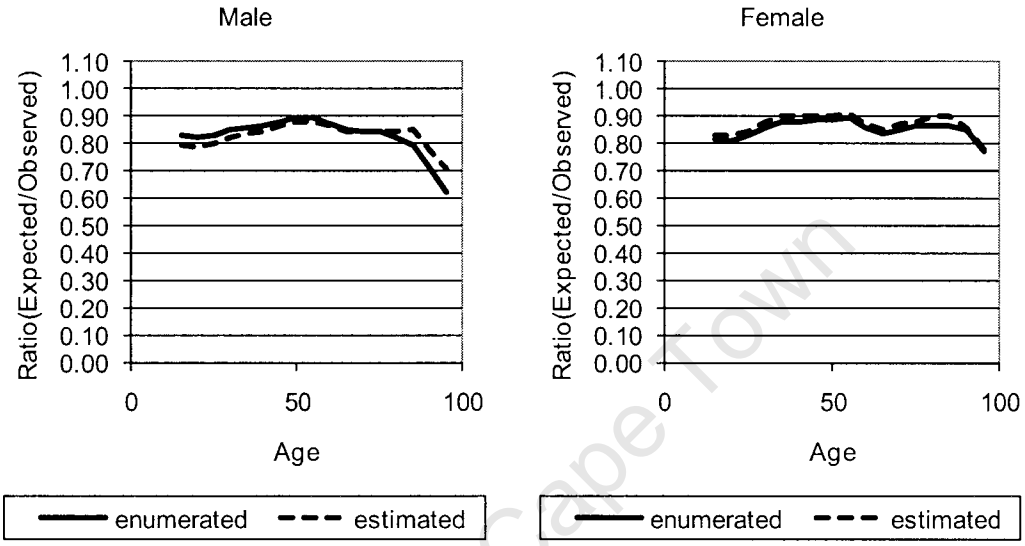
The completeness of reporting of female deaths estimated from the data in Figure 4.14 is 98.1 per cent relative to the census/survey population and 94.5 per cent relative to the population numbers estimated from deaths. Estimation of the old age female population reduced the estimate of the completeness of death reporting. This suggests that the census/survey population numbers at old age have problems and the distortion in the numbers leads to over estimating the completeness of the reporting of deaths. The point deviating from linearity at the last age group is probably again the result of a fall in the completeness of the reporting of deaths.

The results from the application of the GGB method also suggest that the 2001 census coverage was better than the 1996 census coverage and the 2007 survey coverage was better than the 2001 census coverage, as shown by negative intercepts. This suggests that there has been an improvement in the enumeration process in each subsequent census or large scale population count.

Life expectancies at age 100 estimated from the mortality rate of that age group after correcting for incompleteness are 1.09 and 1.23 for males and females respectively for the period between the 1996 and 2001 censuses. For the period between the 2001 census and the 2007 survey, the life expectancies at age 100 are virtually the same at 1.09 and 1.24 for males and females respectively.

Results from the SEG+delta method using both the census/survey and the population estimated from deaths are as shown in Figure 4.15 and Figure 4.16 below.

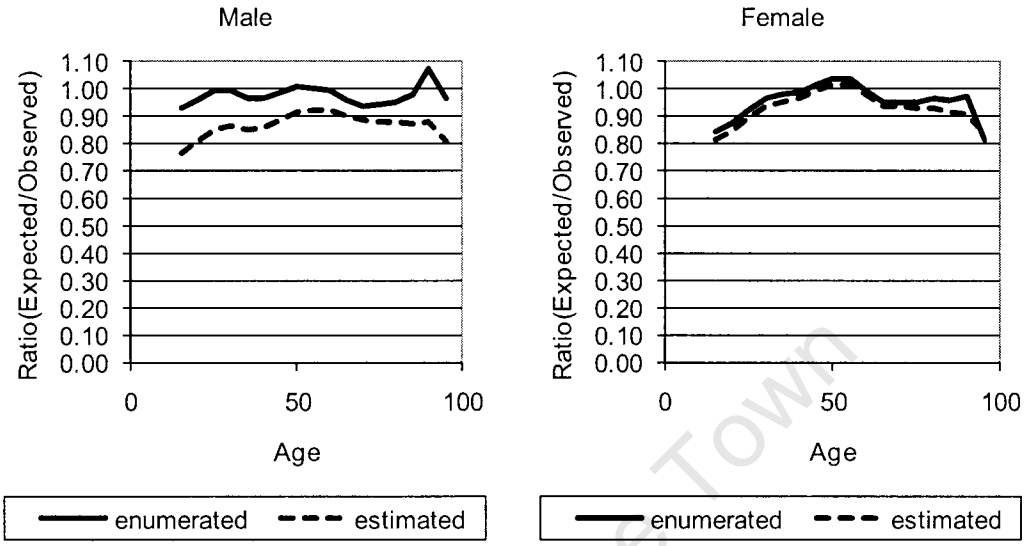
Figure 4.15 Application of the SEG+delta method: Male and female population, 1996 and 2001 censuses



Application of the SEG+delta method using the census/survey population yields results that suggest that either there is age exaggeration in the census enumeration or the completeness of the reporting of deaths is declining with age at the oldest age groups, or both. The fall in the completeness of reporting of deaths and/or age exaggeration is observed from age 90 in both males and females over the period between the 1996 and 2001 census as shown in Figure 4.15 above. The extent that use of the population estimates still does not correct for the fall in estimated completeness suggests that most of this effect is due to a fall in the completeness of reporting of deaths at the oldest ages.

In the period between the 2001 census and the 2007 survey, results from the SEG method suggests that the completeness of reporting of deaths is overstated with some points derived from the SEG method showing completeness of more than 100 percent as shown in Figure 4.16 below.

Figure 4.16 Application of the SEG+delta method: Male and female population, 2001 census and 2007 survey



Application of the SEG+delta and the GGB methods to the data with new estimates of the population points to new conclusions. The unusual patterns observed in the census/survey population data are now to some extent explained by differential enumeration of the old age population by age over time. It has been observed that this error in population enumeration led to lower growth rates of the population between the 1996 and 2001 census, and higher growth rates between the 2001 census and the 2007 survey. The effect of the error in census enumeration is more visible in the plot for males over the period between the 2001 census and the 2007 survey as shown in Figure 4.13 and Figure 4.16 above. From this finding in particular it would appear that the 2001 census may have been under enumerated at old ages.

The preceding discussion was on the patterns observed only and the estimates of the completeness of reporting of deaths used to adjust the deaths were also derived. Table 4.7 below shows the numerical estimates of the completeness of reporting of deaths relative to both the census/survey population and the old age population estimated from deaths.

Table 4.7 Estimates of the completeness of reporting of deaths derived from the SEG+delta method

	Sex	1996-2001	2001-2007
k_o (SEG+delta and census/survey population)	Male	0.846	0.970
	Female	0.862	0.964
k_a (SEG+delta and re-estimated population)	Male	0.845	0.878
	Female	0.881	0.939

The estimates of the completeness of reporting of deaths derived using the population estimated from deaths appear to be reasonable because there is greater consistency between the estimates of the two periods, and there is a reduction of the estimates from implausibly high (in comparison with estimates from other sources) completeness in the most recent period.

The completeness of reporting of deaths for females is generally higher than that for males regardless of the data used. The completeness of reporting of deaths is generally improving with time as reflected by the estimates in Table 4.7 above where the completeness of reporting of deaths estimates over the period between the 1996 and 2001 censuses are less than the death reporting completeness over the period between the 2001 census and the 2007 survey. The plot of the series, $\hat{N}(x)/N(x)$, in Figure 4.15 and Figure 4.16 from the SEG+delta method suggest that the completeness of reporting of deaths is falling with age at the advanced ages even after re-estimating the old age population.

4.5 Comparison of the extinct generations method and the synthetic extinct generations method

One striking observation is that the estimates of the completeness of reporting of deaths derived from the method of extinct generations for the period between the 1996 census and until the population at the census date aged 75 last birthday is almost extinct at age 110 as shown in Table 4.1 are higher than the estimates of death reporting completeness between the 1996 and 2001 censuses derived using the SEG+delta method as shown in Table 4.7. Part of the explanation for this difference is most probably the difference in the fundamental definitions of the two methods and also the general improvement in death reporting completeness over time.

By definition, the method of extinct generations follows a real cohort from the census date until its extinction and then the deaths are aggregated to get the size of the

cohort at the census date. The SEG+delta method assumes that deaths will grow at the growth rate being experienced by the population and then those deaths are aggregated to estimate the size of the population at the middle of the time interval between the two censuses. The growth rates that are assumed are derived from two censuses and therefore an over enumeration of the first census or an under enumeration of the second census will derive low growth rates and hence lead to an estimate of too few deaths relative to the population distribution. As a result, the SEG+delta method is giving low estimates of completeness of reporting of deaths relative to the estimates of the population.

These results suggest that the 2001 census was under enumerated resulting in a shrinking of the growth rates between the 1996 and the 2001 censuses, and an increase in the growth rates between the 2001 census and 2007 survey.

4.6 Mortality rates

One of our main objectives is to determine mortality rates and the results are presented in Table 4.8 below.

Table 4.8 Mortality rates derived from the census/survey population and estimated population by sex and age

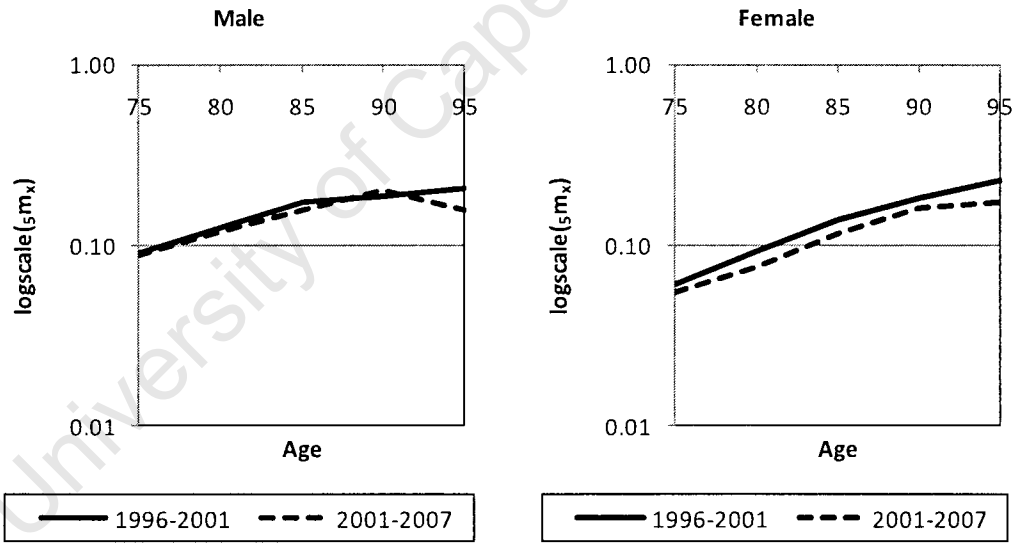
Period		${}_5m_x$-census/survey		${}_5m_x$-estimated		Per cent deviation	
1996-2001	Age(x)	Male	Female	Male	Female	Male	Female
	75	0.09266	0.06087	0.09140	0.06010	1.4	1.3
	80	0.12322	0.09214	0.12499	0.09363	-1.4	-1.6
	85	0.17778	0.14439	0.17290	0.13897	2.8	3.9
	90	0.18253	0.17202	0.18534	0.17990	-1.5	-4.4
	95	0.18170	0.20317	0.20726	0.22534	-12.3	-9.8
	100 and above	0.16678	0.15771	0.23711	0.25096	-29.7	-37.2
2001-2007	75	0.08219	0.05572	0.08935	0.05454	-8.0	2.2
	80	0.10372	0.07418	0.11910	0.07583	-12.9	-2.2
	85	0.14632	0.11959	0.15812	0.11476	-7.5	4.2
	90	0.19100	0.16995	0.20028	0.16173	-4.6	5.1
	95	0.11429	0.13766	0.15701	0.17314	-27.2	-20.5
	100 and above	0.20771	0.15640	0.56592	0.38377	-63.3	-59.2

Table 4.8 above shows the mortality rates derived using both the census/survey population and the population estimated from deaths. One can see a general under estimation of old age mortality from the census/survey population data relative to the population data estimated from deaths as reflected by the negative per cent deviation. The level of under estimation increases with age and this may be the result of age exaggeration in the population. Mortality under estimation is more in males than females. The fact that

the completeness of reporting of deaths for males is less than that for females may also contribute to the differential in mortality under estimation by sex. However, there are a few age groups where mortality derived from the census/survey population data suggests that mortality was over estimated. Possible reasons for these results are that since these age groups correspond to some of the age groups where preference of year of birth in the reported deaths was observed, this bias will cause heaping of deaths in those age groups hence increasing mortality rates.

Under estimation of mortality is worse in the period between the 2001 census and the 2007 survey. Mortality estimates of males at old age are generally higher than that of females. Analysis of the above estimates of mortality for the two time periods is as shown in Figure 4.17 below.

Figure 4.17 Mortality estimates for the two time periods by age and sex



Results from Figure 4.17 above suggests that mortality at old age may be falling as shown by the estimates between the 1996 and 2001 censuses which are generally less than those for the period between the 2001 census and the 2007 survey for females. For males however, the difference is not as marked.

Figures 4.18 and 4.19 below compare the new estimates of mortality with estimates from other sources. Dorrington, Moultrie and Timaeus (2004) used the 1996 and 2001 censuses to estimate mortality over the period between the two censuses and their life tables extend up to age 90. Other mortality estimates are derived from the UNPD and

USCB population estimates using survival probabilities and on the assumption that there is no migration at the advanced ages. The census survival ratio method (United Nations, 2002), was applied on the published population numbers by age and sex from the UNPD and USCB to estimate mortality rates. The population estimates for the years 1995 and 2000 from the UNPD were used to estimate mortality rates for the first period and the population estimates for the years 2000 and 2005 were used to estimate mortality for the recent period for comparison. Figures 4.18 and 4.19 below show the plots of the mortality estimates, one for unadjusted population numbers (Unadj), one for the adjusted old age population (Adj), one (DMT) from estimates by Dorrington, Moultrie and Timaeus (2004), estimates derived from UNPD and USCB population projections.

Figure 4.18 Mortality estimates between the 1996 and 2001 censuses by age group

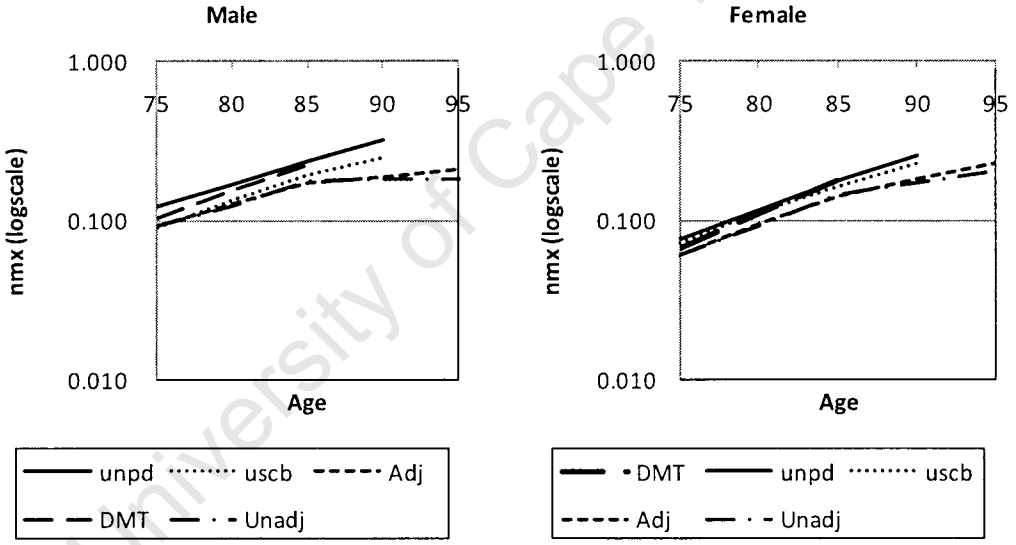
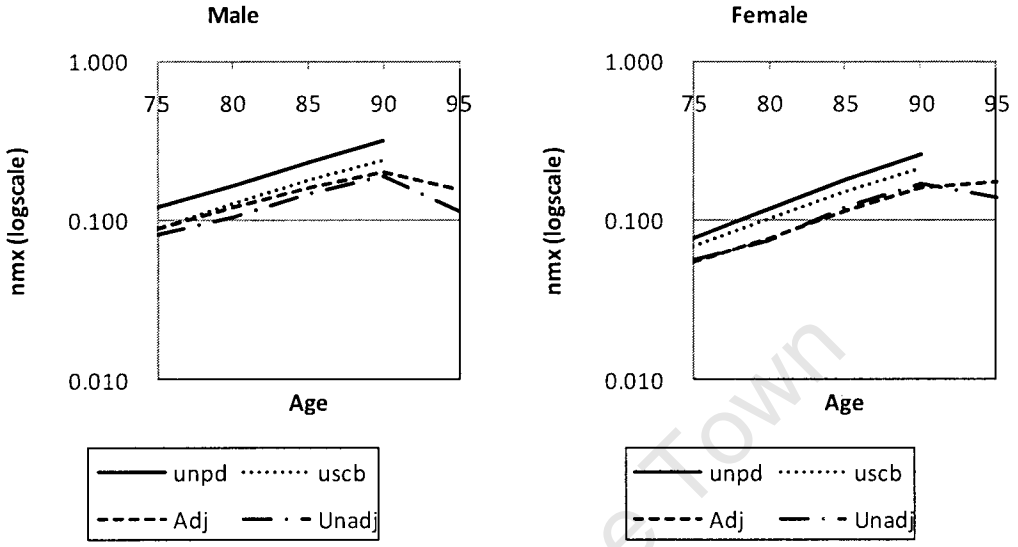


Figure 4.19 Mortality estimates between the 2001 census and 2007 survey by age group



One can see that the census/survey population numbers lead to lower estimates of mortality relative to the re-estimated old age population numbers at ages above 90 in the case of the period between the two censuses and at ages above 95 in the case of the most recent period. The change in trend in the estimated mortality rates at these ages is probably a result of the fall in the completeness of the reporting of deaths. Estimates derived by Dorrington, Moultrie and Timaeus (2004) are higher than the adjusted estimates as shown in Figure 4.18 and this may be partly due to the fact that they factored in migration and the rates are graduated using Brass's logit relation (Brass, 1968) and the General Standard. The mortality rates inferred from the UNPD and USCB population projections are all generally higher than the adjusted estimates for both periods.

5 Discussion and conclusion

5.1 Introduction

The aims of this study were to estimate the South African oldest old age mortality and in the process re-estimate numbers of people in the population at the advanced ages. The research also investigated errors that are common in the demographic data of the old age population. This chapter discusses the results obtained and reflects on the extent to which they meet the objectives of the research. This chapter is organized as follows: Section 5.2 looks at the re-estimated population numbers and compares them with the census/survey population, section 5.3 considers the data quality of both the census/survey population numbers and the reported deaths, and section 5.4 discusses results from the death distribution methods used to adjust for under reporting of deaths. The next section, section 5.5 discusses the mortality rates calculated and compares them with other estimates. Section 5.6 gives the limitations of the study followed by section 5.6 which gives the main conclusions reached from this study and recommends further research.

5.2 Population estimates

The South African population numbers in general and the old age population in particular are obtained from censuses and other nationally representative large scale surveys. Some organizations have designed population projection models so that they can project the population numbers at some point in time in the future, for policy making. Among the organizations are the Actuarial Society of South Africa, USCB and UNPD. A comparison of the census/survey population numbers and the estimates from the different sources shows wide differences of the numbers at old age. Due to the irregularities in the data of this old age population numbers by age and among other practicality concerns, researchers and other users have resorted to aggregating the population estimates at some high enough ages.

This research used indirect methods to re-estimate the oldest old age population numbers (aged 75 up to 100 and above) and compared these estimates with the census/survey numbers. The method of extinct generations which uses deaths arising from a cohort to re-estimate the size of the cohort at some point in time has been observed to produce superior estimates of the population at old age relative to the census/survey numbers and was used in this current research. Results from the method of extinct

generations suggest that there is no systematic difference between the population numbers estimated from deaths and the census/survey population numbers except at the oldest age groups. The results obtained suggest that there is some element of over counting the population at the highest age groups from age 95 and above. The over count of the population at these ages leads to under estimating mortality. The method of extinct generations also produced a smooth distribution of the population by age unlike the census/survey population.

Since the 2007 survey did not enumerate the whole population, the national population estimates from the survey were obtained by adjusting the realized sample such that the national estimates are consistent with previous census/survey data. The adjustment done to the sample might have had a smoothing effect on the national estimates. In addition to this, the current research estimated the 2007 survey population by extrapolating the census estimates (census estimates from the method of extinct generations) forward. The fact that both the published national estimates of the population in the 2007 survey and those estimated by this research were a result of adjustments which might have had a smoothing effect may explain why the two estimates fit well as shown in Figure 4.6.

A comparison of the census/survey population and estimates from the organizations mentioned above suggests that the excess of the census/survey population relative to the estimates increases with age. The excess of old people is high in the comparison of the estimates from USCB and UNPD. This finding suggests that USCB and UNPD are over estimating the South African old age mortality and as a result their projections produce lower estimates of the population.

The data on reported deaths used to re-estimate the old age populations are not free from errors and the following section evaluates the extent of the bias resulting from these errors.

5.3 Data quality

This research investigated the errors that are common in the data of the old age population. The most common errors are age overstatement and age digit preference in both the reported deaths and the census/survey population, and under reporting of deaths.

The methods used to assess the quality of the data on reported deaths depend on a standard population whose mortality pattern is similar to that of the population under

study. After having established that the censuses are wrong at the extreme ages, it was suggested that ASSA estimates are accurate enough to use as the standard population since the ASSA estimates of the population are consistent with the census/survey population numbers up to age 80. This means that the conclusions reached are made relative to ASSA estimates. Even if ASSA estimates are incorrect, conclusions reached on year of birth preference will not change as they have also been reached using a method which does not depend on ASSA.

The difference between the distribution of the reported deaths at old age and ASSA estimates is acceptably small. The marginal age overstatement in the age of reported deaths is improving relative to ASSA over time. The aim of investigating age overstatement in the reported deaths is to see its effect on the re-estimated population numbers. A simulation done to investigate this effect showed that age overstatement in reported deaths has only a marginal effect on the re-estimated population. This is because the age exaggeration in reported deaths notable at a particular point in time nearly cancels out when summed across different periods for the different cohorts. From this, we can conclude that the difference between estimates of the population derived using the method of extinct generations and the census/survey population are real and not due to age overstatement in the reported deaths.

The next data quality concern in reported deaths is age digit preference. The research observed that there was a tendency by the population to prefer certain years of birth rather than certain ages. This type of error in data has to do with periods of mass registration where individuals end up being allocated birth years connected to certain historical periods or years with certain terminal digits. The years of birth that were preferred are 1910, 1914, 1918, 1920 and 1930. From these results, the years 1914 and 1918 may have something to do with World War I and the years 1910, 1920 and 1930 have terminal digit zero or are multiples of 10. The research observed that the same birth year preference is also present in the census/survey population.

From this finding, we can conclude that some differences between the estimates of the population derived using the method of extinct generations and the census/survey population are due to heaping that occurred when birth dates were allocated for registration purposes and are now entrenched. The heaping observed at certain years of birth cannot be attributed to random fluctuations since this is observed in cohorts over time.

Age digit preference and birth year preference in the reported deaths and the census/survey population numbers cause fluctuations in mortality rates. Re-estimating the population numbers smoothens the heaping in the population, but the errors still remain in the reported deaths.

The population born in the years that were preferred will be extinguished and therefore it is expected that data for the whole population in general and the oldest old in particular will improve. This is so because of the new system of continuous registration of vital statistics as opposed mass registration.

Under reporting of deaths is another possible cause of bias in the estimation of mortality and is discussed in the following section.

5.4 Completeness of reporting of deaths

As mentioned previously, deaths are usually under reported and the population numbers are most probably under enumerated. As a result, the data on reported deaths and population numbers cannot be used to estimate mortality at face value. The GGB and SEG+delta methods were used to adjust for under reporting of deaths before estimating mortality rates. The difference between the estimates of the completeness of reporting of deaths from the GGB and SEG+delta methods are small (less than 1 per cent) to have much impact on the estimates of mortality and in this research, the estimates from the SEG+delta method are used.

The estimates of the completeness of reporting of deaths derived from this research are consistent with those from other studies. The estimates of the completeness of reporting of deaths over the period between the 1996 and 2001 censuses derived from this research are 84.5 per cent and 88.1 per cent for males and females respectively. Similar estimates derived by Dorrington, Moultrie and Timaeus (2004) over the same period are 83.5 per cent and 86.7 per cent for males and females respectively. The small differences may be due to the fact that Dorrington, Moultrie and Timaeus (2004) factored in migration in applying the methods, they used lower open intervals and did not re-estimate the population at the old ages. This finding shows that the adjustment made on the population numbers has little impact on the completeness of reporting of deaths, but has a significant impact on the mortality rates at the advanced ages.

The completeness of reporting of deaths at old age is not a problem as concluded by Anderson and Phillips (2006). The problem is the method they used and they produced implausibly high estimates of the completeness of reporting of deaths.

The two methods also show certain patterns in the reported deaths and the census/survey population. Application of both methods on the census/survey population show that there is either age exaggeration in the population data or the completeness of reporting of deaths is falling with age. After adjusting the old age population, the SEG+delta method and the GGB method concluded that the 2001 census appears to have under enumerated the old age population. The SEG+delta and GGB methods show that the completeness of reporting of deaths is falling with age at the advanced ages even after adjusting the population numbers.

The SEG+delta method was applied to the data with adjusted old age population and this old age population is estimated from deaths whose completeness of reporting is falling with age. This would mean that the re-estimated oldest old age population is under estimated as well. On estimating completeness of the reporting of deaths, there is a double effect that the deaths at the oldest old age are under reported and the oldest old age population is under estimated as well.

5.5 Mortality rates

The main objective of this research is to produce better estimates of mortality at old age. The old age population estimates (aged 75 up to 100 and above) derived using the method of extinct generations and the census/survey population aged between 15 and 74, together with the reported deaths in the corresponding age groups were used with the SEG+delta method to produce estimates of mortality. The SEG+delta method was used to adjust for under reporting of deaths and it was chosen because it is a period analogue of the method of extinct generations which traces cohorts. Use of both the method of extinct generations and the SEG+delta method to re-estimate the population helped to evaluate the assumptions underlying each method. The SEG+delta method is very sensitive and depends heavily on growth rates whereas the method of extinct generations depends on survival ratios for non extinct cohorts.

The mortality estimates derived for the age groups between 75 and 90 are close to estimates from other sources. Since the estimates derived need graduation before being

used it was decided to assume final estimates of mortality from one of the UNPD, USCB and mortality estimates by Dorrington, Moultrie and Timaeus (2004) and the mortality rates are smoothed in one way or the other.

Since these three estimates of mortality are close to each other, it is reasonable to recommend the estimates derived using the enumerated population numbers, that is estimates from Dorrington, Moultrie and Timaeus (2004). The problem with these estimates is that they are limited to age groups 85 to 89 and are available only for the whole period between the two censuses. However, since the UNPD mortality estimates fit well with estimates from Dorrington, Moultrie and Timaeus (2004) for age groups 75 to 89 for both sexes, it is recommended that the UNPD mortality estimates be used for the advanced ages.

The research also observed that even if the estimates are consistent, the accuracy of mortality estimates falls with age across the advanced ages and this decrease in precision is higher in males than in females. The decrease in precision may be attributed to falling completeness of reporting of deaths at these ages.

5.6 Limitations of the study

Certain reported ages need to be treated with caution as there are an implausibly number of people aged above 110. The super-centenarians are also observed in the data for reported deaths. However, the number of super-centenarians may not be a significant issue since their numbers are small relative to the numbers in the age groups with which we are concerned.

The method of extinct generations relies on accurate vital statistics by age. The death statistics used to estimate the populations were obtained from Statistics South Africa and which acknowledges that some deaths may yet be reported and therefore the statistics are not completely accurate. The research adjusted the estimates to cover for the deaths which are still to be reported, and it may be debatable which data has more errors between the census/survey population and the reported deaths.

The method of extinct generations was developed to estimate the old age population which is closed to migration. The research assumed that migration from age 75 and above is negligible. The extent of international migration relative to the population at

ages as low as 75 could not be verified because there are no accurate estimates of migration (Dorrington, Moultrie and Timaeus, 2004).

The research could not be extended to an analysis by race because there was a significant proportion (about 25 per cent) of the population with no population group on the death certificates and in the case of all except the African and possibly the White population the numbers are small and probably more vulnerable to the various errors.

The research tried to take as much care as possible to ensure that some of the limitations were adjusted for. The verification of the extent of migration at ages as low as 75 by single ages is one of the limitations beyond which the research could not go.

5.7 Conclusions and recommendations

From the results above, the conclusion must be that the mortality rates derived after re-estimating the oldest old age population and after allowing for the fall in the completeness of reporting of deaths are lower but not significantly different from the mortality rates inferred from the UNPD and USCB population projections, and mortality estimates by Dorrington, Moultrie and Timaeus (2004) up to at least age 90. The current research recommends mortality estimates from the UNPD since they match estimates derived from the enumerated population almost closely. However, the indirect methods derived better estimates of the oldest old age population relative to the census/survey population numbers.

The research has identified a number of areas where further research might be needed. It is recommended that the impact of using more accurate estimates of age (year of birth) be assessed. Analysis by population group/race is also recommended to see the differentials in mortality at old age. Further studies could be done to assess the impact of the assumptions of linearity in the estimation of the population from deaths using the synthetic extinct generations method.

6 References

Actuarial Society of South Africa. 2005. *Aids and Demographic Model 2003.* Cape Town: Actuarial Society of South Africa. www.actuarialsociety.org.za.

Anderson, B. A. and Phillips, H. E. 2006. *Adult mortality (age 15-64) based on death notification data in South Africa: 1997-2004.* Pretoria: Statistics South Africa, 2006

Andreev, K. F. 2004. "A Method for Estimating Size of Population Aged 90 and over with Application to the 2000 U.S. Census Data", *Demographic Research* **11**(9):235-262.

Bennett, N. G. and Horiuchi, S. 1981. "Estimating the Completeness of Death Registration in a Closed Population", *Population Index* **47**(2):207-221.

Bennett, N. G. and Horiuchi, S. 1984. "Mortality Estimation from Registered Deaths in Less Developed Countries", *Demography* **21**(2):217-233.

Bourbeau, R. and Lebel, A. 2000. "Mortality Statistics for the Oldest-Old: An Evaluation of Canadian Data", *Demographic Research* **2**(2)

Brass, W. 1968. *The Demography of Tropical Africa.* Princeton, New Jersey: Princeton University Press.

Brass, W. 1975. *Methods for Estimating Fertility and Mortality from Limited and Defective Data* Chapel Hill North Carolina: Carolina Population Centre.

Coale, A. J., Demeny, P. and Vaughan, B. 1983. *Regional Model Life Tables and Stable Populations.* New York, N.Y./London England Academic Press.

Dorrington, R. E., Moultrie, T. A. and Timaeus, I. M. 2004. *Estimation of mortality using the South African census 2001 data.* Cape Town: Centre for Actuarial Research: University of Cape Town.

Dorrington, R. E. and Timaeus, I. M. 2008. "Death Distribution Methods for Estimating Adult Mortality: Sensitivity. Analysis with Simulated Data Errors, Revisited," Paper presented at Paper presented at the 73rd Annual Meeting of the Population Association of America. New Orleans, Louisiana, United States, 17-19 April 2008.

Gomes, M. M. F. and Turra, C. M. 2009. "The number of centenarians in Brazil: Indirect estimates based on death certificates", *Demographic Research* **20**(20):495-502.

Hill, K. H. 1987. "Estimating census and death registration completeness", *Asian and Pacific Population Forum* **1**(3):8-13.

Hill, K. H. and Choi, Y. 2004. "Death Distribution Methods for Estimating Adult Mortality: Sensitivity Analysis with Simulated Data Errors", Paper presented at Adult Mortality in Developing Countries Workshop., Paper presented at. The Marconi Center, Marin County, California, July 8th to 11th 2004.

Himes, C. L., Preston, S. H. and Condran, G. A. 1994. "A Relational Model of Mortality at Older Ages in Low Mortality Countries", *Population Studies* **48**(2):269-291.

Horiuchi, S. and Coale, A. J. 1982. "A Simple Equation for Estimating the Expectation of Life at Old Ages", *Population Studies* **36**(2):317-326.

Jdanov, D. A., Jasilionis, D., Soroko, E. L. *et al.* 2008. *Beyond the Kannisto-Thatcher Database on Old Age Mortality: An Assessment of Data Quality at Advanced Ages. MPIDR Working Paper WP 2008-013.* Demographic Research.

Jdanov, D. A., Scholz, R. D. and Shkolnikov, V. M. 2005. "Official population statistics and the Human Mortality Database estimates of populations aged 80+ in Germany and nine other European countries", *Demographic Research* **13**(14):335-362.

Kannisto, V. 1988. "On the Survival of Centenarians and the Span of Life", *Population Studies* **42**(3):389-406.

Makino, K. 2004. "Social Security Policy Reform in Post-Apartheid South Africa A Focus on the Basic Income Grant," Paper presented at 19th IPSA World Congress. Durban, July 2003.

Mesle, F., Vallin, J. and Andreyev, Z. 2002. "Improving the Accuracy of Life Tables for the Oldest Old: The Case of France", *Population* **57**(4/5):601-629.

National Research Council. 2006. *Aging in Sub-Saharan Africa: Recommendations for Furthering Research.* Barney Cohen and Jane Menken, Eds. Committee on Population, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press

Omran, A. R. 2005. "The epidemiological transition: A theory of the epidemiology of population change", *The Milbank Quarterly* **83**(4):731–757.

Preston, S. H., Coale, A. J., Trussell, J. *et al.* 1980. "Estimating the completeness of reporting of adult deaths in populations that are approximately stable", *Population Index* **46**(2):179-202.

Preston, S. H., Elo, I. T. and Stewart, Q. 1999. "Effects of Age Misreporting on Mortality Estimates at Older Ages", *Population Studies* **53**(2):165-177.

Preston, S. H., Heuveline, P. and Guillot, M. 2001. *Demography: Measuring and Modelling Population Processes.* Oxford: Blackwell.

Rosenwaike, I. 1968. "On Measuring the Extreme Aged in the Population", *Journal of the American Statistical Association* **63**(321):29-40.

Rosenwaike, I. 1979. "A New Evaluation of United States Census Data on the Extreme Aged", *Demography* **16**(2):279-288.

Rosenwaike, I. 1981. "A Note on New Estimates of the Mortality of the Extreme Aged", *Demography* **18**(2):257-266.

Statistics South Africa. 1998. *The people of South Africa : population census 1996 , calculating the undercount in Census '96 /Statistics South Africa.* Report No 03-01-18 (1996). Pretoria: Statistics South Africa, 1998.

Statistics South Africa. 2001. *Recorded deaths, 1996.* Stats SA Report No. 03-09-01. Pretoria: Statistics South Africa, 2000.

Statistics South Africa. 2004. *Census 2001: Post-enumeration survey: Results and methodology.* Report No. 03-02-17(2001). Pretoria: Statistics South Africa, 2004.

Statistics South Africa. 2008a. *Community Survey 2007: Unit records (Metadata).* Report No. 03-01-21 (2007). Pretoria: Statistics South Africa, 2008.

Statistics South Africa. 2008b. *Mortality and causes of death in South Africa, 2006: Findings from death notification.* Pretoria Statistics South Africa, 2008.

Statistics South Africa. 2009. *Interactive data*: http://www.statssa.gov.za/ Accessed: 4 February 2009.

Thatcher, R. 1992. "Trends in Numbers and Mortality at High Ages in England and Wales", *Population Studies* **46**(3):411-426.

Thatcher, R., Kannisto, V. and Andreev, K. 2002. "The Survivor Ratio Method for Estimating Numbers at High Ages", *Demographic Research* **6**(1):1-18.

United Nations. 1955. *Methods of Appraisal of Quality of Basic Data for Population Estimates.* Population Studies. Manual II, Series A, Population Studies No. 23. New York: United Nations:

United Nations. 2002. *Methods for estimating adult mortality.* ESA/P/WP.175.New York: United Nations Population Division.

United Nations. 2005. *Estimating and projecting national HIV/AIDS epidemics.* Working group on Global HIV/AIDS and STI Surveillance. UNAIDS. Geneva, Switzerland:

United Nations (2007) World Population Prospects: The 2006 Revision Population Database. New York: Department of Economic and Social Affairs.http://esa.un.org/unpp. Accessed: 4 February 2009.

United Nations Department of Economic and Social Affairs. 2007. *World Population Ageing 2007*. New York: United Nations.

US Census Bureau (2005) International Database. Washington DC: International Programs Center. http:// www.census.gov/ipc/www/idb/country/sfportal.html. Accessed: 4 February 2009.

Vincent, P. 1951. "La Mortalité des Vieillards", *Population* **6**(2):181-204.

Willcox, D. C., Willcox, B. J., He, Q. *et al.* 2008. "They Really Are That Old: A Validation Study of Centenarian Prevalence in Okinawa", *Journal of Gerontology: Biological Sciences* **63A**(4):338-349.

Wilmoth, J. R. 1995. "Are Mortality Rates Falling at Extremely High Ages: An Investigation Based on a Model Proposed by Coale and Kisker", *Population Studies* **49**(2):281-295.

Yi, Z. and Vaupel, J. W. 2003. "Oldest-Old Mortality in China", *Demographic Research* **8**(7):215-244.