An assessment of the feasibility of using administrative data in producing mid-year population estimates for South Africa

Mbongeni C. Hlabano
University of Cape Town

Thesis submitted to the Faculty of Commerce in partial fulfilment of the Degree of Master of Philosophy in Demography
University of Cape Town

November 2015

PLAGIARISM DECLARATION

This research is my original work, produced with supervisory assistance from my supervisor. I have used the Harvard convention for citation and referencing. Each contribution to this dissertation from the works of other people has been acknowledged, cited and referenced. In addition, this dissertation has not been submitted for any academic or examination purposes to any other university.

Signature:
Date: 27/11/2015

ABSTRACT

The production of mid-year population estimates is an important undertaking which informs various stakeholders in policy formation and decision making. For instance, national governments use mid-year estimates to allocate seats in parliament to various constituents and public health sectors use them to monitor and improve service delivery. Mid-year population estimates undoubtedly serve very important purposes that affect lives of many people. As such, national statistical offices in various countries are given the mandate to produce annual mid-year population estimates.

Statistics South Africa (Stats SA) assumes the function of producing and publishing official mid-year estimates of the population in South Africa. Stats SA produces its mid-year estimates using DemProj, population projection software which is part of the SPECTRUM suite of policy models developed by the Futures Institute. However, Stats SA does not publish full details of its adaptation of Demproj when producing its mid-year estimates as it regards this as proprietary. Concerns have been raised about the accuracy of the official mid-year estimates in terms of age distribution, particularly for ages below 40 last birthday in 2011 (e.g. Dorrington 2013). As such, this research critically analyses the method used by Stats SA to produce mid-year estimates and assesses the feasibility of using administrative data to produce mid-year estimates for South Africa.

The base population is adapted from the 2001 census population. Birth and death registration data are used in a cohort component approach to produce alternative mid-year estimates for South Africa for the years 2002-2011. Prior to using these data, they are adjusted for incompleteness of registration. Levels of completeness of birth and death registration are estimated by extrapolating earlier estimates of completeness from various researchers. The mid-year estimates obtained are compared with those published by Stats SA in order to assess the relative quality of the two series of mid-year estimates. The mid-year estimates for 2011 are also compared with the mid-year population estimated from the 2011 census. These comparisons help identify the mismatches to the census and their possible causes and as such, these may lead to improved population estimates in the future, and a viable alternative method to that currently being used by Stats SA.

ACKNOWLEDGEMENTS

I hereby extend my heartfelt gratitude to Professor Rob Dorrington for the resolute supervision and much appreciated advice and suggestions culminating in the successful completion of this dissertation. I also extend my gratitude to the Hewlett Foundation for providing me with the necessary financial aid which enabled me to successfully complete my masters studies. I would also like to convey many thanks to the staff within the Centre for Actuarial Research (Professor Tom Moultrie, Doctor Visseho Adjiwanou, Zerina Matthews) for the assistance they gave me, in one way or another, during the course of my studies.

I also extend my appreciation and gratitude to my colleagues for the fruitful discussions and all the assistance they gave me which have certainly gone a long way in helping me reach this point. Last, but certainly not least, I would like to make special mention of Professor Rob Dorrington for his extensive contribution in the method used to estimate completeness of birth registration and the method used to estimate cohort deaths from period deaths.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Background

The need for population estimates more recent than those from the last available census has long been appreciated by various government ministries and public and private organisations (Sterns 1935). Not only do such numbers serve as a foundation for important national statistics such as the production of per capita indices, but they also provide denominators necessary for estimating important demographic indicators of a population such as mortality, fertility and migration rates.

Mid-year population estimates serve numerous functions in various sectors of society from governments to private and public companies. Government ministries, and the public health sector require mid-year estimates for planning and monitoring service delivery, resource allocation and economic policy development (Office for National Statistics 2001). Some countries have developed legislative requirements to produce mid-year population estimates (Australian Bureau of Statistics 2009), underlining their importance. Mid-year estimates are also essential in informing government policy on the appointment of the number of seats in the parliament (Australian Bureau of Statistics 2009). Mid-year population estimates are also used by different stakeholders in the formation, monitoring and appraisal of policies, systems and programs that affect the lives of millions of people. The importance of their accuracy cannot be stressed enough.

Countries that produce mid-year estimates are furnished with standards to adopt when producing population estimates and encouraged to produce them in order to facilitate national and international policy assessments and inform decision making (International Monetary Fund 2007). However, countries are not obliged to produce the estimates, unless explicitly mandated by an act of parliament, hence some countries do not produce mid-year estimates.

In May 2013, Statistics South Africa (Stats SA) published the official mid-year estimates of the population of South Africa, which were the first since the 2011 census. Concerns were expressed regarding inconsistency of the population estimates with the 2011 census age distribution (e.g. Dorrington 2013). Consistency of a series of post-censal population estimates with census population estimates, in terms of the age distribution, is particularly important so as to avoid contradictory policy and planning. Inaccuracy of estimated numbers in specific age groups, for example, affects national planning for purposes such as child immunisation coverage and provision of classrooms

and teachers (Dorrington 2013). It also affects the estimation of important age-specific demographic indicators such as infant mortality, and other important national statistics which could have far reaching and adverse implications.

1.2 Statement of the problem

Stats SA have been producing mid-year estimates of the population in South Africa using DemProj. In order to check for consistency with other series of population estimates in terms of age and sex distribution, and other demographic features, it is important to compare their mid-year estimates with the population estimates from census series. When Stats SA's 2011 mid-year estimates are compared with the population estimates derived from the 2011 census, inconsistencies are found particularly in terms of the age distribution for ages below 40 last birthday in 2011. Such distortions may mislead users of the mid-year estimates and can have negative implications when used to inform policy making and decision making. Thus, it is important to analyse critically the method used by Stats SA to produce mid-year estimates in order to shed light on some of the causes of inconsistencies in their estimates.

1.3 Research objectives

Main objective

This main objective of this research is to assess the feasibility of using administrative data to produce mid-year estimates for South Africa using the cohort component method. Births and deaths from vital registration, and estimated migration will be used to project the population from a 2001 population base, estimated from the 2001 census. Since vital registration was incomplete during the projection period (2001-2011), it will be necessary first to estimate completeness of birth and death registration.

Specific objectives

The specific objectives are:

- To review the methods used by selected countries to produce mid-year population estimates;
- To assess the reasonableness of adopting practices used by other countries in using administrative data to produce mid-year population estimates;
- To analyse critically the method used by Stats SA to produce mid-year population estimates;

- To assess the consistency of Stats SA's series of mid-year estimates in terms of the age and sex composition and aggregate numbers, with other series of population estimates such as the census

1.4 Significance of the study

Mid-year population estimates are important and they serve a wide range of functions which affect many people. Their accuracy is thus paramount. Critically analysing the method used by Stats SA is therefore an important undertaking which would shed light on some of the inaccuracies that have been observed in their series of mid-year estimates. Assessing the feasibility of using administrative data to produce mid-year estimates is also important as it would possibly present a viable alternative to the method used by Stats SA which may even produce more accurate mid-year estimates.

1.5 Organisation of the dissertation

This dissertation is presented in five chapters. The next chapter (Chapter 2) explores existing literature relating to the production of population estimates and includes a critical review of the method currently being used by Stats SA to produce mid-year estimates. Chapter 3 presents the data sets used in the analysis, and the methods used in the study. Chapter 4 presents the results from applying the methods, including completeness estimates of vital registration and the series of mid-year estimates obtained using administrative data. It also includes analyses and discussions of the results obtained and an analysis of the quality and accuracy of the mid-year estimates produced using administrative data. Chapter 5 concludes the thesis by summarising the results and assessing the work presented in the dissertation. It also includes a section on a suggested method of estimating international migration annually, inspired by the ratio correlation method.

2. LITERATURE REVIEW

2.1 The production of current population estimates

The United Nations (UN) sets out standards regarding the production and quality of current population estimates, as well as recommendations to ensure that such estimates are internationally comparable and internally consistent (United Nations Population Division 1952). Furthermore, the International Monetary Fund (IMF) also sets out standards on dissemination of data, including population data, and prescribes practices relating to the coverage, periodicity and timeliness of such data, access by the public, and the integrity and quality of the disseminated data (International Monetary Fund 2007). In addition to adhering to these standards and recommendations, countries undertake the exercise of producing inter-censal (the period between censuses) population estimates for several important reasons such as informing national budget allocations and tracking the development of national projects.

Population estimates are used for various planning purposes by countries that do produce them. The most commonly cited use for population estimates is formulation, monitoring and evaluation of government policies such as those relating to healthcare, social assistance, social services and childcare (Australian Bureau of Statistics 2009, Statistics Canada 2012, Office for National Statistics 2007, Long 1993). Population estimates are also used for the distribution of central government funds to various government departments, with billions of dollars distributed annually to subnational territories on the basis of these population estimates (Long 1993, Statistics Canada 2012). It is quite clear that population estimates provide essential information for administration and planning in the government and private sectors.

2.2 Recommended practices in the dissemination of various data sets, including population statistics

The International Monetary Fund (IMF) publishes standards for data dissemination in the Special Data Dissemination Standards (SDDS). The standards are developed in order to facilitate the availability, to policymakers and various government ministries, of well-timed and comprehensive statistics, thereby helping countries in their development of sound economic policies (International Monetary Fund 2007).

The IMF recommends the production of various statistics. Among the statistics recommended for production are current population estimates which play an important role of providing the denominators needed in calculating per capita indices and other

important financial indicators. Regarding the dissemination of the recommended statistics, the SDDS identifies four main dimensions (requirements) and these include (1) data coverage, periodicity and timeliness, (2) access by the public, (3) integrity of the disseminated data, and (4) quality of the disseminated data (International Monetary Fund 2007, 1). For each of the four requirements, the SDDS sets out best practices that users of the standards are encouraged to observe and monitor.

The practices recommended by the SDDS relating to coverage, periodicity and timeliness of disseminated data mainly apply to fiscal and other financial data. However, in general, the practice that would relate to the production of current population estimates is that data should be disseminated on a timely basis, which is regarded as pivotal in ensuring transparency in policy formation and review (International Monetary Fund 2007). The need for timely population estimates is relevant in maximising the usefulness of the estimates, particularly in resource allocation and the monitoring of policies, programs and short-term developments.

It is important for the public to have sufficient access to government statistics and methods used to produce them since such statistics are considered to be public goods (International Monetary Fund 2007). To promote adequate access to such data, the SDDS recommends that statistical agencies furnish the public with, and adhere to, advance release calendars (ARCs) of future release dates of official statistics and the simultaneous release of data to all interested parties (International Monetary Fund 2007). The SDDS highlights the importance of ARCs as they allow users to plan their analyses and other related activities. Simultaneous release refers to providing users with data at the same time. In this regard, the SDDS encourages statistical agencies to release data in electronic format in order to facilitate simultaneous access by all users (International Monetary Fund 2007).

To promote data users' confidence in released data and in the organisation releasing the data, the SDDS specifies requirements relating to the integrity of the data. Relating to this, the SDDS points out transparency in a statistical agency's practices and procedures, be they administrative or related to changes in methodology, as a fundamental factor in cultivating data users' confidence when using the data (International Monetary Fund 2007). To facilitate this requirement, the SDDS prescribes four practices namely: (1) dissemination of the terms and conditions under which official statistics are produced, (2) identification of internal government access to data before release, (3) identification of ministerial commentary on the occasion of statistical

releases, and (4) provision of information about the revision and advance notice of major changes to methodology (International Monetary Fund 2007). These practices are meant to give users of official statistics a platform for voicing their concerns regarding any practice within in the process of producing these data.

The requirement relating to data quality is essentially objective, and quality standards are based on the type of statistics produced. For statistics tracking short-term developments, timeliness is considered an important requirement. For statistics detailing broader interrelationships between variables, greater care needs to be taken in determining quality aspects (International Monetary Fund 2007). Since population statistics basically serve both these purposes, extra care is required in their production. In addition, because population estimates serve different purposes, different approaches need to be adopted when producing them.

To facilitate quality assessment of the data for users, the SDDS calls for the dissemination of documentation on methodology and sources used in preparing statistics and the provision of statistical cross-checks and assurances of reasonableness (International Monetary Fund 2007). It is recommended that when there is enough information to allow for a quantitative statement of the margins of error in a population estimate to be made, such a statement should be published together with the estimate (United Nations Population Division 1952).

Although these standards are not mandatory for countries or statistical agencies responsible for the production of official statistics, they enhance the quality and usefulness of statistics and population estimates. This being the case, the IMF, through the SDDS, merely encourages any countries using the standard to observe and maintain the practices as prescribed.

2.3 The production of population estimates

Producing population estimates requires the estimation of the base population and the three components of population growth, fertility, mortality and migration. The United Nations Population Division (1952) published a manual detailing recommended practices for producing current population estimates. Although the manual is quite old and has not been updated, it contains fundamental practices for estimating population numbers using various data sources.

Information regarding population growth and size can be obtained by means of periodic census enumerations, administrative records of births, deaths and migration, or, in a few countries, by means of a continuous population register. These sources can be

supplemented in some areas by records of school attendance, occasional sample enumerations, taxation data, social insurance records and voting registers among other sources (United Nations Population Division 1952).

Sources of population data are not readily available in all countries, and in countries where they are available, they are not always up to date or reliable. The most commonly used source of population data is the periodic census enumeration. Census enumerations are quite costly, which is why most countries undertake them once every five or, more commonly, ten years. Moreover, the final results of a census enumeration are usually not entirely up-to-date at the time they are released (United Nations Population Division 1952).

The need for population estimates more recent than those from the last available census has been emphasised (e.g. Sterns 1935), and countries appreciate this fact. Current population estimates provide knowledge which is crucial in developing, monitoring and maintaining policies that pertain to various groups within a population, and in resource allocation as carried out by different ministries within a government (United Nations Population Division 1952).

The importance of ensuring that users of population estimates know how much they can rely on their accuracy should not be taken lightly. It is the responsibility of the producer of the estimates to ensure that estimates are not misused by stating as clearly as possible, the magnitude of possible errors (United Nations Population Division 1952). There are several methods, quantitative and qualitative, that statistical offices can use to determine the accuracy of population estimates.

When an estimate is based on certain assumptions, it is important to make several independent estimates based on other reasonable assumptions, then conduct a quantitative analysis of the differences between the estimates produced (United Nations Population Division 1952). Extreme values obtained may be used to approximate the limits of the range of the population estimates, which should be stated, and the estimates produced which are considered to be the most accurate may be taken as the best estimate (United Nations Population Division 1952).

There are various ways of presenting margins of error once the accuracy of the population estimates produced has been determined. One such way is to publish a quantitative assessment of the margins of error that estimates are subject to if the available information allows for this. Such assessment may be in the form of a percentage of possible error in each direction, of the population estimate (United

Nations Population Division 1952). However, with population estimates (even from a census), it is usually unlikely that quantifications of error be made accurately and stated confidently. In this regard, the United Nations Population Division (1952, 6) recommends that if it is felt that the margin of error cannot be confidently expressed as a percentage of the associated figure, then certain qualifications may be used instead, such as "estimate believed to be fairly accurate", or "approximate estimate", or "very approximate estimate, possibly subject to large error".

Efforts to state margins of error are important because the failure to indicate the approximate nature of certain statistics may give an impression that the estimates are exact, or at least nearly so. In any case, such commentary on possible inaccuracies is an indication that sufficient efforts are being made to appraise the quality of the estimates, and hence efforts are also being made to improve the accuracy (United Nations Population Division 1952).

2.3.1 Internal consistency of population estimates

The maintenance of internal consistency of population estimates is important. Internal consistency is mainly concerned with the population definition and the area covered in estimating the population. If estimates are obtained by updating the census population using vital statistics and migration statistics, the reference area in the census population should be identical to that of the vital statistics and migration statistics (United Nations Population Division 1952). This is particularly relevant when producing sub-national population estimates since sub-national administrative divisions are changed from time to time.

Common sources of inconsistency include failure to account for changed sub-national boundaries, failure by vital registration systems to cover adequately all the areas within national boundaries and the difficulty in recording all migrant movements in and out of a country (United Nations Population Division 1952). In such cases, incomplete statistics of births, deaths and migration need to be inflated or adjusted accordingly before they can be used to produce population estimates.

2.4 Commonly used methods of producing current population estimate

Different countries have different data sources and resources available for producing population estimates, hence the methods used for producing post-censal estimates vary from one country to another. Most countries can be grouped by three main methods of producing post-census population estimates.

The first method, used by relatively few countries, is to base population estimates on a continuous population register (United Nations Population Division 1952). The population register basically updates population numbers for components of population growth at regular intervals, such that at stipulated time points, it is possible to estimate the population size. Births and immigrant arrivals are added to the register, while deaths and emigrant departures are deleted from the register, with the actual occurrences being recorded within specified time intervals (monthly, quarterly or semi-annually) of occurrence (United Nations Population Division 1952). Consolidating population growth components from various sources and regions within a country at convenient and sufficiently frequent intervals such as once or twice a year would then yield a highly accurate estimate of the population (United Nations Population Division 1952). Current population estimates are thus produced using the data from the register and applying a cohort component approach. Some of the countries known to be using this approach to produce population estimates include Belgium, Denmark, Finland, Netherlands, Sweden and Austria (Ormiston-Smith, Smith, and Whitworth 2006).

The second method is used by countries with sufficiently accurate statistical systems, though without a central population register, to track components of population growth. Such countries have separate sources of information on the components of population growth in place which are sufficiently complete (United Nations Population Division 1952). Using the population numbers from the latest census, post-censal population estimates are obtained using the cohort component method. Numbers of births and deaths are obtained from vital registration systems and estimated numbers of immigrants and emigrants from various sources such as surveys, since there is no central registration system for migrants in most countries. If vital registration data and migration data are obtained from statistical systems which are sufficiently complete, then current population estimates obtained this way would most likely be quite accurate (United Nations Population Division 1952). Some of the countries known to be using this method include Greece, Ireland, Portugal, France, UK, USA and Australia (Ormiston-Smith, Smith, and Whitworth 2006)

The third method is used by countries that have undertaken censuses, regularly or irregularly, but their vital registration system is either severely incomplete or non-existent. Population estimates for countries in this group can be produced by means of mathematical extrapolation, which is not as reliable as the other methods since population change is often too dynamic to conform to a mathematical formula (United

Nations Population Division 1952). Prior to 2004, South Africa used to use this method before switching to using SPECTRUM.

2.5 Using administrative data to produce population estimates

The second method as described in the immediate previous section is the main focus of this research, hence it will now be described in greater detail. Administrative data are compiled by various government departments and public private organisations to monitor events and activities that are relevant to their functions and operations. Such data are primarily collected for non-statistical purposes, but they may be used for statistical purposes (Nordbotten 2010). Even though administrative data mainly serve the purpose for which they are collected, such as proof of identification for individuals in the case of civil registration, they could also be used in the production of population estimates.

The main issues surrounding the use of administrative data are obtaining permission to use the data (if the data belongs to a non-affiliate of a statistical agency, e.g. mobile phone usage data from a private network provider), timeliness of the data and concerns about the quality of the data (International Institute for Vital Registration and Statistics 1981). To alleviate some of these issues, statistical agencies have, over the years, developed initiatives, such as collaborating with owners of the data in the data collection, to facilitate efficient and effective usage of administrative data (United Nations Department of Economic and Social Affairs 2010).

The main administrative source of birth and death data is a civil registration and vital statistics (CRVS) system. However, CRVS systems, particularly in developing countries, have been marred by incomplete registration of vital events. It is argued that in Africa and Asia, because such systems have not been sufficiently developed over the past thirty years, millions of people have been born and died without any official record of their existence (Lopez, Mikkelsen, Rampatige et al. 2012).

The International Institute for Vital Registration and Statistics (1981) conducted a survey of 32 developing countries from Africa, Latin America, Asia and Oceania which sought to identify the main obstacles to achieving satisfactory registration of vital events. Funding was identified as the most common obstacle and "the underlying cause of many of the immediate problems facing developing countries, such as lack of adequately trained staff" (International Institute for Vital Registration and Statistics 1981, 1). Another main obstacle identified by the International Institute for Vital Registration and Statistics (IIVRS) was the organisation of the CRVS system. Regarding

this, responding countries cited lack of appreciation by government officials at sufficiently high levels of the importance of complete and accurate vital statistics (International Institute for Vital Registration and Statistics 1981). This is a particularly counterproductive obstacle because without sufficient government support, CRVS systems would not get enough funding and resources to improve their completeness.

2.5.1 Aspects of administrative data quality

Penneck (2007) identifies some of the main aspects of data quality pertaining to administrative data. These include relevance, accuracy, timeliness, accessibility and comparability of the administrative data. Relevance essentially refers to the fact that administrative data are collected based on definitions and coverage relevant to the administrative system, which may not be sufficient in producing population estimates. Thus, when using administrative data to estimate population numbers (or for any statistical purpose), there is an inherent need to control for the insufficient coverage prior to using the data (Nordbotten 2010).

One of the main aspects of the quality of administrative data is timeliness. The process by which administrative data are collected involves the integration of data from multiple sources (Nordbotten 2010), which is an elaborate process that takes time and as a result, administrative data are usually released sometime after they are collected. Furthermore, the administrative data used to estimate components of population growth used in population estimation come from different sources. Thus, if the release of data from one of the sources is delayed, the entire process of producing current population estimates is consequently delayed (Ruotsalainen 2004). These factors should be closely monitored when producing current population estimates because the estimates are often required to monitor short term developments or for resource allocation for a near future.

2.5.2 Estimating current migration

Estimating migration generally involves estimating in-migration (migration into a sub-national region or country) and out-migration (migration out of a sub-national region or country). When producing national population estimates, estimates of net international migration (the difference between numbers of in-migrants and out-migrants for the country) are required. For subnational population estimates, estimates of net internal migration (the difference between numbers of in-migrants and out-migrants for each subnational region in a country) are required. High quality estimates of both

international and subnational migration are essential in the production of population estimates (Jensen 2012).

International emigration is regarded as the most difficult component of national population change to estimate due to various reasons (Jensen 2012). Some of the main reasons are the lack of central migrant registration systems, and the fact that the emigrant population is not resident in the country, and hence cannot be identified in the country's censuses or surveys (Jensen 2012). Apart from this, censuses (and some surveys) are not carried out annually, and are therefore insufficient in estimating international emigration. International in-migration and subnational migration on the other hand can be estimated using proxy data from various administrative data sources in a country. These are data that do not refer directly to the migrants, but may be used as an indicator of the extent of increase or decrease in numbers of migrants between successive periods. For international in-migrants, such data may be, for example, visa application statistics. For subnational migrants, such data may be, for example, water consumption or sales of a staple commodity within a subnational region.

In countries with good and nationally representative administrative data sources, it is be possible to estimate current migration from proxy data. The Office for National Statistics (ONS) in the United Kingdom, for example, uses patient registers to track change of addresses of patients in the UK and Wales, which they use to estimate sub-national migration flows (Office for National Statistics 2007). In countries with a universal healthcare system, patients normally re-register with a new doctor when they move, hence the data from patient registers are "considered to provide a good proxy indicator of migration" (Office for National Statistics 2007, 22).

For countries without sufficiently accurate and nationally representative proxy data for estimating subnational migration, "conjectural" (United Nations Population Division 1952, 10) population estimates may have to be used instead. These are population estimates not based directly on data relating to the population itself (Bryan 2004, United Nations Population Division 1952). Procedures for obtaining conjectural population estimates may vary from guessing to converting data on population density or the consumption of a staple commodity to a population estimate by applying a factor which represents a ratio of the population to the reference point (Bryan 2004, United Nations Population Division 1952). If such population estimates are obtained, they may be used to estimate migration by making use of the demographic balancing equation. However, the estimates obtained are usually subject to wide margins of error (Bryan

2004, United Nations Population Division 1952). Examples of such methods include the ratio correlation method, and two variants, the difference correlation method and the average ratio method.

The foundation of the idea of ratio correlation methods was laid by Snow (1911) in his discussion about using multiple correlation (correlation between one dependent variable and several independent variables) to produce post-censal population estimates. He posits that the relationship between symptomatic indicators (proxy data) and the corresponding population remained unchanged over time (Snow 1911). Using this premise, Snow demonstrates that coefficients from a multiple regression model built using data from one decade can be used with data from a subsequent decade with the insertion of new values of the independent symptomatic variables. Additionally, he posits that the variables should be constructed on the basis of relative change rather than absolute change (Snow 1911). Symptomatic variables could include births, deaths, school enrolment, tax returns, mobile phone usage or any other good indicator of geographic population distribution.

Swanson and Beck (1994) describe the ratio-correlation method using the following equation:

$$\left(P_{i,t}\Big/\sum_i P_{i,t}\right)\Big/\left(P_{i,t-z}\Big/\sum_i P_{i,t-z}\right)=a_0+\sum_{i,j}\left(b_j*\left(S_{i,j,t}\Big/\sum_i S_{i,j,t}\right)\Big/\left(S_{i,j,t-z}\Big/\sum_i S_{i,j,t-z}\right)\right)+\varepsilon_i$$

where:

a_0 = the intercept term to be estimated

i = subarea ($1 \leq i \leq n$)

j = the symptomatic indicator ($1 \leq j \leq k$)

t = year of the most recent census

z = the number of years between the censuses whose data is used to construct the model

b_j = the regression coefficient to be estimated

$P_{i,t}$ = the population in subarea i at time t

$S_{i,j,t}$ = the symptomatic variable for subarea i at time t

ε_i = the error term

Schmitt and Crosetti (1954) used the United States 1930-1940 inter-censal data to build their multiple regression model. In determining which symptomatic variables to include in their model, they used zero order correlation coefficients (Pearson's correlation coefficients) relating population changes in each county to various series of symptomatic variables available. In each instance, the dependent variable was the percentage of the state's population residing in the county in 1940 divided by the corresponding percentage in 1930 and the independent variable was the corresponding ratio for the symptomatic data (Schmitt and Crosetti 1954). They selected the independent variables with the highest correlation to include in the model. These include live births, registered vehicles and public school enrolment (Schmitt and Crosetti 1954). To test the model, they substituted post-censal symptomatic data (data from the period after a census) to obtain post-censal population estimates for each county. Their tests indicated that the ratio-correlation method yields estimates with an average error of 7.4 per cent for the ten year post-censal period which is lower than the averages for other methods, such as the vital statistics method, the arithmetic projection method and the apportionment method (Schmitt and Crosetti 1954).

About ten years later, Goldberg, Rao, and Namboodiri (1964) also tested the accuracy of the ratio-correlation method, using a model which had more independent variables. Their tests yielded an average error of 3.7 percent for the estimates from their model for the 1950-1960 post-censal period, which is also lower than the averages from the other methods they tested (Goldberg, Rao, and Namboodiri 1964).

Thus, the method can yield reasonable sub-national population estimates which can then be used to estimate sub-national migration using the population balancing equation. However, the method does have weaknesses which have been discussed by some researchers (Bryan 2004, Ericksen 1974, 1973, Swanson 1978b, D'Allesandro and Tayman 1980, Tayman and Schafer 1985, McKibben and Swanson 1997). For example, Bryan (2004) points out that substantial time lags in the availability of symptomatic indicators compromises the quality of population estimates obtained using the ratio-correlation method and related techniques. Accuracy of the estimates obtained depends on several factors including the quality of data on symptomatic indicators and the validity of the central underlying assumption (that the observed statistical relationship between the independent and dependent variables in the past inter-censal period will persist in the current post-censal period) of the regression based models (Swanson and Tayman 2011).

Tayman and Schafer (1985) stress the importance of exercising good judgement when applying the ratio-correlation method because analysts need to consider the difference in coverage of each of the symptomatic variables. The regression methods also have very limited application in estimating the age and sex composition of populations (Swanson and Tayman 2011), since different age groups and genders are correlated differently with symptomatic variables. However, it is possible to disaggregate the total population numbers obtained using the method by applying the age and sex distribution from the most recent census (Swanson and Tayman 2011).

2.5.2.1 Methods used by various countries to estimate numbers of international migrants

Most countries don't maintain a central register of international migrants. CSOs typically use data from various sources to estimate international migration. The Australian Bureau of Statistics (2009), for example, uses data from passenger cards filled out by all passengers arriving and departing at airports and other ports of entry to estimate net international (overseas) migration. They also use data from passports and visa applications to identify international visits/moves lasting more than a year which would constitute international migration.

The United States Census Bureau (2013) uses data collected by the American Community Survey (ACS) which has place of birth and previous place of residence questions designed to identify both sub-national and international migration.
For England and Wales, international migration is estimated using data from the International Passenger Survey (IPS) administered to approximately 1% of all passengers arriving and departing at airports in England and Wales (Office for National Statistics 2007). They also use data from the UK Home Office to identify visitor switchers and migrant switchers. Visitor switchers are defined as individuals who enter or leave the UK as visitors intending to stay less than a year but actually stay in the UK or destination country for longer than a year, effectively switching from being non-migrants to migrants (Office for National Statistics 2007). Migrant switchers are defined as individuals classified as migrants in the IPS but actually stay in the UK or destination country for less than a year. In addition the ONS uses data from the Irish Central Statistical Office to estimate migrant flows between the Republic of Ireland and the UK.

2.5.3 Using mobile phone usage data to estimate migration

Innovative use of symptomatic and/or administrative data has seen the development of new methods of estimating internal migration. One such method was proposed by Blumenstock (2012), which involves inferring patterns of internal migration from

mobile phone call records of 1 500 000 Rwandans. Blumenstock (2012) posits that using mobile phone data enables the identification of migration patterns of those individuals whom statistical agencies have typically found difficult to survey (undocumented citizens, illegal immigrants and people living in very remote areas).

Blumenstock and his research team used data collected from global positioning systems installed in smartphones, whilst for less advanced mobile phones without such software, location was triangulated from the nearest cell phone tower which processed calls from these phones. Due to the privacy concerns surrounding the use of individuals' exact location data, only anonymous data without any identifying feature of demographic information could be supplied by the network provider. However, they used data from a geographically stratified sample of mobile phone users to disaggregate patterns of migration by demographic type (Blumenstock 2012).

With these data, they used individuals' history of phone calls to infer approximate monthly locations. Using a model for inferring migration by using inequalities to distinguish between types of movements (temporary or permanent), they calculated aggregate numbers of sub-national migrants in Rwanda (Blumenstock 2012).

The model developed by Blumenstock (2012) could be further refined to give more perspectives about internal migration. Like the ratio correlation method, Blumenstock's model gives aggregate numbers of migrants in each subnational region. The age distribution from the most recent census can be used to disaggregate numbers by age group. However, one problem is that the approach requires data not freely available and legal requirements need to be met to obtain the data.

2.5.4 Estimating numbers of births and deaths

Estimates of births and deaths are obtained from a CRVS system. For most countries, registration of births and deaths is compulsory and governed by legislation, for instance the Provincial and Territorial Vital Statistics Act in Canada (Statistics Canada 2012). Other countries also make a note of similar legislation (Australian Bureau of Statistics 2009, Office for National Statistics 2007, United States Census Bureau 2013).

There are several challenges associated with using civil registration data to produce population estimates. Some of these include late registration, incompleteness of civil registration systems and the time lag associated with the official release of vital statistics from civil registration. These issues are addressed in different ways by CSOs. The Australian Bureau of Statistics (2009) for instance, publishes three sets of population estimates for the same date to address the lag in the release of vital statistics.

Birth and death registration data are first available five to six months after a reference quarter and these are used to produce preliminary estimates. Revised estimates are then produced 21 months after a reference quarter when more revisions have been made to the vital registration data to cater for late and missing registration. The final estimates are then produced after the next five-yearly census. Similar approaches are used by Canada and France (Statistics Canada 2012, National Institute of Statistics and Economic Studies 2009).

The Office for National Statistics (2007) of the U.K. uses a slightly different approach. They produce population estimates for England and Wales once, 14 months after the reference date. This is meant to cater for any late registration and allow for any revisions made to the vital statistics from the time they are released. They also make adjustments to deaths for the period 12 months prior to the reference date to cater for missing death registration. They do this by adding to the provisional deaths, the difference between the provisional deaths for the period 12 months prior to the reference date and the final deaths for the period 12 to 24 months before the reference date.

2.6 Methodology used by Statistics South Africa in producing mid-year population estimates

Mid-year estimates of the population in South Africa have been produced and published since 1966. They were previously produced by the Department of Statistics, which then became the Central Statistical Services, and then, finally, Statistics South Africa. From 1966 to the early 1980's, mid-year estimates were only published in a two page document giving population totals by population group and sex. There was no detail given on the method used to estimate the population numbers. However, in the 1986 publication, some details were given about the method used to produce the mid-year estimates. Mathematical extrapolation was used to produce estimates of aggregate population totals by sex and race. Expected growth rates based on census results were used to project the population from a base of the last available census (Central Statistical Service 1987). By 1991, population numbers were produced with more detail, including numbers by province. In 1998, mid-year estimates were for the first time produced using the cohort component method, with an estimate of the 1996 population as the base, and fertility and mortality assumptions derived from the 1996 census (Statistics South Africa 1998).

During the 2001-2011 inter-censal period, Stats SA have annually produced mid-year estimates of the population in South Africa at national and sub-national levels using two main methods and various data sources as they became available. To produce the 2002 mid-year estimates, Stats SA used an estimate of the population in 1996 as the base since data from the 2001 census was not yet available (Statistics South Africa 2002). Using the 1996 census totals and 2001 projected population totals, they estimated an "inferred" growth rate by assuming exponential population growth (Statistics South Africa 2002). The inferred growth rate for the five year period was calculated as $r_{inferred} = \ln(P_{2001}/P_{1996})$, where P_{2001} is the projected 2001 population and P_{1996} is the 1996 census population. Using the inferred growth rate, they shifted the 1996 census population to the middle of the year and then extrapolated the inferred annual growth rate to estimate the population as at the middle of 2002 (Statistics South Africa 2002). Since they used an aggregate growth rate, they only produced total numbers not disaggregated by age, but disaggregated by sex, race and geographic location.

For their 2003 mid-year estimates, Stats SA used an approach similar to the approach for 2002. However, data from the 2001 census were now available hence they used an estimate of the population in 2001 as the base (Statistics South Africa 2003). They calculated age specific growth rates using the 1996 and 2001 census populations, which they extrapolated to estimate the mid-year population in 2003 in five year age groups by assuming exponential population growth (Statistics South Africa 2003). For 2003 and earlier years, Stats SA indicate that they opted for mathematical extrapolation due to the lack of sufficiently accurate fertility, mortality and migration data with which to carry out annual projections (Statistics South Africa 2002, 2003).

From 2004 onwards, Stats SA stopped using mathematical extrapolation. To produce the national estimates, they started using population projection software, SPECTRUM, developed by the Futures Institute (Statistics South Africa 2014). For the provincial estimates, they started using a sub-national cohort component approach. Fertility, mortality and migration assumptions used as input for their model were obtained from research by various authors (Van Aardt and Van Tonder 1999, Moultrie and Timæus 2003, Moultrie and Dorrington 2004, Dorrington, Moultrie, and Timæus 2004, Phillips, Phoshoko, and Cronje 2004, Udjo 1997, 1998, 1999, 2004b). Stats SA also changed the base population for national mid-year estimates, changing from the 2001 census population to the 1970 census population. They later changed their base to an estimate of the population in 1985. However, it is not clear in which year they

switched to the 1985 base population because in their publications of MYEs for the years 2005-2012, there is no explicit statement pointing out which base population was used. However, for the 2013 MYEs, they clearly state that they used the 1985 population as a base. For provincial population estimates, they continued to use an estimate of the population in 2001 as the base.

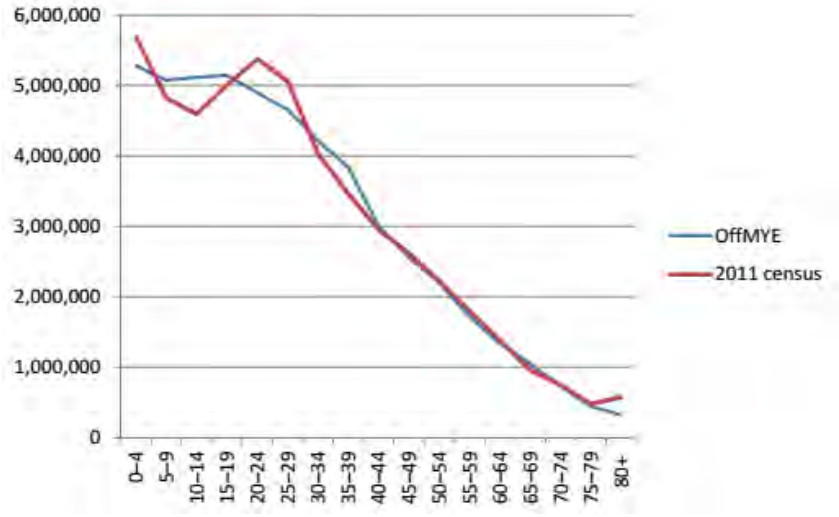
This review focuses on the approach currently being used by Stats SA to produce mid-year estimates, which they have adapted since 2004. Statistics South Africa (2014) point out that each series of MYEs they produce is unique (unrelated), due to assumptions made and data inputs used, although all their series use 1985 as the base for national MYEs. Fertility and mortality assumptions are revised each year as new births and deaths data become available (Statistics South Africa 2014). Due to this, and the fact that different versions of the SPECTRUM software update assumptions and the software itself as new information becomes available, Stats SA note that population and demographic data in each of their statistical releases form a new time series (Statistics South Africa 2014). Accordingly, they stress that users of their statistical releases should only compare data by year in the same release and not data from different releases (Statistics South Africa 2014).

For the national estimates, Statistics South Africa (2014) use the DemProj and AIM (AIDS Impact Model) components of the SPECTRUM suite of models. DemProj uses the cohort component method to project the population, whilst AIM projects the demographic impact of the AIDS epidemic on the population over time (Stover 2003). To obtain a better understanding of the output from both these components of SPECTRUM, it is essential to understand the inputs. Unfortunately Stats SA refuse to provide such detail as they regard this level of detail to be "proprietary" (personal communication, Diego Iturralde, Executive Manager: Demography, Stats SA).

Fertility, non-AIDS mortality, migration and HIV prevalence/incidence are used as inputs for the model, and the projection is prepared such that the 2011 projected total population by race and sex for each province match those of the 2011 census (Statistics South Africa 2014). Even though they managed to match the 2011 census aggregate totals by sex, race and province, Stats SA acknowledge that there are differences in the sex ratios by age and in the age distributions by sex, race and province between their MYEs and the 2011 census population (Statistics South Africa 2014). However, they do not discuss the implications of these differences.

Dorrington (2013) expresses concerns about the differences in these age distributions and their implications, offering alternative mid-year estimates for the years 2001-2013, with an age distribution consistent with that of the 2011 census. Figure 2-1 shows the differences between the age distributions of the official 2013 mid-year estimates (males and females combined) of the population in 2011 and the 2011 census population.

Figure 2-1 Comparison of numbers by age group for 2011 from the 2013 official mid-year estimates and the 2011 census



Source: (Dorrington 2013, Figure 1), OffMYE represents the official mid-year estimates

Inconsistencies for ages 40 last birthday and under are apparent and they indicate that the fertility, migration, and possibly even mortality, allowed for in the projection to produce the 2011 MYEs was notably different from that indicated by the 2011 census age distribution. There is also a notable difference for ages 80+ last birthday, with the census 80+ last birthday being higher than the projected. There is usually pronounced age exaggeration in the census, which is probably the case here, hence the mid-year estimate is much more likely to be closest to the correct number of the 80+ last birthday age-group. Some implications of these differences include providing incorrect denominators used in the estimation of demographic rates of fertility and mortality, and misestimating the necessary capacity of immunisation programs if the estimated numbers of births and children are wrong (Dorrington 2013).

2.6.1 Incorporating the impact of AIDS into the national projection

AIM is used to incorporate the impact of AIDS on the projections. Data collected annually on pregnant women at the antenatal clinics since 1990 are the principal source of HIV prevalence used as input (Statistics South Africa 2014). However, the data

collected are not nationally representative since the sample is only of pregnant women attending public antenatal clinics. HIV prevalence amongst women attending public health facilities is expected to be higher than prevalence amongst women attending private health facilities (Statistics South Africa 2014). Furthermore, HIV prevalence amongst pregnant women is most likely different from HIV prevalence in the general population (Statistics South Africa 2014). Due to these sources of bias when using data from antenatal clinics, adjustments are made to account for (1) the differences in HIV prevalence amongst pregnant women attending the antenatal clinics and pregnant women attending private facilities, and (2) the difference in prevalence between pregnant women and the general adult population in order to minimise the bias (Statistics South Africa 2014).

Prior to applying adjustments to the data from antenatal clinics, the data are first standardised for race because the racial distribution of attendees in the clinics is different from the population distribution. Stats SA standardise the HIV prevalence of women attending the antenatal clinics using the racial distribution of the South African adult population (15-49) from Stats SA's latest "South African Statistics" publication (Statistics South Africa 2009b). Ideally, this standardisation should be based on the racial distribution of all pregnant women in South Africa, but since these data are not available, Stats SA use the racial distribution of the population. Correction factors used in the adjustment are calculated as the ratio of the unadjusted prevalence divided by the race standardised prevalence, which is done for each province (Statistics South Africa 2014).

Stats SA then use the Epidemic Projection Package (EPP), also incorporated into SPECTRUM, to produce a national HIV epidemic curve based on the adjusted data from antenatal clinics. HIV prevalence implied by the national HIV epidemic curve is used as input in the projection to obtain a series of mid-year estimates for the period 2002-2013 (Statistics South Africa 2014).

2.7 Completeness of birth registration in South Africa

The registration of births in South Africa has, over the years, been marred by late registration and missing births. The Department of Health (DOH), the DHA, Stats SA and various other government departments have implemented various initiatives to alleviate this. In 1995, the DOH set up a national committee to develop a National Health Information System Strategy for South Africa (NHIS/SA) (Mbananga, Madale, and Becker 2002). The main objective of the Health Information System (HIS) is to

facilitate the health system management by improving health service data collection and maintenance.

In 1999, a national District Health Information System (DHIS) rollout strategy was initiated (Statistics South Africa 2009a). The DHIS rollout included software that allows health facilities, including clinics and hospitals, to enter data relating to their services (Statistics South Africa 2009a). The data collected are transferred to provincial and national departments. The DHIS has significantly improved birth registration in recent years since its inception by involving clinics and hospitals in the birth registration process since the data transferred from clinics and hospitals is used to complement vital registration data (Statistics South Africa 2009a). The DOH and the DHA also set up committees at national and provincial levels to develop initiatives to help improve birth registration in public facilities (Statistics South Africa 2009a). These efforts have, over the years, collectively helped improve the completeness of birth registration.

2.8 Completeness of death registration in South Africa

The South African vital registration is incomplete, but gradually improving. Several researchers have conducted work using indirect techniques and different data sources to estimate plausible levels of completeness of South African death registration (Dorrington 1989, Timæus 1993, Dorrington, Bourne, Bradshaw et al. 2001, Nannan, Dorrington, Laubscher et al. 2012, Dorrington, Moultrie, and Timæus 2004, Dorrington and Bradshaw 2011, Darikwa and Dorrington 2011).

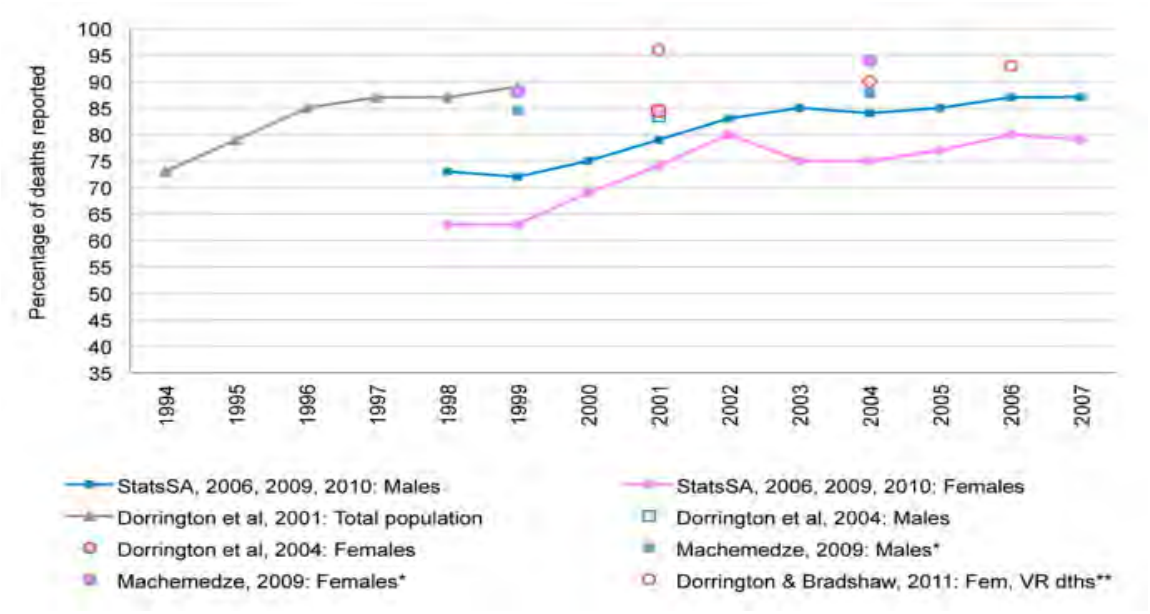
In the context of South African mortality research, it is important to incorporate the impact of the HIV/AIDS pandemic and how it affects different population groups. Dorrington et al. (2001) consolidated various data sources in South Africa, and taking into account the heterogeneity of the four population groups in South Africa, they applied the Synthetic Extinct Generations Method (SEG) (Bennet and Horiuchi 1984) to estimate national completeness of death registration. Using deaths from vital registration and the Population Register, and population estimates from the ASSA600 AIDS and Demographic Model, they estimated that national completeness of adult deaths had increased from 73 per cent in 1994 to 89 per cent in 1999/2000 (Dorrington et al. 2001).

When data from the second all-inclusive South African census in 2001 became available, Dorrington, Moultrie, and Timæus (2004) took the opportunity to use these data, together with the population counts from the 1996 census and registered deaths during the inter-censal period, to apply Hill's Generalized Growth Balance (GGB)

method (Hill 1987), adapted to allow for the impact of migration on the population. They estimated national completeness of death registration for the period 1996-2001 to be at 83.4 per cent for males and 84.5 per cent for females (Dorrington, Moultrie, and Timæus 2004).

The next available, and nationally representative, data source of demographic data of the population in South Africa needed in the estimation of death registration completeness (age distribution, migration) was the 2007 Community Survey. In their investigation of maternal mortality in South Africa, Dorrington and Bradshaw (2011) estimated national female death registration completeness for the period 2001-2007 using population counts from the 2001 census and the 2007 Community Survey. They estimated national completeness to be 91 per cent using the GGB method and 89 per cent using the SEG method (Dorrington and Bradshaw 2011). Figure 2-2 shows the recent trends of estimated completeness of adult death registration from various researchers. There has been an apparent improvement over time, with recent estimates by Dorrington and Bradshaw (2011) and Machemedze (2009) exceeding 90 per cent for the early 2000's.

Figure 2-2 Estimated completeness of adult death registration in South Africa between 1994 and 2007



Source: (Joubert, Rao, Bradshaw et al. 2013, Figure 1)

Completeness of child death registration is less easy to estimate due to a lack of reliable child mortality estimates, and the fact that there is no indirect method for doing this (Bradshaw and Dorrington 2007). However, mortality estimates obtained directly

using births and deaths from vital registration compared with empirical estimates can give an indication of the level of completeness of infant and child death registration. In their review of South African under-5 mortality statistics, Nannan et al. (2012) compared direct estimates of infant and child mortality to empirical estimates from previous research (e.g. Darikwa and Dorrington 2011). They conclude that there are indications that completeness of child death registration has risen sharply, reaching levels of 90 per cent for infants and 60 per cent for children aged 1-4 last birthday (Nannan et al. 2012). Thus completeness of child death registration is still not as high as it is for adults.

Darikwa and Dorrington (2011) derived estimates of completeness of child death registration using a multi-stage method with registered deaths from civil registration, data from the 2007 Community Survey, the 2001 census and previous estimates based on mortality data from the 1996 census and the 1998 SADHS. For the period 1996-2006, they estimated that infant death registration had improved from 43 per cent to 89 per cent; for children aged 1-4 last birthday, from 43 per cent to 57 per cent; and for children under five years old, from 44 per cent to 78 per cent (Darikwa and Dorrington 2011).

From the literature reviewed, there are indications that completeness of death registration has improved over time for both adults and children. These improvements present an opportunity to be able to use vital registration data to produce post-censal population estimates, with minimal adjustment to the vital registration data.

3. DATA AND METHODS

This chapter examines the data sets used in the analyses. This involves interrogating the quality of the data used to estimate various components of population growth in order to determine quality of current population estimates that could be produced using these data. The chapter will describe various methods used to analyse and make adjustments, where necessary, to the data from different sources before they can be used to produce mid-year population estimates.

The starting point will be the examination of the base population which is derived from the 2001 census population distribution.

3.1 The base population

The base population is derived from the 2001 census population in the following manner. First, since the 2001 census date was 9/10 October, it is necessary to adjust the single year age distribution from the census to reflect the estimate of the population as at the middle of the year (0.2726 years prior to the census). This is done by multiplying the 2001 census distribution by $e^{-0.2726r(x)}$, where r(x) is the continuously compounded average annual growth rate of the population aged x between 1996 and 2001. This adjustment assumes constant annual exponential growth of the population between 1996 and 2001.

A demographic analysis of this base population is then carried out to determine the quality of the population data. In order to identify any age heaping, undercount or any inconsistencies (with the 1996 census) in the 2001 census, the enumerated 2001 population is compared to an expected population which is projected from the 1996 census population. Fertility, mortality and migration assumptions used for this purpose are obtained from the ASSA2008 AIDS and Demographic Model. Having identified any problems with the 2001 census numbers for specific cohorts, it is necessary to make adjustments to the census numbers to correct the problems in order to obtain a suitable base for post-censal estimates. Adjustments are based on cohort size and sex ratio anomalies, relative consistency checks against the 2001 projection and evidence of age misreporting and age exaggeration at the older ages.

3.2 Sources of vital statistics in South Africa

The main sources of vital statistics in South Africa are the civil registration system, population censuses and sample surveys. When using data from these sources to produce population estimates, it is important to consider the completeness of the reporting and the problems associated with the use of such data. Some of these problems include misreporting of age at death, delays of death records under medico-legal investigation and delays in releasing the data (Joubert, Rao, Bradshaw et al. 2012). Different sources provide data with varying levels of completeness and there are different problems regarding the use of these data.

The Department of Home Affairs (DHA) administers the registration of births which is governed by the Births and Deaths Registration Act of 1992 (Republic of South Africa 1992). Births are captured using the DHA-24 forms for births of persons less than 30 days old and the DHA-24/LRB forms for births of persons aged 31 days and older (Statistics South Africa 2012c). Birth data are then edited, analysed and released annually by Statistics South Africa.

The DHA also administers death notification forms (BI-1663) which serve the purpose of registering deaths in accordance with The Birth and Deaths Registration Act of 1992 (Nannan et al. 2012, Republic of South Africa 1992). All death notification forms (DNFs), irrespective of the deceased's citizenship or whether the person was on the population register or not, are processed by Statistics South Africa who capture the cause of death data and other socio-demographic and health data (Joubert et al. 2012, Statistics South Africa 2012b). Data analysis and the dissemination of information are done by Statistics South Africa.

The register of deaths is updated annually as new information becomes available to improve the completeness of the data (Statistics South Africa 2012b). Mortality data are released with a time lag of approximately two years (Joubert et al. 2012) in order to allow for processing, late registrations and adequate analysis. This means population estimates produced using these mortality data would be available only at best towards the end of the second year after a reference date.

Another source of mortality data is the population register of South African citizens (with identity numbers or whose births have been registered) and permanent residents, maintained by the DHA since 1972 (Khalfani, Zuberi, Bah et al. 2005). A rapid mortality surveillance (RMS) system was set up in 1999 by the Medical Research Council (MRC) of South Africa and the University of Cape Town (UCT) to capture and monitor trends in death data from the population register (Joubert et al. 2012). The

main purpose of the RMS system is to allow for mortality trends by age and sex to be monitored within a few months after a date of death (Joubert et al. 2012). RMS reports are published with a time lag of approximately one year, although information about deaths is available within a few months of occurrence (Joubert et al. 2012).

3.3 Births

Birth registration data were obtained from a STATA file (Births1998-2012_F1.dta) downloaded from Stats SA's website[1]. The STATA file gives births by status of registration which is either "current" or "late". "Current" births are births that were registered during the year of birth up to February of the year after the year of birth, while "late" births are those that were registered later than this period. Current births in the STATA file actually correspond with the births as published in Stats SA's "Recorded live births" series (Statistical release P0305).

Stats SA normally publishes current births with approximately a one-year time lag. Thus for population projection purposes one would only have access to the births up to the year before the year for which the mid-year estimate is being produced, hence, only these births were used in the analysis. Table 3.1 below gives the births by status of registration, as published by Stats SA in 2012.

Table 3.1 Birth registration by status of registration, South Africa 1991-2011

Year of registration	Number of birth registrations			Percentages		
	Total	Current	Late	Total	Current	Late
1991	537 999	238 053	299 946	100.0	44.2	55.8
1992	501 461	228 445	273 016	100.0	45.6	54.4
1993	557 995	199 460	358 535	100.0	35.7	64.3
1994	667 107	246 345	420 762	100.0	36.9	63.1
1995	809 439	260 880	548 559	100.0	32.2	67.8
1996	998 798	295 719	703 079	100.0	29.6	70.4
1997	1 046 095	309 723	736 372	100.0	29.6	70.4
1998	1 216 337	273 180	943 157	100.0	22.5	77.5
1999	1 363 800	344 700	1 019 100	100.0	25.3	74.7
2000	1 407 833	409 707	998 126	100.0	29.1	70.9
2001	1 433 432	477 489	955 943	100.0	33.3	66.7
2002	1 517 671	557 573	960 098	100.0	36.7	63.3
2003	1 677 415	621 887	1 055 528	100.0	37.1	62.9
2004	1 475 809	728 283	747 526	100.0	49.3	50.7
2005	1 380 496	793 788	586 708	100.0	57.5	42.5
2006	1 346 119	860 263	485 856	100.0	63.9	36.1
2007	1 199 712	858 866	340 846	100.0	71.6	28.4
2008	1 277 763	915 674	362 089	100.0	71.7	28.3
2009	1 254 707	879 707	375 000	100.0	70.1	29.9
2010	1 294 694	889 691	405 003	100.0	68.7	31.3
2011	1 202 377	911 353	291 024	100.0	75.8	24.2

Source: (Statistics South Africa 2012c, 3)

[1] www.statssa.gov.za/

The numbers of late registrations are births from the respective years registered after the period referred to as current, up to 2012. Considering this, to extract "Year 1" births (births registered up to the end of the year after the year of birth) from the STATA file, two-way tabulations were used in STATA to give registered births by Year of Birth and Year of Birth Registration. Births registered in "Year 0" (births registered during the year of birth) and "Year 1" for each projection year were then obtained from this tabulation.

The following approach was used to adjust the birth registration data for incompleteness. The method adopted for estimating completeness of birth registration uses age-specific fertility rates (ASFRs) estimated from data on children ever borne reported by women during the 1996 and 2001 censuses. These estimates of fertility are used because at the start of the period for which mid-year estimates will be produced using administrative data (2002-2011), data from birth registration, including late registration, would not be available.

ASFRs from the 1996 census data were obtained from Moultrie and Timæus (2002). For the 2001 census data, ASFRs were obtained from Moultrie and Dorrington (2004). Table 3.2 shows the fertility rates for the 1996 and 2001 census.

For each year between 1996 and 2001, the ASFRs are linearly interpolated on the assumption that fertility is not expected to change significantly within five years. These rates are shown in Table 3.2. The numbers of women by age group in the age range 15-49 required to estimate the implied numbers of births are projected from the 1996 census age distribution using survival factors from the ASSA 2008lite population projection model.

Table 3.2 Estimates of fertility for South African Women for years 1996-2001

Age group	*Age Specific Fertility Rate*					
	1996	*1997*	*1998*	*1999*	*2000*	*2001*
15-19	0.078	0.075	0.073	0.070	0.068	0.065
20-24	0.151	0.146	0.141	0.136	0.131	0.126
25-29	0.156	0.153	0.151	0.149	0.146	0.143
30-34	0.125	0.124	0.123	0.122	0.121	0.120
35-39	0.087	0.085	0.082	0.080	0.077	0.075
40-44	0.042	0.040	0.037	0.035	0.032	0.030
45-49	0.007	0.008	0.008	0.009	0.009	0.010
TFR	3.230	3.153	3.076	2.999	2.922	2.840

Source: 2001 and 1996 derived from Moultrie and Dorrington (2004) and Moultrie and Timæus (2002)
Note: For the years 1997-2000, fertility rates have been linearly interpolated using rates from 1996 and 2001

After calculating the implied numbers of births using the fertility estimates, completeness of birth registration in the year of birth is estimated as the number of

births registered in the year divided by the expected numbers of births for that year. Table 3.3 shows these estimates of the completeness of birth registration in the year of birth.

Table 3.3 Estimates of completeness of registration of births in the year of birth for the years 1996-2001

Year	1996	1997	1998	1999	2000	2001
Completeness	25.7%	27.0%	24.0%	30.5%	36.6%	42.8%

Completeness of registration of births by the end of the year of birth has been improving over time. To extrapolate the trend in completeness of birth registration for years after 2001, a logistic curve is fitted to the estimates of completeness. This is done using the LOGISTIC workbook prepared by the U.S. Census Bureau (1994). The lower bound asymptote is set at 10 per cent and the upper bound asymptote is set at 100 per cent. It is assumed that at least 10 per cent of births in a particular year are registered during the same year, hence an asymptote of 10 per cent is used for the lower bound. There is no literature to support the adoption of this asymptote, but it is adopted because it is arbitrarily low and it is a suitably conservative measure to ensure that the logistic fit does not rise too rapidly over the projection period. This further ensures that true numbers of births are not underestimated.

Estimates of the true numbers of births for the years 2002-2011 are computed by dividing the observed births by the corresponding estimate of completeness. However, due to the two-year lag in the availability of death data from Stats SA, by the time mid-year estimates are being produced for a particular year, "Year 1" births would be available for that year. Thus, "Year 1" births for the years 1996-2001 are used to fit a logistic curve to estimate "Year 1" completeness for 2002 births. Observed "Year 0" births for each year are initially used to estimate the true number of births for that year using the fitted estimate of completeness for "Year 0". When "Year 1" observed births for that particular year are assumed to have become available, they are used to improve the estimate of the true number of births for that year. This procedure can be used until the estimate of the true numbers of births obtained using "Year 0" completeness estimates becomes stable. Estimates of completeness using "Year 0" births become stable when "Year 0" births are now sufficiently complete due to improvements in vital registration, such that the true numbers of births estimated using "Year 0" births become similar to those estimated using "Year 1" births. Asymptotes for estimating completeness by the end of the year after the year of birth ("Year 1") are set at 20 per cent for the lower bound and 100 per cent for the upper bound. A slightly higher lower

asymptote was used to allow for at least 10 per cent of births being registered during the year of birth, and at least 10 per cent births being registered at the start of the second year after the year of birth.

The following is an illustration of how the method is applied. The number of births in 2002 are initially estimated by dividing 2002 births registered in 2002 by the "Year 0" estimate of completeness for 2002 to obtain and estimate of 1 269 515 births. However, this initial estimate of births based on "Year 0" births is not quite satisfactory, given that there is uncertainty surrounding how quickly estimates of completeness increase from "Year 0" to "Year 1" (39.68 per cent to 67.06 per cent for 2002). However, this is probably a consequence of the logistic fit and the indirect estimates of completeness of birth registration adopted to produce the fit. Nonetheless, because there is insufficient data to obtain any better estimate of the completeness of birth registration, the estimates from the logistic fit had to suffice in estimating the true numbers of births for the projection years. Since estimates of "Year 1" are essentially more complete (more births registered) than "Year 0" estimates, the number of births estimated for 2002 using "Year 1" estimated completeness are assumed to be more reliable (1 121 105). The final estimates of true numbers of births for a particular year are thus those that are estimated using "Year 1" estimates of completeness.

3.4 Deaths

Death registration data were also obtained from Stats SA's website. The deaths are given by year of registration and age at death. This allows for the tabulation of deaths that would be available at the time Stats SA would publish registered deaths. In the past Stats SA have published registered deaths for a particular year with a time lag of approximately two years, meaning mid-year estimates produced using registered deaths would be available at least two years after a reference date. Trends in deaths registered in previous years can be used to extrapolate deaths during the projection year and estimate deaths which can be used to produce preliminary mid-year estimates.

3.5 Completeness of death registration

As was the case with births, it is also necessary first to adjust deaths for incompleteness prior to using them in the cohort component projection method to produce mid-year population estimates for the projection period 2001 to 2011.

3.5.1 Estimating completeness of adult death registration

Estimates of completeness of adult death registration for the projection period are obtained by fitting a logistic curve to estimates of completeness from previous research. Dorrington, Moultrie, and Timæus (2004) estimated completeness of adult death registration for the period 1996-2001 using population distributions by age and sex from the 1996 and 2001 censuses together with deaths from vital registration. They used the Generalised Growth Balance method as proposed by Hill (1987) to obtain their estimates. They estimated that national completeness of death registration was 83.4 per cent for males and 84.5 per cent for females for the 1996-2001 period.

Further, by assuming that completeness in death registration increased linearly over the period such that mortality of those aged over 65 remained constant for the period, they produced estimates of completeness for each year (personal communication, Professor Rob Dorrington). These estimates are used to fit a logistic curve using the LOGISTIC workbook from the Population Analysis Spreadsheets developed by the United States Census Bureau (U.S. Census Bureau 1994), setting asymptotes at zero per cent for the lower bound and 95 per cent for the upper bound for both males and females. Zero per cent is used as a lower bound, unlike the ten per cent for births, because it is arbitrarily low and it is also a conservative measure to ensure that the fit does not rise too rapidly. Setting the lower bound at ten per cent results in the fit rising to rapidly. Although completeness may ultimately reach 100 per cent, the upper bound was set to 95 per cent in order to restrict the increase in completeness during the projection period. Extrapolation of the assumption of linear increase in adult death registration completeness for the period 1996 to 2001 adopted by Dorrington, Moultrie, and Timæus (2004) makes the logistic fit rapidly increase such that completeness by 2011 would be too close to 100 per cent, which is unlikely in reality. In any case, the reviewed literature notes that completeness of adult death registration is estimated to have been around 90 per cent during the early 2000's and just below 95 per cent by 2006 (Dorrington and Bradshaw 2011, Machemedze 2009). This also informed the decision to set the upper bound to 95 per cent.

3.5.2 Estimating completeness of infant and child death registration

Completeness of death registration for those aged 0-4 last birthday is estimated in a way similar to that of adult death registration. Annual estimates of completeness of infant and child death registration are obtained from Darikwa and Dorrington (2011). Even though the research uses data that would otherwise not have been available for producing mid-year estimates for the years 2002 to 2006 (before data from the 2007 CS

was available), only the components of their research that refer to past data from 2001 and prior years are used.

They estimated completeness by comparing deaths implied by empirical estimates of infant and child mortality to deaths of infants, 1-4 year olds and under 5 year olds from vital registration. They used registered deaths from vital registration, the age distribution of 0-4 year olds from the 2001 census and the 2007 Community Survey, and past research based on the 1998 South African DHS. Using these data, they applied an innovative multi-stage method to derive estimates of infant and child mortality (Darikwa and Dorrington 2011).

Annual estimates of completeness of infant death registration, and 1-4 year olds death registration from their research for the years 1996-2000 are used to fit logistic curves to estimate completeness for the projection period. New logistic curves are fitted for the years 2007-2011 in the projection period, which incorporated data from the 2007 Community Survey since the data would have been available to use in estimating completeness of death registration.

Logistic curves are fitted using the LOGISTIC workbook. For infants, the lower bound is set at zero per cent whilst the upper bound was set at 95 per cent. For the 1-4 year olds, the lower bound was set to zero per cent and the upper bound to 70 per cent. For 0-4 year olds, the lower bound is set to zero per cent and the upper bound to 85 per cent. As was done for estimating completeness of adult death registration, the upper bounds for the three categories are set below 100 per cent even though ultimately, 100 per cent may be possible. This restriction is done to ensure that completeness would not get too close to 100 per cent within the 2001-2011 projection period. The lower bounds are set at zero per cent for the same reason as that for the fit for adult death registration. The choice of upper bounds for the asymptotes is informed by the literature reviewed, particularly Darikwa and Dorrington (2011) who estimated completeness to increase from 43 per cent to 89 per cent for infants, from 43 per cent to 57 per cent for 1-4 year olds and from 44 per cent to 78 per cent for 0-4 year olds during the period 1996-2006. This is done for the curve fitted using data available prior to the 2007 Community Survey.

For the curves fitted using the 2007 Community Survey (CS) data as well, the upper bound is maintained at 95 per cent for infants. For the 1-4 year olds, the upper bound is increased to 80 per cent whilst a restriction for completeness by 2011 is set to be below 70 per cent. For the 0-4 year olds, the upper bound is increased to 90 per cent,

whilst restricting completeness by 2011 to be below 85 per cent. As was done when fitting curves with data available prior to the 2007 Community Survey, the choice of asymptotes and the restrictions are made with the same logic.

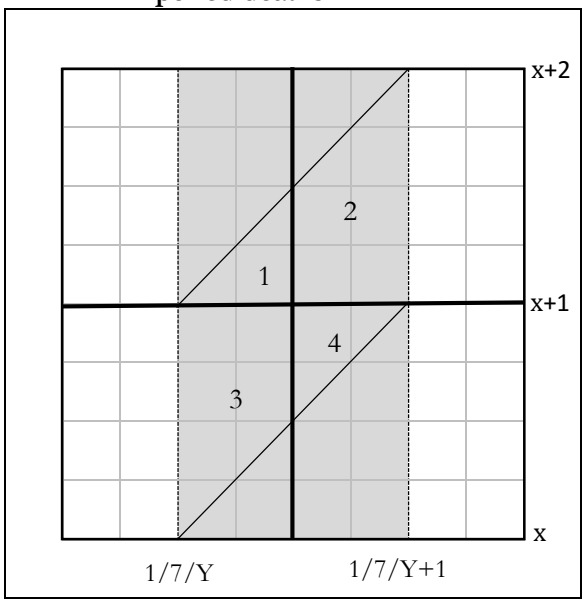
3.6 Estimating deaths by cohort between 1 July Year (Y) and 30 June Year (Y+1)

Registered deaths in each year are adjusted for incompleteness by dividing deaths at each age by the corresponding estimate of completeness. Completeness of adult death registration is assumed to apply also to ages 5-14 last birthday. This assumption is usually guided by the logic that mortality rates are normally very low at these ages, hence any errors arising from this assumption are likely to be small. Deaths from vital registration are available by age at death. In the projection using the cohort component method, deaths are subtracted from the mid-year population by age and sex. In order for this to work correctly, it is necessary to adjust the age at death to the age the deceased would have been on 30 June in Year (Y+1). This is necessary because one projects cohorts, so adjusting deaths to the age the deceased would have been by 30 June in Year (Y+1) ensures that we would be subtracting the correct number of deaths.

The age the deceased would have been on 30 June in Year (Y+1) can be estimated if one has the birthday of the deceased, which is available to Stats SA, but not available publicly. This would be done by creating a variable which calculates the age the deceased would have been by 30 June by subtracting the birthday of the deceased from 30 June Year (Y+1) in century month code (CMC) in STATA. Unfortunately, the data from Stats SA for the years 2001 to 2008 do not include the date of birth of the deceased. The data sets only include the date of death and the age the deceased was at death. However, by assuming that birthdays and deaths are evenly distributed throughout the year, it is possible to approximate the cohort deaths using the period deaths. Figure 3-1 shows a lexis diagram demonstrating how this approximation can be done.

The deaths aged x last birthday and $x+1$ last birthday that we have from Stats SA are represented by the four shaded rectangles, two for each year. The cohort deaths we require are represented by the parallelogram enclosed by the diagonal lines. If we assume that deaths are evenly distributed over age and time, then we may estimate the cohort deaths as follows.

Figure 3-1 Lexis diagram demonstrating how cohort deaths can be estimated using period deaths



We are interested in the cohort of deaths aged *x* last birthday in year Y in the enclosures labelled 1 and 3. 75 percent of half of the deaths aged *x* last birthday in year Y can be expected to be aged *x* last birthday as at 1 July year Y. The cohort of deaths in the enclosure labelled 1 on Figure 3-1 are approximately 25 per cent of the half of the deaths aged *x+1* as at 1 July year Y.

For the deaths in year Y+1, we are interested in the deaths occurring before the first of July. Using the same approach described for deaths in year Y, 75 per cent of half of the deaths aged *x+1* last birthday in year Y can be expected to be aged *x+1* last birthday as at the first of July in year Y+1. This cohort of deaths is represented by the enclosure labelled 2 on Figure 3-1. Similarly, the cohort of deaths in the enclosure labelled 4 can be approximated as 25 per cent of the half of the deaths aged *x* last birthday as at the first of July year Y+1.

The final approximation for the required cohort of deaths can then be given as follows:

$$\begin{aligned}\text{Required cohort of deaths} = \quad & 0.75*(0.5*D(x,Y)) + 0.25*(0.5*D(x+1,Y)) \\ & + 0.75*(0.5*D(x+1,Y+1)) \\ & + 0.25*(0.5*D(x,Y+1)) \\ = \quad & 0.375*D(x,Y) + 0.125*D(x+1,Y) \\ & + 0.375*D(x+1,Y) + 0.125*D(x,Y+1)\end{aligned}$$

where:

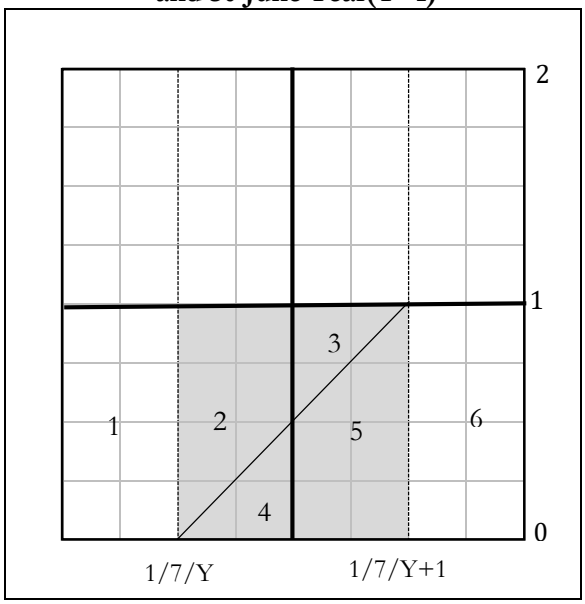
D(*x*, Y) = deaths aged *x* last birthday in year Y

The above approximation is used to estimate the required cohort deaths for those aged two last birthday and above. For those aged zero last birthday at death during a particular year, a greater proportion would have been born in that year as well hence another approximation in necessary.

3.6.1 Estimating infant cohort deaths

Registered infant deaths from Stats SA are given by age last birthday at death, the domain of which may be represented in a lexis diagram in the form of a unit square (single age and single unit of time). However, to estimate infant cohort deaths from these period deaths, it is necessary to split the data in the unit square into two lexis areas. An illustration is given in Figure 3-2.

Figure 3-2 Estimating infant cohort deaths from period deaths between 1 July Year(Y) and 30 June Year(Y+1)

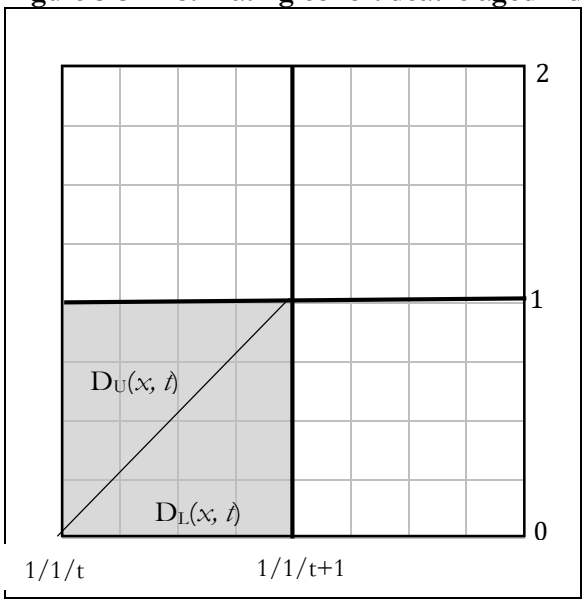


What we have from Stats SA are the deaths in year Y aged [0, 1), represented on the lexis diagram by the areas marked 1, 2 and 4; and the deaths in year Y+1 aged [0, 1) represented on the lexis diagram by the areas marked 3, 5 and 6. However, for the projection we are concerned with the deaths from 1 July year Y to 30 June year Y+1, represented on the lexis diagram by the shaded areas marked 2, 3, 4 and 5. Therefore, assuming that infant deaths are uniformly distributed over the two calendar years, Y and Y+1, we may combine the deaths in the second half of year Y with the deaths in the first half of year Y+1. The period deaths we would have are those is the areas marked 2, 3, 4 and 5; and from these, we require the cohort deaths from births, i.e. those in areas 4 and 5. Deaths are not evenly distributed across age and time because mortality falls

rapidly with age over the first year of life. Estimating the cohort deaths in the areas marked 4 and 5 thus needs to take this into account.

Wilmoth, Andreev, Jdanov et al. (2007) developed a regression equation for estimating cohort deaths from period using data from the Human Mortality Database. An illustration of how they developed their equation is given in Figure 3-3.

Figure 3-3 Estimating cohort deaths aged x during Year (t) from period deaths



Relating our projection to their formulation, our period would be starting from the middle of year Y up to the middle of year Y+1, corresponding to their period starting at the beginning of year (*t*) up to the end of year (*t*). The cohort deaths we require, in the areas marked 4 and 5 in Figure 3-2, would thus be corresponding to the lower area marked $D_L(x, t)$ in Figure 3-3.

Estimating the cohort deaths in the lower area marked $D_L(x, t)$ only requires an approximation of the proportion of deaths in this area. This proportion is given as follows.

$$\pi_d(x,t) = \frac{D_L(x,t)}{D_L(x,t) + D_U(x,t)}$$

where

$\pi_d(x,t)$ = the true proportion of deaths aged x last birthday at time t in the lower triangle

$D_L(x,t)$ = the true number of deaths aged x last birthday at time t in the lower area

$D_U(x,t)$ = the true number of deaths aged *x* last birthday at time *t* in the upper area

The true numbers of deaths in the both the lower and upper areas are usually not known, hence it becomes necessary to get an estimate of the proportion of deaths in the lower area, $\hat{\pi}_d(x,t)$. This proportion may then be used to estimate the numbers of deaths in the lower and upper triangles from period deaths as follows.

$$\hat{D}_L(x,t) = \hat{\pi}_d(x,t) * D(x,t)$$

$$\hat{D}_U(x,t) = \left[1 - \hat{\pi}_d(x,t)\right] * D(x,t)$$

where

$D(x,t)$ = the observed period deaths aged *x* last birthday over the year *t* to *t+1*

Wilmoth et al. (2007, 12) developed a regression equation[2], which differs by sex, to approximate the true proportion of deaths in the lower area. Their equation is adapted as follows for our projection:

$$\hat{\pi}_d(x,t) = 0.4710 + \hat{\alpha}_x^F + 0.7372 * \left[\pi_b(x,t) - 0.5\right] - 0.0112 * \ln(IMR(t)) - 0.0688 * \ln(IMR(t)) * I(x=0) - 0.268 * \ln(IMR(t)) * I(x=1)$$

$$\pi_b(x,t) = \frac{B(t-x)}{B(t-x) + B(t-x-1)}$$

$$IMR(t) = \frac{D(0,t)}{\frac{1}{3}B(t-1) + \frac{2}{3}B(t)}$$

where

$\hat{\alpha}_x^F$ = the estimated age effects for the female version of the equation (given)

$\pi_b(x,t)$ = the birth proportion at time *t*

$I(\bullet)$ = one if the logical statement in the brackets is true, zero

[2] $\hat{\pi}_d(x,t) = 0.4710 + \hat{\alpha}_x^F + 0.7372 * [\pi_d(x,t) - 0.5] + 0.1025 * I(t = 1918) - 0.0237 * I(t = 1919) - 0.0112 * \ln(IMR(t)) - 0.0688 * \ln(IMR(t)) * I(x = 0) - 0.268 * \ln(IMR(t)) * I(x = 1) + 0.1526 * \left[\ln\left[\left(IMR(t)\right) - \ln(0.01)\right] * I(x = 0) * I(IMR)(t) < 0.01\right]$

otherwise

IMR(t) = the infant mortality rate (both sexes combined) at time t

B(t) = number of births (both sexes combined) during year t

In general, the regression equation estimates the proportion of deaths in the lower area, marked $D_L(x, t)$, by capturing the effects of trends in infant mortality. The above equation is used to estimate cohort deaths for the ages 0 last birthday and 1 last birthday.

3.7 Migrants

International migration for the projection will be estimated using estimates of annual numbers of migrants during the 1996-2001 inter-censal period. Numbers of net international immigrants are estimated using data from the ten per cent samples of the 1996 and 2001 censuses obtained from Stats SA's website[3]. Specifically, place of birth tabulations are extracted from the census data which gives age distributions of the foreign-born population as at the 1996 and 2001 censuses. These age distributions are used to estimate net international migration during the inter-censal period by applying the residual method.

Numbers of South African emigrants during the inter-censal period are also estimated using the residual method. Age distributions of the South African-born (SAB) populations in five of the most popular destinations of South African emigrants (Australia, United Kingdom, New Zealand, Canada, United States of America) in 1996 and in 2001 are obtained from various sources. For Australia and New Zealand, age distributions of South African-born individuals in 1996 and 2001 were obtained from the United Nations Statistics Division (UNSD) as given in the UN Data website[4]. The website gives a database of foreign-born populations by country/area of birth, age and sex. The data are consolidated from various sources including censuses and surveys in various countries. Data for the other countries were not available from the UNSD database hence they were obtained from other sources.

Estimates of the SAB in UK in 1996 were obtained from the Office for National Statistics' website[5]. However, only total numbers of the SAB were available for 2001 and 2011. To estimate the total SAB population in 1996, the growth rate of the SAB

[3] www.statssa.gov.za/

[4] http://data.un.org/Data.aspx?d=POP&f=tableCode%3A44

[5] http://www.ons.gov.uk/ons/index.html

between 2001 and 2011 is extrapolated by assuming exponential growth of the SAB population during the period. The SAB population is estimated using the following equation:

$$SAB_{1996} = SAB_{2001} * e^{-5r}$$

where:

r = the compounded annual growth rate of the aggregate SAB population in the UK between 2001 and 2011.

Of course this gives only a rough estimate of the SAB population in the UK in 1996, but this has to suffice because no other estimate of the SAB population in 1996 is available. Since the estimated number of the SAB for 1996 is only a total, the number was disaggregated using the 2001 Australian age and sex distribution of the SAB in Australia. This assumption is that the age distribution of the SAB population in Australia is accurate and sufficiently resembles the age distribution of the SAB population in the UK. This is a key assumption as the accuracy of the estimate of net emigration of the SAB population mainly depends on it. However, the assumption is nonetheless crude because people probably migrate to different destinations for different reasons, hence the age distributions may not be similar. For the SAB population in UK in 2001, the estimated age and sex distribution was obtained from the UNSD database.

The SAB population in the USA in 1996 and 2001 was estimated using data obtained from the United States Census Bureau's (USCB) website[6]. The data were only available for the years 1990 and 2000. The SAB population in 1996 was obtained by interpolating using the growth rate of the SAB population during the period 1990-2000, assuming exponential growth. The following equation was used:

$$SAB_{1996} = SAB_{1990} * e^{5r}$$

where:

r = the compounded annual growth rate of the aggregate SAB population in the USA between 1990 and 2000

The SAB population in the USA in 2001 was estimated by extrapolating the growth rate. For Canada, the SAB population in 1996 was estimated using data obtained

[6] http://www.census.gov/#

from Statistics Canada's website[7]. The data were available by age and sex. For 2001, the data on SAB population was obtained from the UNSD database.

Estimated numbers of the total SAB population in 1996 and 2001 were then obtained by summing the numbers from the five countries by age and sex.

3.8 Estimating net numbers of FB immigrants

The residual method, as adapted by Bashir and Robinson (1994) is used to estimate net numbers of immigrants and net numbers of emigrants.

Expected survivors of the FB population during the 1996-2001 inter-censal period are estimated using survival ratios obtained from the ASSA2008 Demographic and AIDS Model. The age distribution, in five year age groups, of the expected FB population in 2001 was estimated as follows:

$$ {}_5P_{x+5}^{FB2001} = {}_5S_x * {}_5P_x^{FB1996} $$

where:

${}_5P_{x+5}^{FB2001}$ = the expected FB population as at the 2001 census in the age group $[x+5, x+10)$

${}_5S_x$ = the survival ratio for the 1996-2001 inter-censal period of those in the age group $[x, x+5)$ as at the 1996 census

${}_5P_x^{FB1996}$ = the observed FB population as at the 1996 census in the age group $[x, x+5)$

The difference between the observed (enumerated) FB population as at the 2001 census and the expected (estimated survivors) FB population is then taken as the estimate of net number of surviving FB migrants as at the time of the 2001 census.

3.8.1 Estimating net numbers of SAB emigrants

SAB emigrants are estimated in the same way as FB immigrants but using the censuses in overseas countries. Survival ratios of the White population group were obtained from the ASSA2008 Demographic and AIDS Model. It is assumed that only white South Africans emigrated during the 1996-2001 inter-censal period due to the level of uncertainty surrounding emigration from South Africa. The difference between the expected (estimated survivors) SAB population abroad in 2001 and the observed

[7] http://www.statcan.gc.ca/start-debut-eng.html

(enumerated) is then taken as the estimate of net number of surviving SAB migrants as at the 2001 census.

3.8.2 Estimating annual net numbers of international migrants for the projection

The projection requires annual numbers of migrants in single year ages because the base population is projected in single year increments. For the purpose of this projection, annual numbers of migrants are estimated by dividing the inter-censal numbers of migrants by five. Since we were working with five-yearly age groups, we also disaggregate the numbers in five-yearly age groups into single year ages. Beers' six-parameter disaggregation coefficients are used to disaggregate the five-yearly age groups into single year ages (Beers 1945).

Annual net migration for the 2002 to 2011 projection period is assumed to remain the same as that of the 1996 to 2001 inter-censal period. This assumption is hardly sufficient in estimating annual flows of migrants in and out of South Africa, given the posterior knowledge that there were significant inflows of Zimbabweans immigrating to South Africa to escape political and economic turmoil in Zimbabwe. However, because there is no registration system for migrants in South Africa, no annual and nationally representative survey which could be used to estimate migration, and no reliable administrative source of data that could be used to estimate migration, the estimates of migration from the 1996 to 2001 inter-censal period would have to suffice.

3.9 Projecting the base population from one year to the next

Mid-year population estimates are calculated for the years 2002-2011. The cohort component method is used to project the base population, derived from the 2001 census, from one year to the next. Population growth components, namely fertility, mortality and migration, are derived as described in the preceding sections. The base population is estimated as at 30 June 2001.

Mid-year estimates from 30 June year Y to 30 June year Y+1 are calculated as follows. Starting with a base population by single age; increase the ages by one, subtract the estimated cohort deaths and estimated net numbers of emigrants during the 12 month period from 1 July year Y to 30 June year Y+1, and finally, add estimated net numbers of immigrants and births occurring during the 12 month period. The resulting series of mid-year estimates are then compared with Stats SA's series and the 2011 census age distribution to evaluate their quality and consistency

3.10 Summary of assumptions and their consequences on estimates

Table 3.4 gives the assumptions that were used in the method and their possible consequences on the mid-year estimates.

Table 3.4 Method assumptions and their consequences

	Assumption	*Reason*	*Consequence on MYEs*
1	Constant annual exponential growth of the population between 1996 and 2001	Deriving the 2001 base population	Possible over estimate or under estimate of the base population, but probably negligible since the assumed exponential growth rate is only extrapolated backwards by 100 days from the census date to the middle of 2001. Consequence on MYEs thus probably minimal
2	Fertility not expected to have changed significantly over the period 1996-2001	Estimating completeness of birth registration	If fertility had changed significantly over the period, then this assumption would affect estimates of completeness of birth registration. This would subsequently either over inflate or under estimate true numbers of births, ultimately affecting ages 0-10 last birthday in the 2011 MYEs. Since fertility is not changing very rapidly, the extent of any error is not expected to be big.
3	At least 10 per cent of births in a particular year are registered during the same year	Estimating completeness of birth registration	This affects the logistic fit of completeness of birth registration and would have the same effect as that of (2). The effect of the error resulting from this assumption may be significant owing to the substantial numbers of births. If the assumption results in too low estimates of completeness of birth registration, then births will be over inflated resulting in the estimated MYEs being much higher than actual. If the assumption results in too high estimates of completeness of birth registration, then births will be underestimated resulting in the estimated MYEs being much lower than actual.
4	At least 20 % of births in a particular year would have been registered by the following year	Estimating completeness of birth registration	This affects the logistic fit of completeness of birth registration and would have the same effect as that of (2) and (3) above.

	Assumption	*Reason*	*Consequence on MYEs*
5	Estimated completeness of adult death registration applies to ages 5-14 last birthday	Estimating true numbers of adult deaths from period deaths by cohort between 1 July year Y and 30 June year Y+1	This would affect the estimates of true numbers of deaths of those aged 5-14 last birthday, but such an effect is not expected to be very serious. It would be difficult to state the direction of error in this regard as no literature was identified that discusses this
6	Birthdays and deaths are evenly distributed throughout the year	Estimating true numbers of deaths from period deaths by cohort between 1 July year Y and 30 June year Y+1	This affects estimates of true numbers of deaths by age for all ages. Other months probably have greater numbers of deaths and birthdays but the errors in the estimates would most likely balance each other and the effect on the MYEs would be minimal
7	Infant deaths are evenly distributed over the two years Year(Y) and Year(Y+1)	Estimating true numbers of infant deaths by cohort from period deaths between 1 July year Y and 30 June year Y+1	Same as for (6) above
8	Age distribution of SAB immigrants in the UK is similar to the age distribution of SAB immigrants in Australia	Estimating net numbers of SAB immigrants by age group	The assumption is crude and probably distorts the age distribution of SAB immigrants in the UK as immigrants move to different countries for different reasons. The effect on the MYEs is uncertain, but probably not very significant since such numbers are most likely relatively small compared to the population resident in South Africa.
9	Only White South Africans emigrated during the 1996-2001 period	Using survival ratios of the White population group, as extracted from ASSA 2008, to estimate net numbers of South African emigrants	Same as for (18) above. If there was a significant number of non-White SAB emigrants during the period, then this would result in an underestimate of net emigrants, and subsequently an underestimate of the MYEs. This is so because the survival ratios for the White population group are marginally higher than the survival ratios for the aggregate South African population. The higher survival ratios result in higher numbers of surviving SAB emigrants during the period, thus the difference between the resident SAB emigrants in 2001 and the surviving is lower than it would actually be, thus underestimating net SAB emigrants. However the effect is quite small since numbers of emigrants are not quite substantial.

	Assumption	*Reason*	*Consequence on MYEs*
10	Estimated net numbers of FB immigrants and SAB emigrants in the 5 year period between 1996 and 2001 were evenly distributed during the 5 years	Estimating annual rate/numbers of net international migration	Crude assumption and probably distorts the estimates of net numbers of immigrants per annum. The period after 1996 was a transition period for South Africa, hence annual numbers of net immigrants cannot have been evenly distributed. The effect on MYEs is probably a not too significant under estimate of net numbers of immigrants. The mostly affected age groups would be 15-29 last birthday
11	Annual net international migration for the period 2002-2011 remained the same as annual net international migration for the period 1996-2001	Estimating net international migration per annum	Same as for (10) above

4. RESULTS AND ANALYSIS

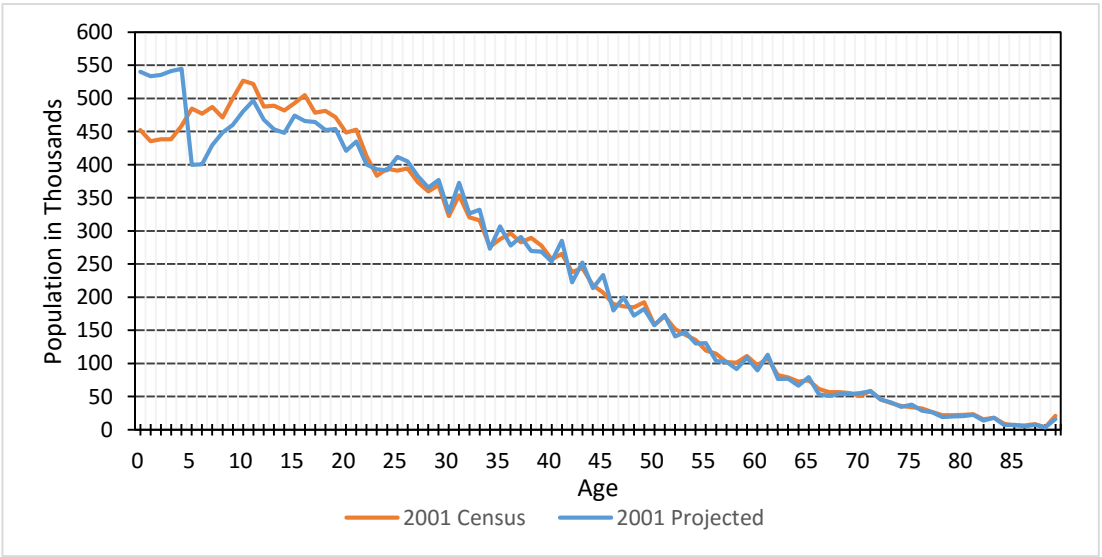
This chapter presents the results of the methods described in Chapter 3. These results include output from the demographic analysis and adjustment of the base population, completeness estimates of birth and death registration, estimates of international migration and the post-censal mid-year estimates produced using administrative data for the years 2002-2011. Results will be presented together with analyses of the results.

4.1 Demographic analysis of the base population

This section presents the results from the demographic analysis of the 2001 base population derived from the 2001 census. The comparison of the observed and expected population distributions is shown in Figure 4-1 for males. The comparison for females is similar and it is available in the appendix. Visual inspection of the census age distribution shows minor peaks at ages ending with 1 (e.g. 21, 31, 41, 51, 61, 71, etc.) which suggests digit preference for these ages. This may be interpreted as preference for birthdays in years ending with 0 (e.g. 1930, 1940, 1950, 1960, 1970, etc.).

There is a notable shortfall in the population aged between zero and five, with the population aged one being approximately 11 per cent less than the population aged five. This suggests that children under five were undercounted in the 2001 census relative to the population at older ages. Comparing those aged 0-4 last birthday in the census with the projected 0-4 last birthday also suggests an undercount in 1996. The 0-4 last birthday in the census are approximately 16 per cent less than the projected 0-4 last birthday.

Figure 4-1 2001 Projected census distribution and 2001 census distribution, males



It is possible that the fertility regime used in the projection is higher than the actual fertility, resulting in the difference between the projected 0-4 last birthday and the census 0-4 last birthday. However, the fertility assumptions used by the projection are similar to other empirical estimates of fertility produced by other researchers using the 1996 and the 2001 censuses (e.g. Moultrie and Timæus 2003, Moultrie and Dorrington 2004). Table 4.1 gives a comparison of the total fertility rates (TFRs) used in the projection with those from previous research.

Table 4.1 Comparison of TFRs for the period 1996-2001

Year	*Source*	
	ASSA2008	*Previous research*
1996	3.30	3.23
1997	3.20	3.15
1998	3.09	3.08
1999	2.99	3.00
2000	2.89	2.92
2001	2.82	2.84

Source: Derived from the ASSA2008 AIDS and Demographic model; Moultrie and Timæus (2003); Moultrie and Dorrington (2004)
Note: The 1996 TFR is as estimated by Moultrie and Timæus (2003) using data from the 1996 census and the 2001 TFR is as estimated by Moultrie and Dorrington (2004) using data from the 2001 census. TFRs for the intervening years are calculated by linear interpolation, assuming fertility changed linearly over the period 1996-2001.

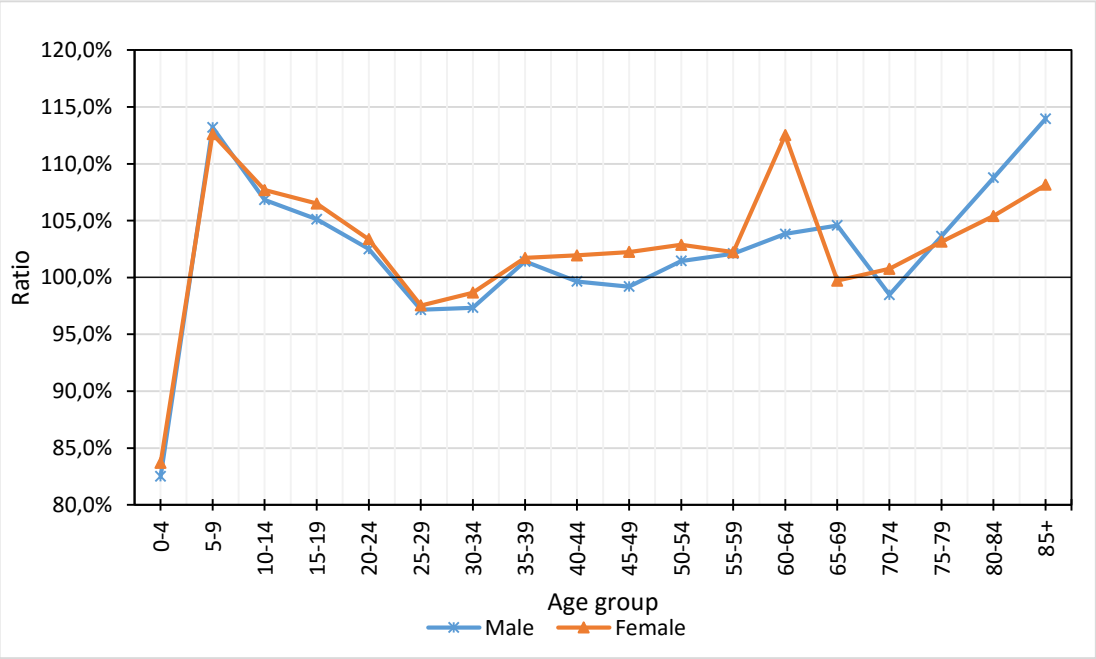
TFRs are the most common measure of period fertility and they give the number of children a woman is expected to have if she were to live to the end of her reproductive life span experiencing the age-specific fertility rates observed in that time period. The TFRs from ASSA2008 for the period 1996-2001 are similar to those estimated by other researchers, although they are slightly higher for earlier years. However, these are all empirical estimates and the difference between them is not pronounced, hence the fertility assumptions used by the projection are assumed to be sufficiently accurate. Considering this assumption, the 2001 census population aged 0-4 last birthday is approximately 20 percent less than the projected 0-4 last birthday. This difference is too vast to be explained by exaggerated fertility rates alone. Thus, it would it would appear that there was a significant undercount of those aged 0-4 last birthday as at the 2001 census.

Further comparison of the census and projected age distributions shows marked differences for those aged 5-22 last birthday. The projected population aged 5-22 last birthday is about 7 percent less than the population aged 5-22 last birthday as at the 2001 census. This suggests that there was an undercount of the population aged 0-17

last birthday in the 1996 census, relative to the 2001 census, since this cohort is the one projected to be 5-22 years last birthday in the 2001 census. The biggest difference is in the 5-9 last birthday age group, also indicated in Figure 4-2 below. This suggests an undercount in the 0-4 last birthday age group in the 1996 census. This undercount is approximately 13 per cent, which is marginally lower than the undercount suggested by the comparison for the 0-4 last birthday as at the 2001 census which is about 16 per cent. The difference may also suggests that migration for the age group was not sufficiently allowed for in the projection, although this is unlikely to be significant as migration is low at these ages. It is also possible that the age-specific mortality regime assumed by the projection for the 0-17 age range in 1996 was higher than the actual, but this would only account for a small proportion of the difference.

Some additional insights about the base population can be obtained from analysing the percentage ratios of census numbers divided by projected numbers. Figure 4-2 shows these ratios by five year age groups.

Figure 4-2 Ratios of 2001 census/projected



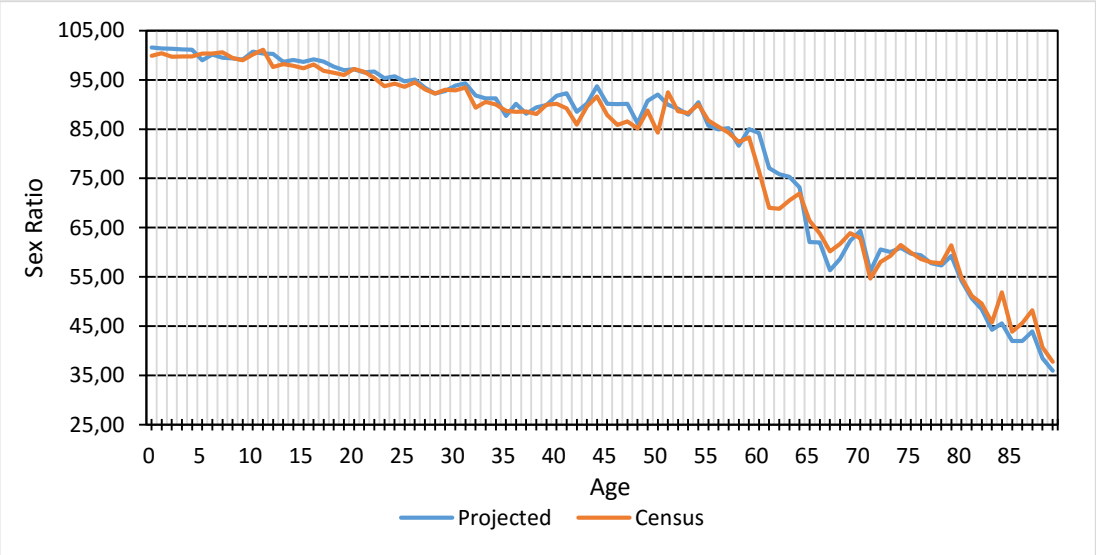
Except for the outlying ratio for the 0-4 last birthday age group (undercount of 0-4 in 2001 census), the ratios are above 100 per cent for most of the age groups. This indicates an excess in the 2001 census relative to the 1996 census, assuming that fertility, mortality and migration assumptions in the projection are correct. It suggests a general undercount in the 1996 census or an over-count in the 2001 census. A proportion of the

shortfall in the projected population could be due to migration insufficiently accounted for in the projection.

For ages 80 and older, there is a generally increasing trend in the ratios. This could be due to too high mortality assumed for the older ages by the projection, but is more likely to be age exaggeration in the 2001 census. There is a notable peak in the 60-64 last birthday age group for females, and in the 65-69 last birthday age group for males, though it is less pronounced. This is probably a result of age exaggeration to get pension, given that the retirement age in South Africa was 60 for females and 65 for males.

Analysing the ratios of the number of males divided by the number of females by single age can also give further insights about the population. Figure 4-3 shows the sex ratios (males/females*100) from the census and the projected populations by single ages.

Figure 4-3 2001 census sex ratios and projected sex ratios



Comparison of the sex ratios for ages 0-4 last birthday show that the projected sex ratios are less than those from the census, which indicates that there was either a relative undercount of female to male children in 2001, or that the sex ratio at birth used in the projection was too low. However, the differences are not pronounced. For the ages 12-42 last birthday, the census sex ratios are slightly less than the projected. This indicates a slight undercount of males in the ages 12-42 last birthday relative to females in the 2011 census. To know definitively whether it was an excess or deficit in either of the genders, further investigations, not within the scope of this research, would be necessary. The sex ratios gradually decline at the older ages with the decline accelerating at ages 60 and

over, both projected and census, which is expected since male mortality is expected to be much higher for males at these ages.

4.2 Demographic adjustment of the base population

Adjustments to the base population are made using cohort analysis adjustment which aims to correct any demographically implausible features in the data.

4.2.1 Adjustment of the 0-4 last birthday age group

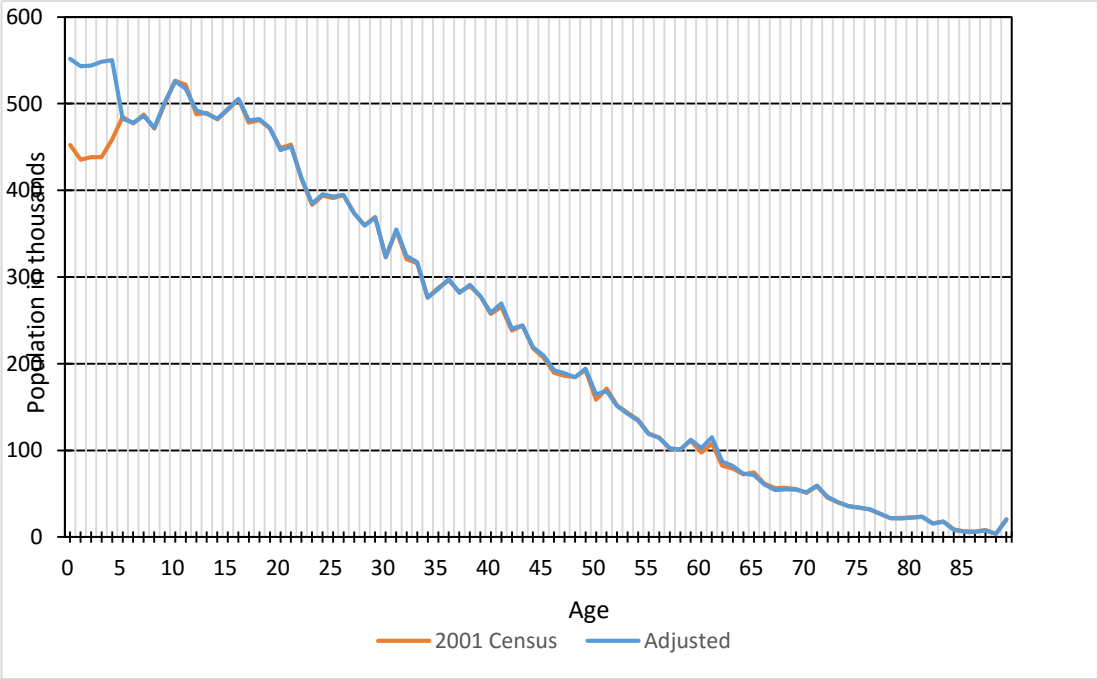
For this age group, comparison of the census numbers to the projected numbers in the 0-4 last birthday age group indicates an undercount in the census of approximately 20 per cent. To adjust for this undercount, it is assumed that the fertility, mortality and migration allowed for in the projection were sufficiently accurate. The projected population aged 0-4 last birthday are thus preferred over the census 0-4 last birthday and adopted into the base population.

4.2.2 Adjustment of sex ratios

The census sex ratios show some sharp fluctuations at various ages. For instance, the sex ratio falls sharply from 83 for age 59, to 68 for age 62, then gradually increase to about 72 for age 64. This is unexpected and determining the cause of this fluctuation would require in-depth research into the distribution of males and females in the affected ages as at the census. Similar fluctuations are observed for ages between 82 and 89 last birthday. This could be due to age exaggeration by the elderly. In any case, sex ratios for most ages are only slightly different between the census population and the projected population. However, the census sex ratios have some unexpected fluctuations, while the projected sex ratios seem more stable with minor fluctuations. It is thus decided to prefer the projected sex ratios over the census sex ratios.

Projected sex ratios by single age are used to redistribute the census population by single age. Figure 4-4 shows a comparison of the census male population distribution to the adjusted male population distribution. There are marked differences for the 0-4 last birthday age group. The differences for the rest of the ages are minor.

Figure 4-4 Comparison of the census male population distribution to the adjusted male population distribution



4.3 Completeness of birth registration

Figure 4-5 shows the estimates of completeness obtained using the LOGISTIC workbook. The results indicate that the completeness of birth registration of births in the year of birth (Year 0) increases to reach about 80 per cent by 2011.

These estimates seem plausible, given that the Department of Home Affairs has, over the years, been making concerted efforts to increase birth registration (Statistics South Africa 2009a).

Observed registered births by "Year 0" and "Year 1" for the years 2002-2011 and the corresponding estimates of the true numbers of births estimated using completeness estimates for "Year 0" and "Year 1" are given in Table 4.2.

Figure 4-5 Estimates of completeness (Year 0) – Observed and fitted (LOGISTIC), 1996-2031

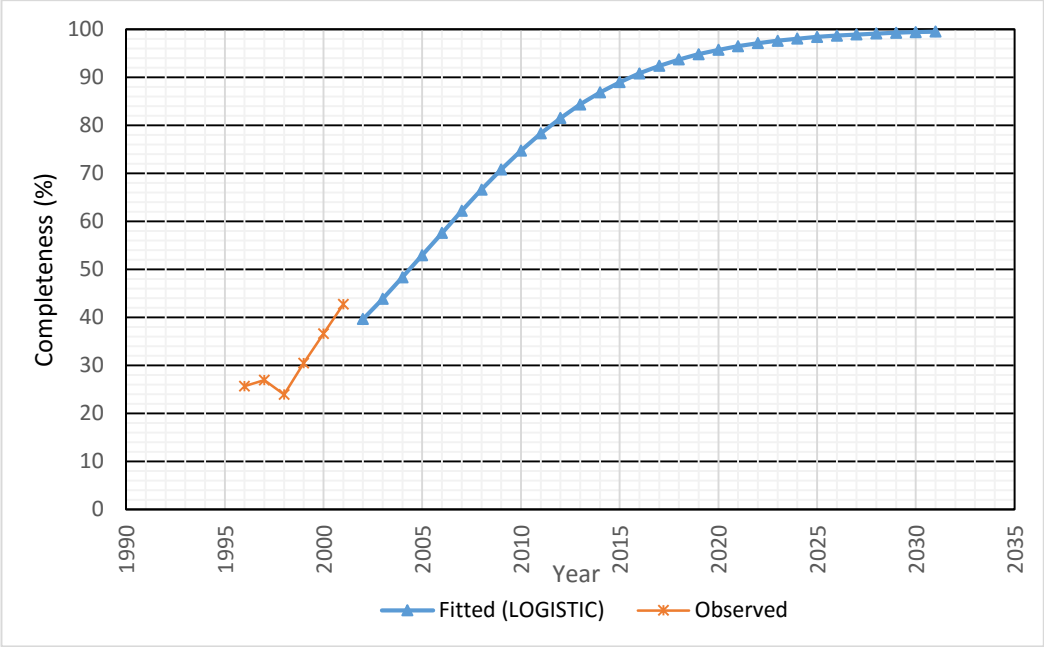


Table 4.2 Observed registered births by "Year 0" and "Year 1", and adjusted births estimated using completeness estimates by "Year 0" and "Year 1": 2001-2011

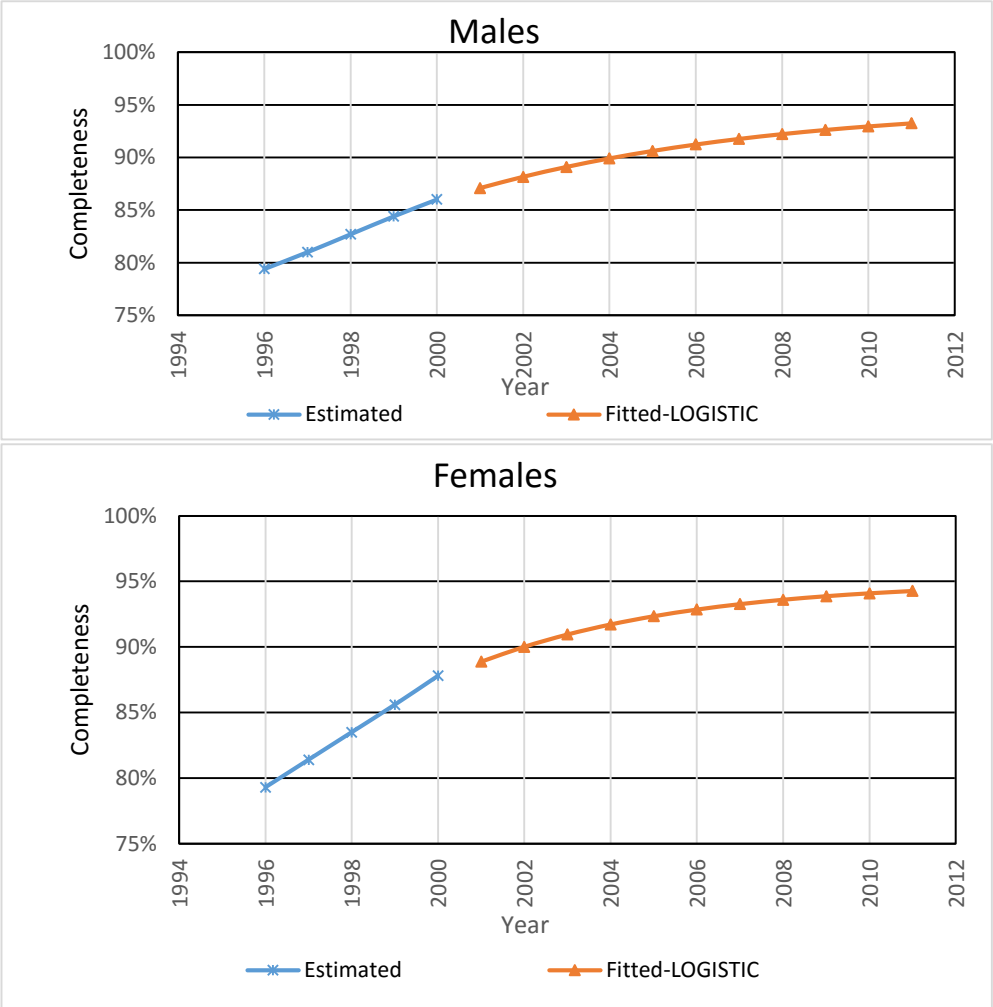
Year	Registered births by:		Estimated true births obtained using completeness by:	
	Year 0	Year 1	Year 0	Year 1
2001	477 489	695 745		1 116 693
2002	557 573	751 777	1 269 515	1 121 105
2003	621 887	787 549	1 201 042	1 070 984
2004	728 283	878 829	1 209 993	1 108 235
2005	793 788	948 119	1 164 436	1 126 299
2006	860 263	986 621	1 155 859	1 119 382
2007	858 866	987 202	1 073 073	1 081 941
2008	915 674	101 7417	1 090 498	1 086 853
2009	879 691	970 771	1 005 710	1 017 790
2010	889 691	969 770	988 279	1 003 056
2011	911 353	948 946	990 492	

The series of true numbers of births estimated using completeness estimates obtained using "Year 1" births ends with 2010 because at the time mid-year estimates are being produced for 2011, only "Year 0" births would have been available for 2011.

4.4 Completeness of adult death registration

The fitted trends in completeness for the projection period are similar for both sexes, with annual estimates of completeness being slightly higher for females. Fitted completeness by 2011 is 93.3 per cent for males and 94.3 per cent for females. Estimates of completeness for the years 1996-2000 and fitted estimates of completeness for the projection period 2001-2011 are shown in Figure 4-6 for males and females.

Figure 4-6 Estimated and extrapolated completeness of deaths registered in the year of death, 1996-2011



The estimates of completeness indicate that completeness of adult death registration has been increasing over time, which seems plausible given the ongoing concerted efforts by the Department of Health and the Department of Home Affairs to improve vital registration. The estimates of completeness can be improved year on year as new information allowing for the estimation of completeness of death registration becomes available.

4.5 Completeness of infant and child death registration

Figure 4-7 shows the estimated and fitted completeness of death registration of infants and 1-4 year olds using data available prior to and after the 2007 Community Survey.

Figure 4-7 Estimated and fitted completeness of infant and child death registration, prior to and after availability of data from the 2007 Community Survey

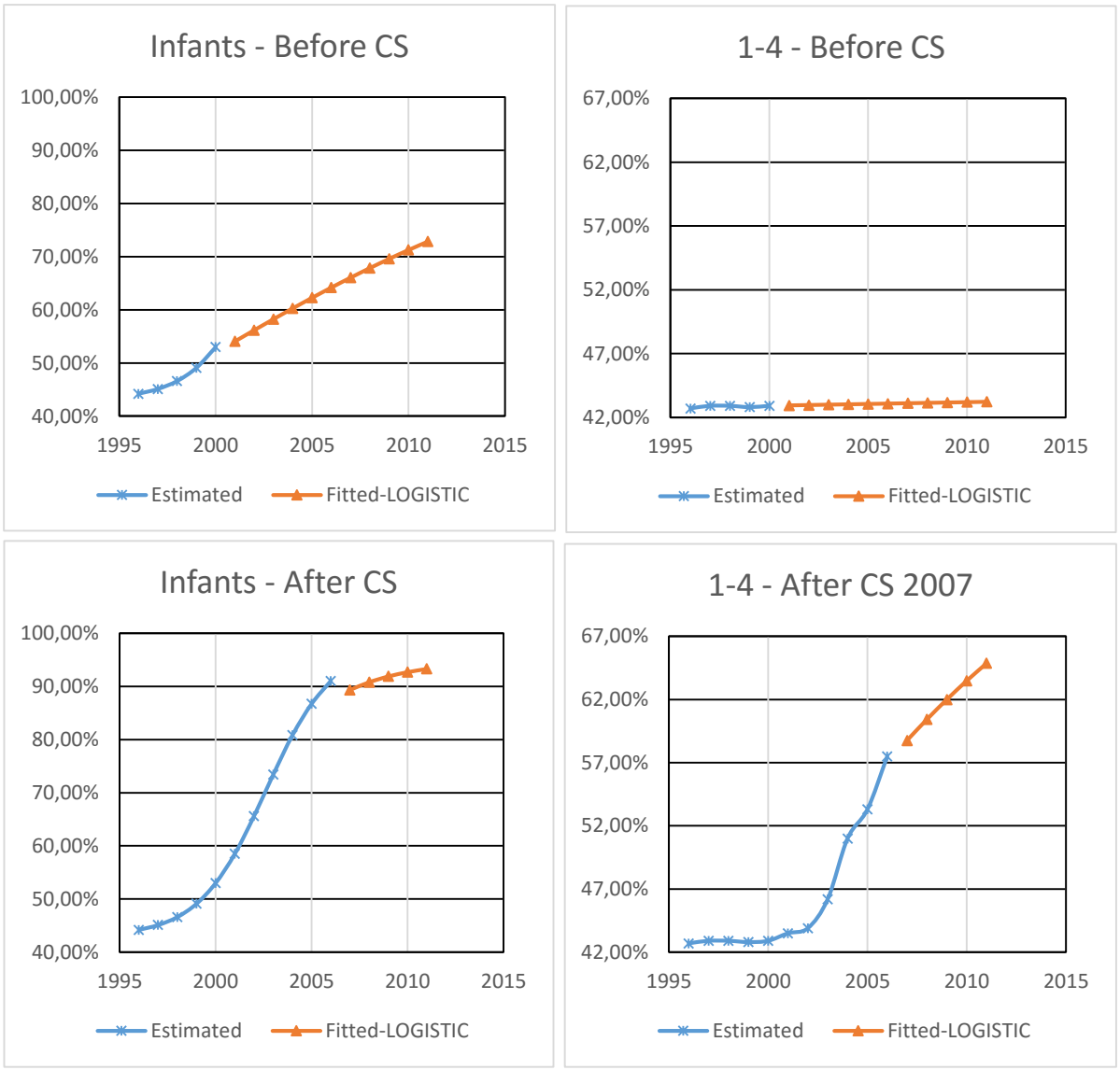


Figure 4-7 shows that completeness was increasing during the projection period, with completeness of infant death registration increasing at a faster rate. From the curve fitted using data available prior to the 2007 CS, it is estimated that completeness of infant death registration increased from about 60 per cent in 2001 to about 91 per cent in 2011. Completeness of deaths of 1-4 year olds on the other hand is estimated to have been flat at around 43 per cent. Using the data from the 2007 CS, completeness of infant death registration is estimated to have increased from around 60 per cent in 2001

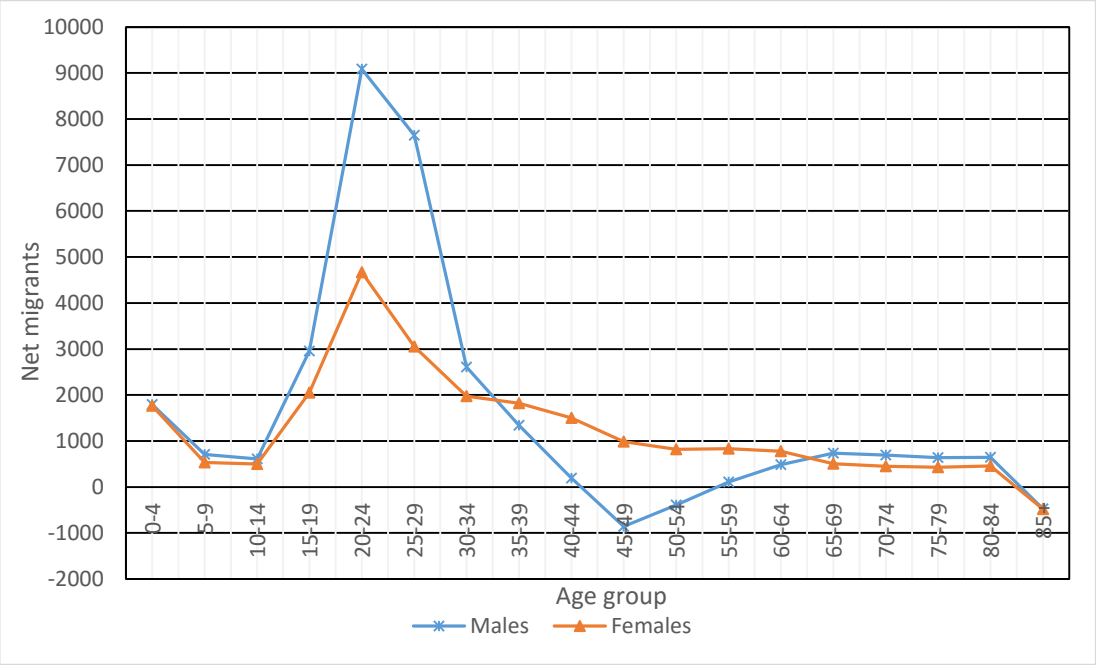
to around 97 per cent in 2011. For the 1-4 year olds, completeness is estimated to have increased from around 42 per cent in 2001 to around 65 per cent in 2011.

These estimates of completeness can be improved with each additional year of vital registration data available, thereby correcting any implausible change in the infant mortality rate or the under-five mortality rate from one year to the next.

4.6 Estimates of numbers of international migrants

Figure 4-8 gives the age distribution of annual net numbers of migrants in South Africa estimated using the method described in Chapter 3.

Figure 4-8 Net annual migrants in South Africa, 1996-2001



The estimated age distribution indicates higher net flows of males, with 28 517 males and 22 636 females. The shape of the distribution resembles a typical age distribution of dominant immigration with some return migration at the older ages. The peaks at the 20-24 and 25-29 last birthday age groups indicate labour force migration. The labour force migration peaks are accompanied by a peak at the youngest age group only if families are migrating with their children. However, there is only a minor peak at the youngest age groups, indicating that this is not the case in South Africa. There is an indication of net out-migration for males in the 45-49 age group which may be workers returning home. There is also an indication of net out-migration for the open age group for both males and females, which is unlikely. Minor contributors to this are probably underestimated net international in-migration for the period, and to a lesser extent,

overestimated net international out-migration, which is a consequence of the assumptions adopted when estimating net numbers of migrants. Differential age exaggeration could also be another factor contributing to this.

Further refinement of estimates of net flows of international migrants would be necessary to improve the estimates of net international migration. However, the main focus of this research is to determine the feasibility of using administrative data in producing current population and estimates, thus the estimated age distribution of net international migrants will suffice, since these numbers are small relative to the size of the population.

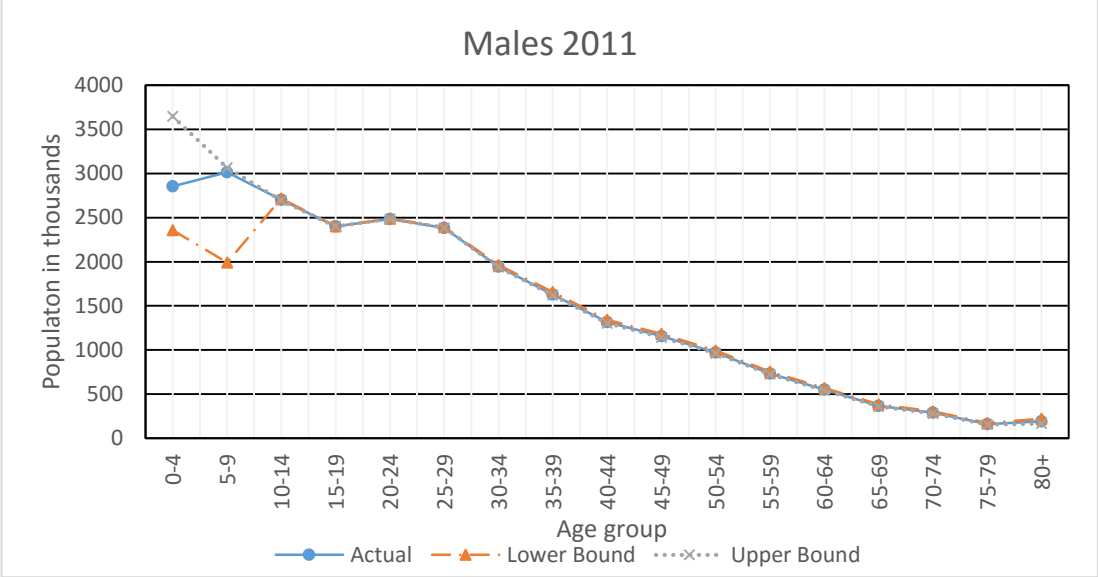
4.7 Mid-year estimates using administrative data

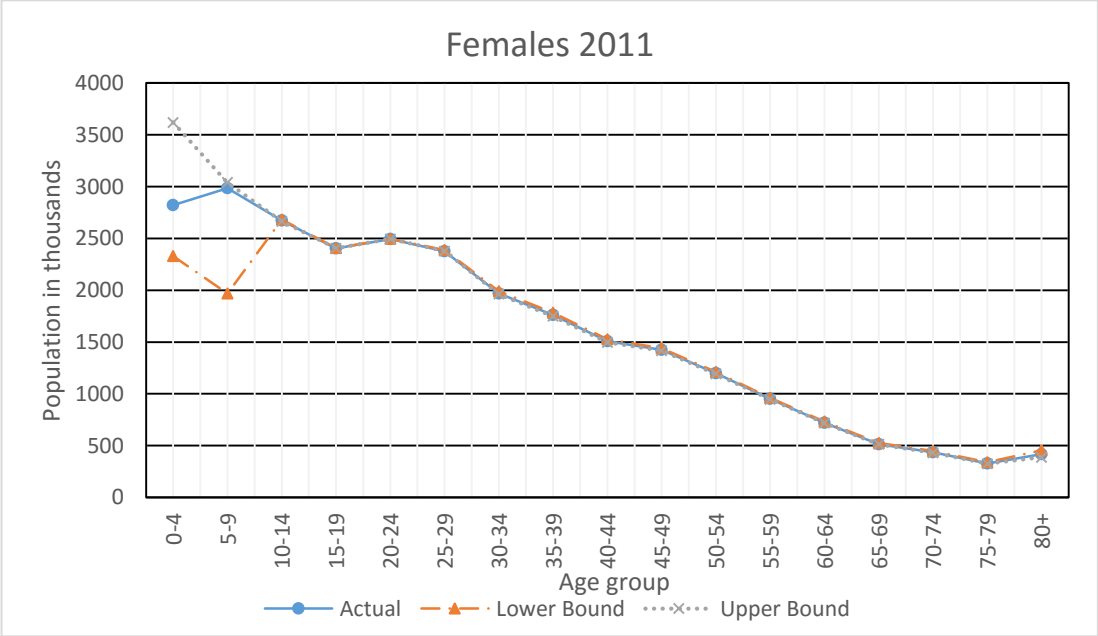
Mid-year population estimates were produced on a year-by-year basis for the years 2002-2011 using the cohort component projection method. The complete series of mid-year estimates produced using administrative data is available in the Appendix.

To assess the extent of uncertainty of the MYEs produced using administrative data, two extreme series of MYEs were produced by adopting different assumptions. The first series is produced assuming that the registered births and deaths are complete, thus no adjustment is applied to the vital registration data. This essentially underestimates the true numbers of births and deaths, thus underestimating the MYEs. This gives a sense of the lower bound of the MYEs produced using administrative data.

The second series is produced assuming that completeness estimated in 2001 remains the same over the remainder of the projection period, thus estimates of completeness are not revised going forward. The estimated level of completeness which is used to divide the registered births and deaths is lower than actual, thus it essentially overestimates the true numbers of births and deaths since completeness obviously increases over the period. As a result, this overestimates the MYEs produced, thus giving a sense of the upper bound of the MYEs produced using administrative data. Figure 4-9 shows a comparison of the three series of the ultimate MYEs for 2011, the lower bound MYEs, the upper bound MYEs and the actual MYEs.

Figure 4-9 Comparison of Lower Bound, Upper Bound and Actual MYEs produced using administrative data, 2011





The uncertainty is, obviously, confined to the 0-9 last birthday age-range. For the 0-4 last birthday age-group, the upper bound estimates are approximately 28 per cent higher than the actual estimates, whilst for the lower bound estimates are about 17 per cent lower. For the 5-9 last birthday age-group, the upper bound estimates are about two per cent higher than the actual estimates, whilst the lower bound estimates are about 34 per cent lower. The upper bound estimates are closer to the actual estimates probably because the assumptions adopted when producing them are more realistic than the assumptions adopted for the lower bounds.

Overall, the upper bound estimates are about 2.8 per cent higher than the actual estimates, whilst the lower bound estimates are about 4.9 per cent lower. The key point from this exercise is that the lower bound estimates particularly exaggerate uncertainty because the assumption is unrealistic. The upper bound is also an exaggeration, but a more realistic indication of the level of uncertainty. This exercise merely indicates that there is uncertainty surrounding the estimated MYEs, though is not very large.

4.8 Comparison with mid-year estimates produced by Stats SA

In order to evaluate the approach used, it is necessary to compare the mid-year estimates to those produced by Stats SA, which uses SPECTRUM. The comparison is based on expected demographic characteristics of year on year population distributions. To determine accuracy, a comparison with the 2011 census population distribution is made. Analysing the differences between the population distributions and determining what could be causing the differences will help improve the estimates.

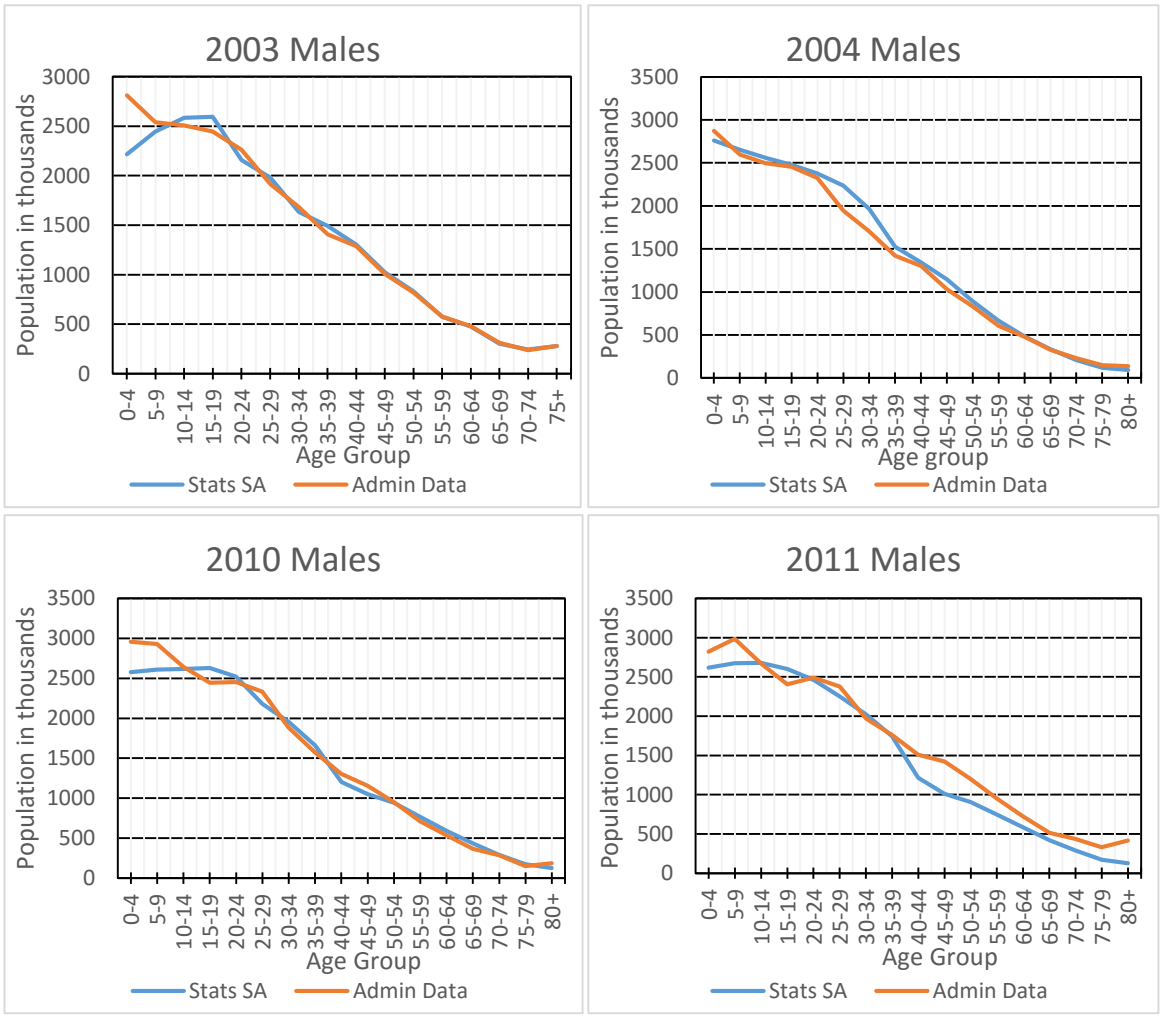
4.8.1 Comparison of age distributions

Figure 4-10 shows comparisons of the age distributions, by five year age-groups for males, of the mid-year estimates produced by the two methods for the years 2003, 2004, 2010 and 2011. These years are chosen for the comparison to show the differences at the earlier years of the projection, and at the later years which highlight the accumulation of the differences. The comparisons for females are similar and they are available in the Appendix. The age distributions of the mid-year estimates produced by Stats SA were obtained from the statistical releases published by Stats SA (Statistics South Africa 2003, 2004, 2010, 2011).

For 2003, there is a notable difference in the 0-4 last birthday age group. The projected numbers in the 0-4 last birthday produced using administrative data are approximately 20 per cent more than the 0-4 last birthday produced by Stats SA for both males and females. This is probably an indication of the difference in the population bases used by the two methods. For their base population for the 2003 mid-year estimates, Stats SA state that they used the 2001 census population only adjusted for the undercount by the Post Enumeration Survey (PES), and they did not make any further adjustments to the age specific numbers (Statistics South Africa 2003). The base population used in the method which uses administrative data has further adjustments

on the PES adjusted population distribution from the 2001 census, in particular, for the undercount of the 0-4 last birthday in the census, as described in section 4.2.1.

Figure 4-10 Comparison of the age distributions produced using vital registration (Administrative) data and those produced by Stats SA; 2003, 2004, 2010, 2011



In the 2004 age distributions, the difference in the 0-4 last birthday is much less than in 2003, with the numbers aged 0-4 last birthday produced using administrative data being only about five per cent higher than the numbers aged 0-4 last birthday from Stats SA. This apparent shift is probably because Stats SA changed their methodology for projecting the population in mid-2004. In 2003, Stats SA extrapolated the inter-censal age-specific growth rates from the 1996 and 2001 censuses to estimate the mid-year population in 2003, essentially assuming that the age-specific growth rates remained constant during the post-censal period (Statistics South Africa 2003). They opted to use mathematical extrapolation due to what they regarded as a lack of reliable fertility,

mortality and migration data from the 2001 census (Statistics South Africa 2003). However, they did mention that the mid-year estimates produced for 2003 were provisional and revised estimates would be produced after more detailed analysis of the 2001 population was completed.

In 2004, and the following years, Stats SA started using DemProj (part of SPECTRUM suite of policy models developed by the Futures Institute) to produce mid-year estimates (Statistics South Africa 2004). The DemProj software uses the cohort component projection method to project populations for countries or regions (Stover and Kirmeyer 2008). The software requires numbers of people by age and sex in a base population, assumptions about the total fertility rate (TFR), the age distribution of fertility, the life expectancy at birth by sex, the most appropriate model life table and the magnitude and pattern of migration for all the years of the projection (Stover and Kirmeyer 2008). The age distribution of the base population is normally obtained from national censuses (possibly after correction of any shortcomings) while assumptions about fertility, mortality and migration are adopted from demographic research based on data from censuses and national surveys (Stover and Kirmeyer 2008).

The base population used to project the 2004 mid-year population was the 1970 census age distribution. Stats SA state that they "constructed" a set of TFRs for their projection in such a way that the projection would reach the 2001 adjusted census population (Statistics South Africa 2004, 7). However, they do not give details of how they constructed their TFRs. They compared their TFRs with TFRs estimated by various researchers (Van Aardt and Van Tonder 1999, Moultrie and Timæus 2003, Moultrie and Dorrington 2004, Phillips, Phoshoko, and Cronje 2004, Udjo 1999, 2004b, Sadie 1993).

For the mortality assumptions, Stats SA state that they used life tables "nearly identical" to the life tables Statistics South Africa (2000) calculated for the four population groups (Africans, Asians, Coloureds, Whites) in South Africa for the years 1985-1994 and 1996 (Statistics South Africa 2004, 10). Stats SA also say for the years after 1995, the life expectancies for the projection are largely influenced by the HIV prevalence assumptions made for the years 1990 and later by SPECTRUM (Statistics South Africa 2004). The life expectancies at birth they used in their projection closely match those estimated by Udjo (2004b, 2003, 2004a).

For the migration assumptions in DemProj, Stats SA used migration estimates published in migration reports they compiled (which they do not cite) and estimates by

Van Aardt and Van Tonder (1999) for White South Africans (Statistics South Africa 2004). They ignored migration for the other population groups due to the lack of data (Statistics South Africa 2004).

The main point is that the methods used by Stats SA to produce mid-year population estimates for 2003 and 2004 each used entirely different assumptions and methods, which results in the differences from the mid-year estimates produced using administrative data. The differences with the age distributions produced by Stats SA gradually get more pronounced as the years in the projection period increase, with Stats SA's age distribution fluctuating below, above or at the same level as the age distribution produced using administrative data by 2011, as shown on Figure 4-11 The shifts in the estimated age distributions produced by Stats SA from one year to the next result in inconsistency in their series of population estimates. Figure 4-11 shows comparisons of annual age distributions of the mid-year estimates produced by Stats SA and those produced using administrative data for males.

Figure 4-11 Comparisons of the consistency of the age distributions produced by Stats SA and those produced using administrative data, Males 2003-2008

From Figure 4-11, we see that the age distributions of the mid-year estimates produced by Stats SA for the years 2003-2005 differ significantly. From 2004 onwards, the shapes of Stats SA's age distributions show closer resemblance from one year to the next. However, there are still some notable shifts in the sizes of the age groups, such as in the age range 5-24 last birthday from 2007 to 2008. This is in contrast to the close similarities between the distributions for 2006 and 2007.

From 2004 onwards, the shifts in Stats SA's age distributions, which are not as marked as from 2003-2004, are probably due, mainly, to regular updates on the source code and bug fixes in the DemProj software. The Futures Institute releases many versions of the SPECTRUM suite of models during the year with updates and bug fixes. Additionally (about every two years), the Futures Institute updates the future assumptions of fertility, mortality and migration as new information becomes available from demographic research as presented in the World Population Prospects[8] (Stover and Kirmeyer 2008). These factors contribute to the inconsistencies observed in Stats SA's age distributions.

For the age distributions produced using administrative data, there are, as might be expected in the absence of any demographic shocks, hardly any noticeable differences in the shapes of the distributions as shown in Figure 4-11. There are only slight increments and decrements in some of the age groups indicative of population growth from one year to the next. There is, however, a slight difference in the oldest age group between 2003 and 2004. The open age group for the 2003 mid-year estimates is 75+ and for 2004 it is 80+, hence the noticeable difference between the two distributions at the last age group. The open age group for 2003 was set to be 75+ in order to facilitate comparison with Stats SA's age distribution for 2003 which is 75+. Otherwise the age distributions produced using administrative data are quite consistent from one year to the next.

Considering the above analysis, in terms of consistency of age distributions of mid-year estimates from one year to the next, at least, the mid-year estimates produced using administrative data perform better than those produced by Stats SA using DemProj. Stats SA do mention, in the description of the methodology for producing the 2013 mid-year estimates, that series of mid-year estimates produced in each of their statistical releases are unique due to changing assumptions and the availability of new

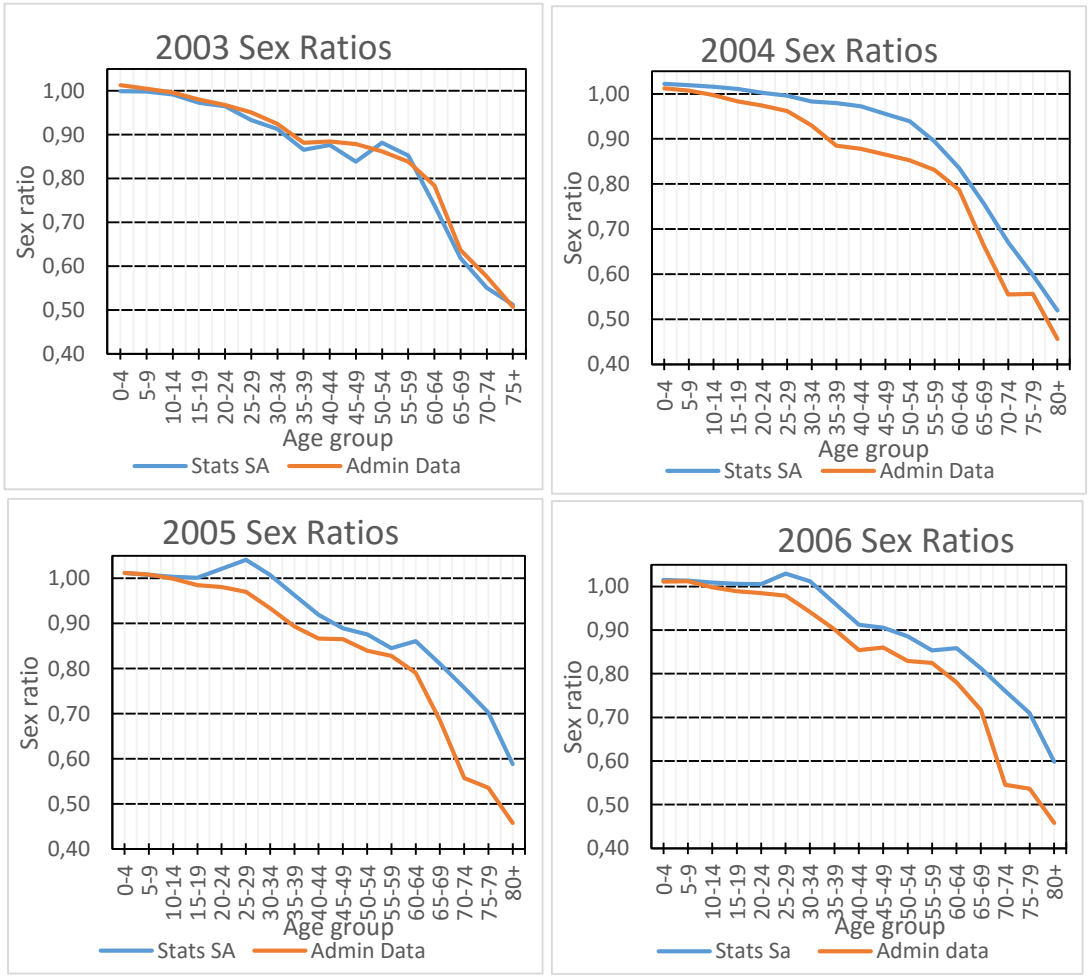
[8] http://esa.un.org/unpd/wpp/

data inputs from one year to the next (Statistics South Africa 2014). They go on to say that users of their mid-year estimates should not compare series of mid-year estimates from different statistical releases, essentially acknowledging the presence of inconsistencies in their age distributions from one year to the next. This does not inspire confidence in using their population estimates or in their method of producing them. Consistency of current population estimates in the absence of demographic shocks, and confidence in them and the method used to produce them were identified in the literature as important characteristics in determining the quality of the population estimates. In this regard, the method used by Stats SA compromises the quality of their population estimates.

4.8.2 Sex ratios

Comparison of sex ratios gives further insights about the quality of the two series of mid-year estimates and the differences between them. Figure 4-12 compares the sex ratios produced by the two methods for the years 2003 to 2006.

Figure 4-12 Comparison of sex ratios produced using administrative data and those produced by Stats SA, 2003-2006



The curves of sex ratios are quite similar in 2003, although from age 35 last birthday onwards, Stats SA's distribution mostly falls slightly below the distribution produced using administrative data. From 2004 onwards, the sex ratios become radically different, with the sex ratios produced by Stats SA being consistently and significantly higher than those produced using administrative data.

The increase in difference is mainly due to the change in the methodology of producing mid-year estimates by Stats SA from 2004 onwards. Changing the base population from the 2001 census population to the 1970 census population certainly has a big impact in the change in Stats SA's sex ratios. The starting age structure of a population projection has a big impact on population growth (and hence the projected age structure) during the first few years of a projection as the variability in vital rates will be relatively small, in which case the cohort component method performs well (Bongaarts and Bulatao 1999, Keyfitz 1971, Stoto 1983). However, with longer projection periods, the precision gained from using the starting age distribution begins to be overcome by the uncertainty in the future vital rates and migration (Goldstein and Stecklov 2002). When Stats SA used the 1970 census population as the base population, the projection period for the 2004 mid-year estimates was quite long (34 years) compared to the projection period when they used the 2001 census population as the base (3 years). Precision in projecting the age structure is lost with the longer projection period, hence the radical difference in their sex ratios from 2003 to 2004.

Although Stats SA do mention that they "constructed" (Statistics South Africa 2004, 7) their fertility and mortality rates for the DemProj projection such that their projection would reach the 2001 census adjusted population, they do not say whether they matched the 2001 census population age distribution. If they did not, it would probably compromise the consistency of their sex ratios just as with the age distribution.

Figure 4-13 gives comparisons of Stats SA's series of sex ratios in two three-year groups as well as for the sex ratios produced using administrative data for the years 2003-2008 to show the gradual increase in the differences in Stats SA's sex ratios. Except for the years 2006 and 2007, the curves of Stats SA's sex ratios are quite different, particularly for the years 2003-2005. There is no resemblance even in the sex ratios for 2004 and 2005, immediately after Stats SA started using DemProj to produce their mid-year estimates. It would seem, however, that Stats SA adopted assumptions of fertility, mortality and migration for producing their 2005 mid-year estimates that were completely different from those that they had used to produce their 2004 mid-year

estimates (Statistics South Africa 2004, 2005). As stated earlier, Stats SA do not explain how they "constructed" their fertility and mortality assumptions for the 2004 projection, nor do they cite their sources for their migration estimates. They do not give any detail

Figure 4-13 Comparisons of the consistency of the series of sex ratios produced by Stats SA and those produced using administrative data, 2003-2008

either in 2005 of how they came up with their fertility, mortality and migration assumptions for their projection.

Without details about how they adopt fertility, mortality and migration assumptions for their projections, it is not possible to understand fully why their sex ratios are markedly different from one year to the next. Nonetheless, it can be said that the inconsistencies in their sex ratios are probably due to the changes to their assumptions. Software updates and bug fixes to the SPECTRUM suite of models probably contribute, but to a much smaller extent than was the case with age distribution, to the inconsistencies in sex ratios as well.

The sex ratios produced using administrative data are quite consistent from one year to the next, with minor differences in some age groups. As with the age distributions discussed earlier, sex ratios normally would not change drastically from one year to the next, unless due to extraordinary circumstances and/or demographic shocks. Therefore, regarding sex ratios, the method which uses administrative data performs better than the method used by Stats SA.

4.8.3 Overall conclusion on expected demographic properties of MYEs

The MYEs produced using administrative data clearly performed better in terms of expected demographic characteristics of population estimates. Since migration accounts for only a small part of the difference, this suggests that estimates of births and deaths derived from the administrative data are probably more accurate (or at least produce more demographically consistent projections) than those produced by the model used by Stats SA and used in their projection using SPECTRUM.

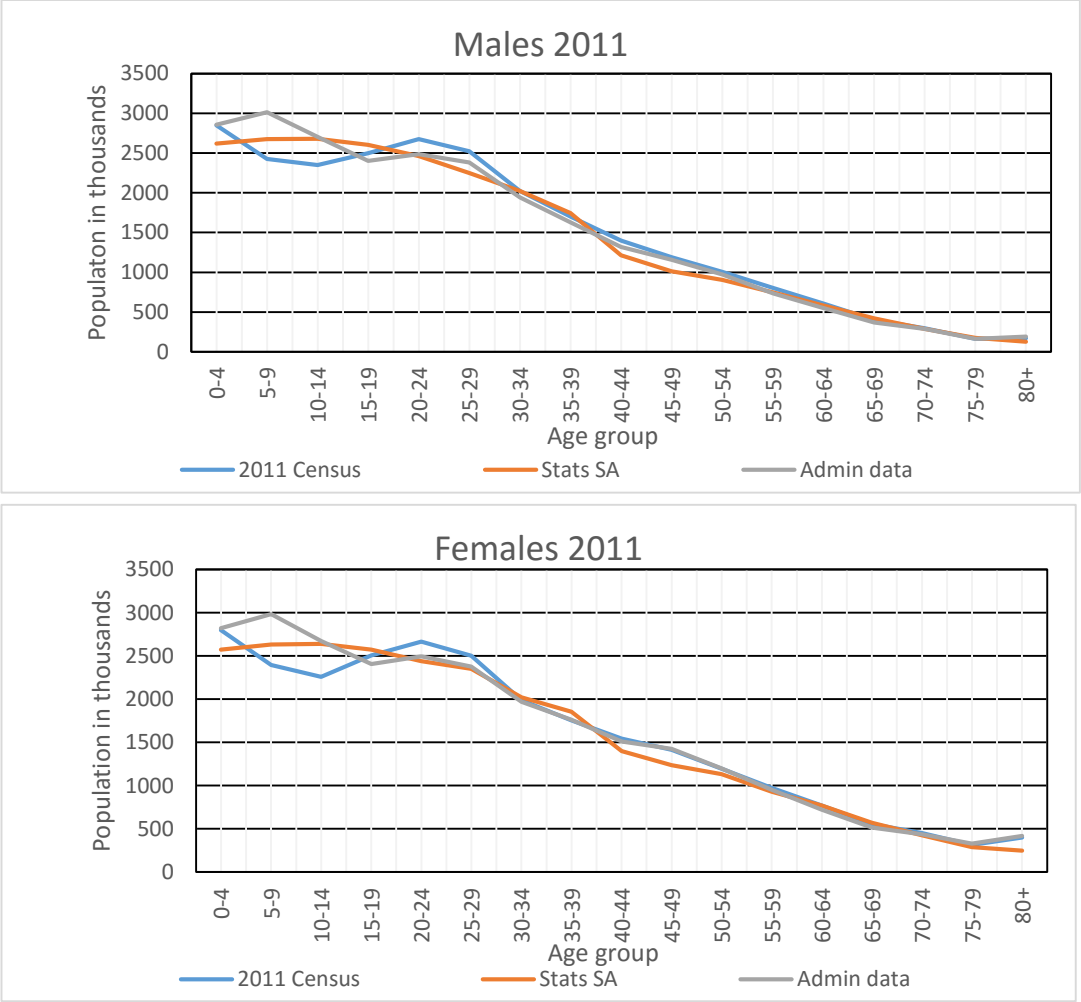
4.9 Comparison with the 2011 census age distribution

Although the mid-year estimates produced using administrative data perform better, in terms of exhibiting expected demographic characteristics of annual population estimates, than those produced by Stats SA using Spectrum, it does not necessarily mean that they are more accurate (closely matched with the census age distribution). In order to determine the level of accuracy, it is necessary to compare the age distributions produced by both methods to the census enumerated age distribution which "best" estimates the true population age distribution, albeit there are some errors in the census counts such as under-count or over-count of some age groups.

4.9.1 Age distributions

Figure 4-14 shows the comparison, in five-year age groups, of the population in mid-2011 estimated from the 2011 census to the mid-year population produced by the two methods, for males and females separately.

Figure 4-14 Comparison of age distributions produced by Stats SA and by using administrative data, to the age distribution from the 2011 census, Males and Females



The total population estimated using administrative data (52 133 453) is about 600 000 more than the census population (51 552 790), while the total population estimated by Stats SA (50 586 757) is about 1 000 000 less than the census population. Although the population estimated using administrative data is closer to the total census population in this regard, both the age distributions produced by Stats SA and the administrative data are markedly different from the census age distribution, particularly for ages below 40 last birthday.

The differences between the numbers produced using administrative data and the census are probably a result of three main factors; (1) under-estimates of completeness of birth registration for the first five years of the projection period resulting in over-inflation of registered births to estimate true numbers of births, (2) over-estimate of the numbers of the 0-4 last birthday age group in the base population resulting in the

projected 10-14 last birthday being considerably more than the 10-14 last birthday enumerated as at the census, and (3) under-estimation of numbers of net migrants during the 1996 to 2001 inter-censal period, most likely causing numbers in the projected distribution to be lower than the numbers estimated from the census distribution, particularly for ages 20-29 last birthday.

For the age group 0-4 last birthday, the census and the projection using administrative data produce similar numbers, while the numbers projected by Stats SA are lower. This indicates that the estimates of completeness of birth registration, used to estimate true numbers of births in the projection using administrative data, had become more stable and accurate, accepting that the census accurately enumerated the 0-4 last birthday age group. It also indicates that the fertility assumed for the years 2006-2011 by Stats SA's projection are lower than true fertility for the period, accepting that the census accurately enumerated the 0-4 last birthday age group. However, for the age group 5-9 last birthday, Stats SA produced numbers much higher than the census numbers which indicates that the fertility assumptions in their projection for the years 2001-2005 are much higher than actual fertility.

The comparison also indicates that true numbers of births estimated from vital registration for the period were too high. This indicates that estimates of completeness of birth registration for the years 2001-2005 are too low, thereby over-inflating the registered births. Furthermore, this is probably an indication that the method used to estimate completeness of birth registration was not yet stable enough in estimating completeness (estimates produced using "Year 0" births) for the first five years of the projection period accurately. For the last five years of the projection period, the method probably became more stable, and actual completeness of birth registration probably became sufficiently high and stable for the numbers produced by the method to match those from the census. However, there is the possibility that the census under-counted the 5-9 age group to some extent, given how big the differences are, especially between the census numbers and the numbers produced using administrative data.

The numbers for the 10-14 age group from both the projected distributions are notably higher than those from the census. This indicates that the adjustment to the 0-4 last birthday age group in the 2001 base population for the projection using administrative data was too high and that Stats SA's numbers of the 0-4 last birthday age group in their base population were also too high. While this also indicates that the fertility rates used to project the 1996 census population during the demographic

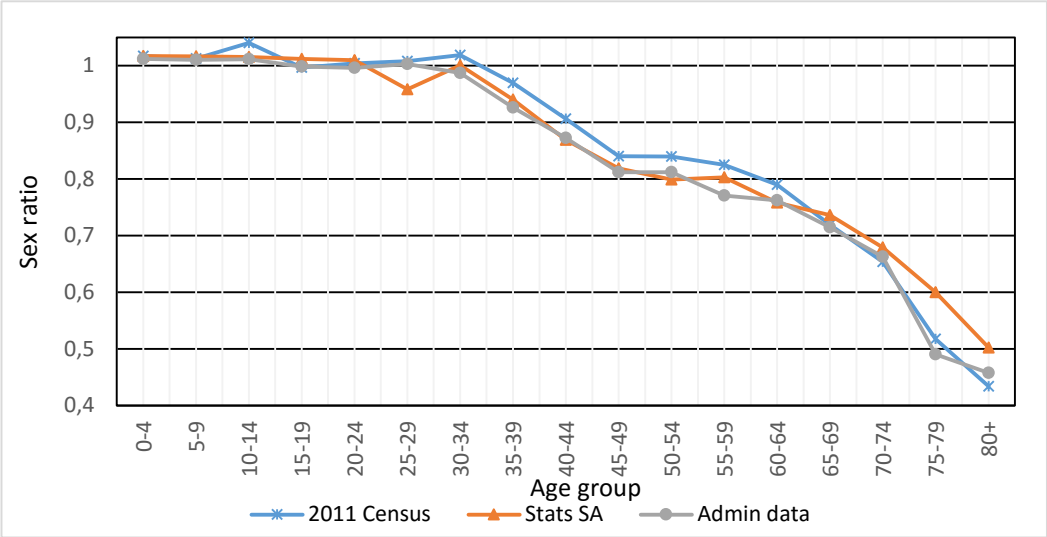
analysis of the 2001 census were too high, it does not nullify the assertion that the 0-4 last birthday were under-counted as at the 2001 census.

For the age ranges 15-44 last birthday for males and 15-29 last birthday for females, the age distribution produced using administrative data is consistently lower than the census distribution, though it maintains a shape similar to that of the census distribution. This indicates that migration used in the projection is probably lower than it actually was, given that the affected ages are mostly young adults who generally make up the majority of those who migrate (Skeldon 2013, Ravenstein 1985). The differences are more pronounced for the males and they extend over an age range of 30 years compared to the 15 years for females, indicating further that male migration was under-estimated to a greater extent. In the projection by Stats SA, their age distribution falls below the census distribution for the age ranges 20-29 last birthday and 40-54 last birthday, meaning their assumed migration is also lower than the actual migration for the period for these ages. This could also mean that the mortality assumptions in their projection were too high. However, Stats SA better match the census numbers for the age ranges 15-19 last birthday and 30-39 last birthday, with the numbers for the 30-39 for males being more or less the same as those from the census. This being the case, Stats SA's projection better approximates migration during the projection period only for the age range 30-39 last birthday in mid-2011. Their estimates of migration are better probably because they used data from the 2007 Community Survey to update their estimates of migration (which wasn't the case for the projection using administrative data). However, the shape of their age distribution is not similar to the census distribution, and the root cause of this is probably the inconsistency of the age distributions of their mid-year estimates from year to year discussed earlier.

4.9.2 Sex ratios

The sex ratios produced by the two methods were compared to the census sex ratios to check which method is more accurate in terms of projecting the relative distribution of males and females over the period. Figure 4-15 shows the comparison of the sex ratios produced by the two methods and those from the census.

Figure 4-15 Comparison of the sex ratios, 2011



The curves of the sex ratios produced by the two methods mostly do not fall particularly close to the curve of sex ratios from the census, although there is quite a broad similarity with the census distribution in terms of shape. For the younger age groups, the three distributions of sex ratios are mostly similar, except the upward fluctuation in the census sex ratios for the age group 10-14 last birthday. This indicates that there was an under-count of female children in the age group, or an over-count of male children. From age 15 onwards, the sex ratios produced using administrative data consistently fall below those of the census, notably for ages 20-44 last birthday. The under-estimation of migration for the projection period by the method using administrative data is probably a root cause of this, with the comparison of the projected and census age distributions indicating that the impact of male migration was greater than the impact of female migration. This being the case, the projection using administrative data under-estimated males in the age group, hence the projected sex ratios fall consistently below the census sex ratios. Despite this, the shape of the curve of sex ratios is similar to that of the census for most of the age groups, even those affected by under-estimation of migration. For the older age groups, the shape of the curve of the sex ratios produced using administrative data bares a close resemblance to the census curve of sex ratios, except for fluctuations at age groups 55-59 and 75-79 last birthday.

Stats SA's distribution of sex ratios is closer to the census age distribution for the age groups 20-24 and 30-39 last birthday. This is probably due to the fact that Stats SA had better estimates of migration as discussed earlier, hence there is a smaller shortfall

of males in the age groups affected by under-estimation of migration. However, Stats SA's distribution of sex ratios is mostly not similar in shape to the census distribution, with some pronounced fluctuations between ages 20-35 and notable divergence from the census sex ratios from age 65 onwards. Stats SA's sex ratios do not progress smoothly with age and the inconsistency of their sex ratios from year to year discussed earlier is probably a root cause of this.

5. DISCUSSION AND CONCLUSIONS

This research set out to determine whether administrative data could be used to produce reasonably accurate mid-year population estimates for South Africa as an alternative to the method being used by Stats SA. Data from vital registration, adjusted for incompleteness, were used to estimate annual births and deaths which were used in a cohort component approach to project the population from the 2001 base population adapted from the 2001 census population. Migration used in the projection was estimated using place of birth data from the 1996 and 2001 censuses and assuming that the age distribution of net migration from the period 1996-2001 remained constant during the period 2001-2011. This chapter reflects on the extent to which the research objectives were met by applying the proposed method of producing mid-year estimates.

Year-on-year consistency of age distributions and sex ratios

Analysis of the results revealed that the age distributions and sex ratios of the mid-year population estimates produced using administrative data are more internally consistent from one year to the next than those produced by Stats SA re-parameterising a different version of the Spectrum model each year. Thus, in terms of expected demographic characteristics of population age distributions and sex ratios from one year to the next, the series of mid-year estimates produced using administrative data is better than Stats SA's mid-year estimates. This finding suggests encouraging prospects of having an alternative to the method used by Stats SA, which relies on a complex model requiring many, often vaguely identified, assumptions.

It should, however, be stated that the MYEs produced using administrative data conformed well to expected demographic characteristics of populations. This is probably a consequence of the fact that migration accounts for a small part of the change in population from one year to the next, hence its effect was not very significant. Furthermore, the low net migration is a clear consequence of the assumptions adopted when estimating numbers of net international migrants. The assumed annual migration was clearly lower than the actual migration for the projection period. The Stats SA's estimate of migration was better than assumed for the administrative projection (as shown by Stats SA's MYEs being more accurate for the ages mostly affected by migration (15-29 last birthday)). This being the case, migration for the projection using administrative data should have been estimated in the same way Stats SA did which would have improved the MYEs.

As far as the year-on-year inconsistencies in the Stats SA series of mid-year estimates are concerned, the main source of inconsistency in the age distributions for the years 2002 and 2003 is the change in methodology, from extrapolating growth rates to using the cohort component method. Changing the base population probably contributes to the year-on-year inconsistencies as well. For the years that follow, 2004-2011, the year-on-year inconsistencies in their age distributions are less pronounced. Although changing their methodology would have been expected to coincide with some changes in the shape of their age distributions, the changes should not be as pronounced as they are. Changing a method of projecting a population should not result in a change in the population's age structure, unless due to demographic shocks. This should obviously have been considered by Stats SA in adopting their new model and assumptions. At least they should have investigated why there were such pronounced year-on-year inconsistencies in the age distributions and taken measures to reduce their extent.

Software updates to the source code, bug fixes for the SPECTRUM suite of models, and occasional changes to the assumptions (fertility, mortality and migration) as new empirical evidence becomes available, are additional sources of the year-on-year inconsistencies in the age distributions and sex ratios. The errors may be minimal from one year to the next, but obviously they could accumulate over the whole or most of inter-censal period. The accumulation of the errors could then manifest in the form inaccuracies in the age distribution when the mid-year estimates are compared to a census population.

5.2 Comparison with the 2011 Census age distribution

When compared with the mid-year population estimated using the 2011 census, both the mid-year population estimates produced using administrative data and by Stats SA did not accurately match the census age-specific numbers for certain age-groups. For the mid-year estimates produced using administrative data, the worst affected ages were those below 45 last birthday for males and 35 last birthday for females. These inaccuracies suggest that the main sources of inaccuracy in the projection using administrative data were (1) over-estimating the true numbers of births in the first five years of the projection (2) too high an estimate of the number aged 0-4 last birthday in the base population (aged 10-14 in 2011), and (3) under-estimation of migration for the age groups 15-44 last birthday for males and 15-34 last birthday for females at the time

of the census. Although some of the differences could be due to errors in the 2011 Census count, these are expected to be small.

Stats SA's mid-year estimates differed from the census numbers by age for ages below 55 last birthday for both males and females. However, their age distribution was slightly more accurate (closer to the census) for ages 30-39 last birthday for males. Since Stats SA do not provide full details of their adaptation of DemProj when publishing their mid-year estimates, it is not clear what factors contribute to the age-specific inaccuracies in their population estimates.

5.2.1 0-9 last birthday

Inaccuracies in the 0-9 last birthday age group are mainly caused by estimates of incompleteness in birth registration and the volatility in quantifying the extent of inaccuracy from one year to the next. Completeness of birth registration for births registered during the year of birth was estimated to be about 78 per cent by 2011. Completeness of vital registration seems to have stabilised for the last five years of the projection period, resulting in the numbers aged 0-4 last birthday produced using administrative data matching with the corresponding census numbers. This further presents encouraging prospects for using vital registration to produce mid-year estimates in future. With constant, or at least near constant, reporting of births and deaths, the method can be expected to produce more accurate results.

The projected number aged 0-4 last birthday was 0.5 per cent higher than 0-4 last birthday estimated from the census, while the projected number aged 5-9 was 24.4 per cent higher. This indicates that completeness of birth registration was substantially underestimated for the first five years of the projection using administrative data, resulting in over-adjustment of registered births to estimate true numbers of births. The root cause can be attributed to the starting point of estimating completeness of birth registration, which was derived using empirical estimates of fertility for the years 1996 and 2001, assuming linear change in age-specific fertility rates for the intervening years, and projected numbers of women in the 15-49 last birthday age group.

Fertility is not expected to have changed markedly during the years 1996-2001. However, it seems that the implied true numbers of births, estimated using projected numbers of births, were too high, resulting in estimates of completeness of birth registration for the years 1996-2001 that were too low. This subsequently resulted in the extrapolation of completeness also yielding underestimated levels of completeness, particularly for the first five years of the projection. Hence the starting point of

estimating completeness of birth registration could have contributed to the projected number aged 5-9 last birthday being much greater than the number from the census. Errors in birth registration data, which provided the numerators in estimating completeness, could also have contributed to the inaccuracies in the projected number aged 5-9 last birthday.

Accurate estimates of completeness of birth registration are, to a large extent, a prerequisite when using administrative data to produce mid-year population estimates. The biggest deviation from numbers estimated from the census is the numbers of those aged 5-9 last birthday as at the middle of 2011. Thus, it would be important to develop other methods of estimating completeness of birth registration, such as comparing registered births to births reported in a census of survey.

The assumptions adopted when estimating completeness of birth registration were thus insufficient. Clearly completeness of birth registration was under-estimated, hence over inflating registered births. The method used probably had a hand in this as well. It was an innovative method suggested for the purpose owing to the fact that there is no formal indirect method for estimating completeness of birth registration. Furthermore, there is no reliable literature on the completeness of birth registration. The method would definitely need to be improved and the assumptions adopted for the method need to be revised. This is an area for further research.

5.2.2 10-14 last birthday

The difference between the projected number aged 10-14 last birthday and the number from the 2011 census indicates that the number aged 0-4 last birthday in the base population was too high. However, as already discussed, the 1996-2001 inter-censal fertility rates used to estimate the 0-4 last birthday age-group in the base population were consistent with other empirical estimates of fertility at the time, which now, in retrospect, appear to have been too high. Identifying why this is so requires further research of fertility data from the 1996 and 2001 censuses. Accuracy of the projected number aged 10-14 last birthday in mid-2011 could be improved if lower fertility rates were used to adjust the 2001 base population.

5.2.3 Underestimation of migration

Under-estimation of migration was identified as an apparent source of inaccuracy in both the age distribution produced using administrative data and by Stats SA when compared to the age distribution of the 2011 Census. Clearly, much needs to be done to

improve the estimates of annual net numbers of migrants. To this end, two methods are suggested, which, it is hoped, will detect the magnitude of increases in the net numbers of migrants from one year to the next in future. However, the methods could not be used in this research for the projection period under consideration, or tested for the post-projection period due to the lack of data. Stats SA only started to publish the necessary migration data in 2012 (Statistics South Africa 2012a), and this with a one-year lag.

5.3 The impact of international migration on producing mid-year population estimates

Comparison of the estimated 2011 mid-year population distribution produced using administrative data to the mid-year population distribution based on the 2011 census indicates that under-estimation of international migration during the projection period adversely affected the accuracy of the mid-year estimates produced using administrative data. This is highlighted by the fact that the population age distribution produced using administrative data is consistently lower than, though similar in shape to, that based on the 2011 census, most notably for the ages 15-44 last birthday. Underestimating international migration also results in an underestimate of births which subsequently results in an underestimate of children during the projection period.

There is a need to develop a method of improving estimates of international migration from one year to the next when estimating post-censal mid-year population estimates, with the initial estimate of migration derived from census data. Skeldon (2013) stresses the importance of improving the quality and quantity of data for estimating migration in order to better inform policy making and public debate on the subject. Estimating migration has been rightfully classified as the biggest challenge in the area of population studies (Skeldon 2013, Jensen 2012), probably due to the lack of reliable data and the fact that no country has a central international migrant registration system. As has been pointed out, estimates of migration of sufficiently high quality are essential if accurate estimates of current population numbers are to be produced (Jensen 2012, United Nations Population Division 1952). Considering all these factors, in the endeavour to produce current population estimates, there is an inherent need to identify possible sources of data that could be used to improve the estimates of the annual volumes of migration (internally and internationally) and if necessary to develop innovative methods to estimate migration from these sources.

5.4 Available data sources which could be used to estimate post-censal international migration in South Africa

Several data sources exist in South Africa which might be of use in estimating the level of migrant flows in and out of South Africa in post-censal years. These include deportation statistics, border entry and exit statistics, visa application statistics and asylum statistics. Unfortunately, the information systems used to capture these statistics in South Africa do not link entries to exits, making it impossible to identify circular migration and determine whether an increase in the number of border entries indicates an increase in overall migrant stock in South Africa (Kiwanuka and Monson 2009). However, if the records collected are sufficiently complete, the statistics can be used to infer a trend in percentage increase in volumes of overall migrant flows from one year to the next. In any case, there is much uncertainty about the accuracy of estimates that could be produced using these information systems. A report by the United Nations Development Programme concluded that the South African government had not done enough to establish data collection mechanisms to estimate flows of international migrants sufficiently accurately enough to inform migration policies (Landau and Wa Kabwe-Segatti 2009).

5.5 Possible approaches to estimating international migration annually

Due to the difficulties in estimating migration, researchers and national statistical offices have devised various methods for estimating migration (Bonaguidi 1990, Hill 1979, Rogers and Castro 1981, Statistics Canada 2012, United States Census Bureau 2013, Australian Bureau of Statistics 2009, Office for National Statistics 2007). The methods used to estimate international migration mainly depends on the data which are available and the quality of those data. Two approaches are suggested here which could be used to estimate and/or infer volumes of net international migration when there is no annual survey to estimate numbers of international migrants or any sufficiently accurate migrant registration system.

The first approach draws from the ratio correlation method (Goldberg, Rao, and Namboodiri 1964, Schmitt and Crosetti 1954, Swanson and Tayman 2011) to estimate net numbers of international migrants annually. The regression equation used in the ratio correlation method is customised to produce conjectural (not based on numerical data relating to the population itself) estimates of subnational populations. Of course the equation cannot be directly used to produce national population estimates as it would mean making use of ratios with denominators summing international populations

and symptomatic variables, which is not possible. However, using the principle of relating ratios of a dependent variable to ratios of independent variables, we may relate the ratios of population numbers at time *t* divided by the corresponding population numbers at time *t-z* to ratios of counts of symptomatic variables at time *t* divided by corresponding counts of the symptomatic variables at time *t-z*. The formulation for this approach is as follows:

$$ {}_nP_x^t \,/\, {}_nP_x^{t-z} = \alpha_0 + \sum_j \left(b_j * \left({}_nS_x^{j,t} \,/\, {}_nS_x^{j,t-z} \right) \right) + \varepsilon_i $$

where:

a_0 = the intercept term to be estimated

j = the symptomatic indicator $(1 \leq \mathrm{j} \leq \mathrm{k})$

t = year of the most recent census

z = the number of years between the censuses whose data is used to construct the model

b_j = the regression coefficient to be estimated

${}_nP_x^t$ = the population in age group $[x, x+n)$ at time *t*

${}_nS_x^{j,t}$ = the symptomatic variable for age group $[x, x+n)$ at time *t*

ε_i = the error term

The above model uses age-specific population numbers from two consecutive censuses and symptomatic data from the corresponding census times. The method uses age-group specific ratios in order to disaggregate the population and have several data points for the independent variable which is necessary when constructing a multiple linear regression model. Symptomatic variables that could be used include births, deaths, school enrolment, enrolment at higher education institutions and vehicle registrations. Selection of the symptomatic variables would depend on how well the ratios of the symptomatic variables correlate with the ratios of the population numbers. Various age groups could have their own age-group specific multiple linear regression model for the ages within them. This is because data for some symptomatic variables would naturally be available for specific age ranges, for instance school enrolment data would be expected to cover mostly ages 6-18 last birthday, and some fewer cases up to ages in the early twenties, but would not account for movements of single adults and only for

families if they move with children of school-going age. The estimated population numbers for the various age-groups can then be summed to get the estimated total population at time *t*.

Once the coefficients of the independent symptomatic variables have been estimated using census data, the next step would be to use these coefficients with post-census data to obtain a conjectural estimate of the total population. The inherent assumption in the method, as with the ratio correlation method, is that the relationship between the symptomatic indicators and the corresponding population remains unchanged over time. The conjectural post-censal population estimates obtained using the method would be inclusive of net migration, births and deaths. This enables the estimation of aggregate net migration from one year to the next using the demographic balancing equation, expressed simply as:

$$P_2 = P_1 + B - D + NIM$$

where:

P_2 = the population at time two

P_1 = the population at time one

B = births during the period between time one and time 2

D = deaths during the period between time one and time 2

NIM = net international migration during the period between time one and time 2

Since an estimate of the population at time one will be available, we can then use the conjectural estimate of the population at time 2, births and deaths estimated from the CRVS system, and the above relationship to estimate aggregate numbers of net international migrants as follows:

$$NIM = P_2 - P_1 - B + D$$

However, the method is circular because it uses an estimate of the population at time 2 obtained by conjecture to estimate net migration, which is in turn used to estimate the population at time 2 using the cohort component method. Nonetheless,

this could be viewed as an "iteration" to obtain a reasonable estimate of the population at time 2 by making use of the "best" of both methods (population estimate by conjecture versus population estimate by the cohort component method). The conjectural population estimate can be used to provide aggregate numbers of net international migration from one year to the next, whilst the cohort component method can be used to incorporate data from the CRVS system to obtain the final population estimate.

If aggregate net numbers of international migrants are estimated in this way, an age distribution of migrants from the latest nationally representative survey, or the latest census can then be used to disaggregate the numbers into age groups. Of course the method does not adopt the originally recommended principle of the ratio correlation method, to use proportions, but as pointed out earlier, using proportions would not make sense when estimating a national population. However, it would suffice in tracking current migration for the production of mid-year estimates, rather than having to wait for the next decennial census to realise that our projection was wrong. The method could not be used in this research due to the lack of reliable symptomatic variable data for the 1996-2001 inter-censal period which would have been necessary to estimate the regression parameters of the model. Information systems which record data on symptomatic variables have significantly improved which presents better opportunities for applying the method when producing mid-year estimates in the future.

The second approach involves using symptomatic variables to track changes in volumes of migration from one year to the next. In this method, symptomatic variables that are specific to international immigrants would have to be identified. These could include border statistics (deportations, entries and exits), visa and permit applications, asylum applications or any other proxy for international migrants. The recorded data on the symptomatic variables are assumed to be constantly complete from one year to the next.

Time series analysis could be used to fit trends on the symptomatic data series which would be used to project future values of the symptomatic variables. Depending on the nature of the symptomatic data series, a suitable time series model would be fitted. Thus, from one year to the next, recorded values of symptomatic variables would be compared to the expected values estimated using the time series model. If the completeness of the recorded values of the symptomatic variables is constant from one year to the next, the comparison can be used to make inferences about changes in the

level of international migration. For instance, if forecasted numbers of work permit applications are five percent higher than the recorded numbers, and completeness of the work permit application records and the rates at which foreign nationals apply for work permits are constant, then it can be inferred that actual migration was lower than expected.

Extensive research would need to be carried out to identify the best and most consistent proxies or symptomatic variables which can be used to make reliable inferences about international migration. Annual symptomatic variable data would be required to build time series models to calculate expected values. Symptomatic variable data could not be obtained for the years prior to 2001, which could have been used to build time series models to estimate expected values for the post-censal years. Thus, the approach could not be adopted for this research.

Stats SA has recently begun publishing an annual statistical release on documented international immigrants, which contains reviews of data sources for measuring migration and analyses of characteristics of immigrants in South Africa. The first publication was in 2012, which has information on international immigrants recorded in 2011, and the second and most recent was in 2013 (Statistics South Africa 2012a, 2013). Thus, for future production of mid-year estimates, these annual publications could provide the data needed to apply the method. However, Stats SA indicate that the data currently being collected, specifically for work permit application do not have some key variables such as sex of the applicant (Statistics South Africa 2012a). But, they report that it is expected that the Department of Home Affairs will capture these variables in the future. The necessary information systems are already in place to capture symptomatic variable data which can be used to estimate international migration annually. Of course some modifications and improvements are still necessary to improve the quality and accuracy of the data collected. Hence this could be an approach to adopt.

The objective of the approach is to enable timely discovery of changes in volumes of international migrant flows from one year to the next. This would provide a basis for adjusting projected numbers of net international migrants to be added to the base population when producing post-censal population estimates. This would alleviate the inaccuracies caused by under-estimation of migration, and not just wait for the next census to realise that under-estimation of migration distorted mid-year population estimates.

5.6 Limitations of current research

There are a number of limitations associated with using administrative data to produce mid-year estimates and reflecting on them is essential in determining the feasibility of adopting the method as a viable alternative to the method used by Stats SA. As pointed out by the literature, birth registration data have, until this year, been published by Stats SA with approximately a one-year lag, whilst death registration data have been published with approximately a two-year lag.

Obviously the time lags in releasing vital registration data limit the usefulness of the data for producing the mid-year population estimates, particularly in tracking short-term developments, as it means they would be produced at least two years after the date to which they refer.

Another challenge associated with using administrative data to produce population estimates is the quality of the data used to estimate population growth from one year to the next. The method suggested for estimating international migration requires symptomatic variable data which could include births, deaths, vehicle registration, school enrolment, water consumption, electricity consumption and mobile phone usage among others. Such data, primarily collected for non-statistical purposes, would probably be affected by errors and would require extensive analysis and adjustment before it can be used.

5.7 Areas for further research

The limitations associated with using administrative data to produce mid-year population estimates can be mitigated. For instance, the lags in releasing birth and death data can be countered by extrapolating past trends in birth and death registration to estimate births and deaths for a reference year which can be used to produce preliminary mid-year estimates with no lag. In this regard, research needs to be done to develop time series models or regression models, using past data, to forecast births and deaths for a projection year before vital registration data becomes available.

There are also other sources of vital registration data available with much shorter lags than those of Stats SA. The Rapid Mortality Surveillance (RMS) system provides death registration data from the population register with a lag of approximately one year (Joubert et al. 2012). Data from the Department of Health Information system provide birth data on births occurring in health facilities. Thus, research needs to be done on how best to integrate the data from these sources to produce preliminary mid-year

estimates. When vital registration data become available for a reference year, they could then be used to produce revised mid-year estimates which was done in this study.

The method can also be extended to produce sub-national mid-year estimates, although this would not be easy. Stats SA publish registered births and deaths by province, which could be used in the cohort component approach. However, there are several problems associated with place of death data such as that some deaths do not occur in the deceased's province of residence. Some births also do not occur in the mother's province of residence. Research on the numbers of births and deaths not occurring in the affected persons' province of residence and substantial data analysis and cleaning would be prerequisites to using the place of death data to produce reasonably accurate sub-national mid-year estimates. Sub-national migration would need to be estimated perhaps using the ratio correlation method and making use of the population balancing equation to estimate net migration as discussed above. However, sufficient research on whether or not the method produces reasonable estimates of sub-national migration would need to be done first.

The method could also be used to produce mid-year estimates by population group. The main challenge would be that since 1991, the Department of Home Affairs has not categorised births by population group (Statistics South Africa 2012d). Research on how to obtain reasonable estimates of births by population group would need to be undertaken. One method could be to make use of projected fertility rates by population group based on data collected in censuses and surveys to project births by population group. Projected births could then be used to apportion registered births and have estimates of registered births by population group. Deaths are published by population group, although population group is missing for 25 per cent of the death data. This is a substantial proportion of the data and research on how to reasonably apportion the missing data would need to be done. Migration would be estimated using the ratio correlation method. Extensive data analysis and cleaning would also be essential before using death data to produce mid-year estimates by population group.

5.8 Conclusions

Overall, the population estimates produced using administrative data were more accurate, in aggregate terms, than the estimates produced by Stats SA. Although the age distribution produced using administrative data did not match the census age distribution for ages below 45 last birthday, it still had a shape similar to the census age

distribution. Some of the causes of the mismatches were identified and hopefully these will lead to improved population estimates in the future.

Also Stats SA's age distribution did not match the census age distribution, in terms of shape, for ages below 55 last birthday. Furthermore, it is not clear how Stats SA "constructed" (Statistics South Africa 2004, 7) their fertility, mortality and migration assumptions for their projection. In this regard, they do not fully apply some of the recommendations given by the United Nations Population Division to provide full details of the methodology used to produce population estimates.

In light of the observations relating to the shortcomings of Stats SA's MYEs and their publication of these, the following recommendations are made. First, Stats SA should be transparent when publishing their MYEs, as recommended by the United Nations Population Division. They should provide sufficient detail about how they adapt SPECTRUM to produce their MYEs and exactly how they "constructed" (Statistics South Africa 2004, 7) their fertility, mortality and migration assumptions, especially for their national projections. Furthermore, they should either provide more detailed output of their MYEs (numbers by single ages), or publish full details of their model as adapted in SPECTRUM so that it can be possible to reproduce their results and investigate the causes of the identified problems with their MYEs. Second, Stats SA should unpack/give explanations clarifying the dissimilarities and changes in their MYEs from one year to the next by either reconciling the numbers of one year to that of the one before, or explaining the changes in numbers from one year to the next. Third, Stats SA should not change their series from one year to the next. They should produce them in such a way that their MYEs are comparable between different years. Changing their series from one year to the next renders them completely inconsistent, which does not inspire confidence in users of their estimates. Fourth, Stats SA should unpack/explain the differences in the age distribution of their 2011 MYEs and the 2011 census population distribution. It is not enough for them to just acknowledge the differences. In any case, it does not make sense for them to fix their MYEs such that they match the 2011 census population total, and yet ignore the differences in the age distributions.

On the whole, using administrative data to produce mid-year estimates appears to be a feasible endeavour with encouraging prospects. Sufficiently accurate estimates of completeness of vital registration and migration are prerequisites when using administrative data to produce mid-year estimates.

REFERENCES

Australian Bureau of Statistics. 2009. Population Estimates: Concepts, Sources and Methods. edited by P. Harper. Australia: Australian Bureau of Statistics.

Bashir, A., and J.G. Robinson. 1994. Estimates of emigration of the foreign-born 1980-1990. In *Working Paper No. 9*: U.S. Census Bureau.

Beers, H.S. 1945. "Six-term formulas for routine actuarial interpolation." *The Record of the American Institute of Actuaries* no. 34 (1):59-60.

Bennet, NG., and S. Horiuchi. 1984. "Mortality estimation from registered deaths in less developed countries." *Demography* no. 21:217-233.

Blumenstock, J.E. 2012. "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda." *Information Technology for Development* no. 18 (2):107-125. doi: 10.1080/02681102.2011.643209.

Bonaguidi, A. 1990. "Measurement of Emigration Using Indirect Techniques." *European Journal of Population* no. 6:113-116.

Bongaarts, J., and R.A. Bulatao. 1999. "Completing the demographic transition." *Population and Development Review* no. 25 (3):515-529.

Bradshaw, D., and RE. Dorrington. 2007. "Child mortality in South Africa - we have lost touch." *S Afr Med J* no. 97 (8):582-583.

Bryan, T. 2004. "Population Estimates." In *The Methods and Materials of Demography*, edited by J.S. Siegel and D. Swanson, 523-560. Amsterdam, The Netherlands: Elsevier Academic Press.

Central Statistical Service. 1987. Statistical News Release. Mid-Year Estimates: 1970-1986. Pretoria, SOuth Africa: Central Statistical Service.

D'Allesandro, F., and J. Tayman. 1980. Ridge Regression for Population Estimation: Some Insights and Clarifications. edited by Office of Financial Management. Olympia, Washington.

Darikwa, T.B., and R. Dorrington. 2011. "The level and trends of child mortality in South Africa, 1996-2006." *African Population Studies* no. 25 (1):159-172.

Dorrington, R.E. 2013. Alternative South African mid-year estimates, 2013. Centre for Actuarial Research Monograph 13, University of Cape Town.

Dorrington, RE. 1989. African mortality rates - an initial estimate. Transactions of the Actuarial Society of South Africa, Actuarial Society of South Africa.

Dorrington, RE., D. Bourne, D. Bradshaw, R. Laubscher, and IM. Timæus. 2001. The Impact of HIV/AIDS on Adult Mortality in South Africa. Tygerber, Cape Town: South African Medical Research Council.

Dorrington, RE., and D. Bradshaw. 2011. "Maternal mortality in South Africa: lessons from a case study in the use of deaths reported by households in censuses and surveys." *Journal of Population Research* no. 28:49-73.

Dorrington, RE., TA. Moultrie, and IM. Timæus. 2004. Estimation of mortality using the South African Census 2001 data. Cape Town: Centre for Actuarial Research, University of Cape Town.

Ericksen, E. 1973. "A Method for Combining Sample Survey Data and Symptomatic Indicators to obtain Population Estimates for Local Areas." *Demography* no. 10:137-160.

Ericksen, E. 1974. "A Regression Method for Estimating Population Changes of Local Areas." *Journal of the American Statistical Association* no. 69:867-875.

Goldberg, D., V.R. Rao, and N.K. Namboodiri. 1964. "A Test of the Accuracy of Ratio Correlation Population Estimates." *Land Economics* no. 40 (1):100-102.

Goldstein, J.R., and G. Stecklov. 2002. "Long-range Population Projections Made Simple." *Population and Development Review* no. 28 (1):121-141.
Hill, K. 1979. "The Use of Information on Residence of Siblings to Estimate Emigration by Age." *Notas de Poblacion* no. 7:71-89.
Hill, KH. 1987. "Estimating census and death registration completeness." *Asian and Pacific Population Forum* no. 1:8-13.
International Institute for Vital Registration and Statistics. 1981. Major Obstacles to Achieving Satisfactory Registration of Vital Events and the Compilation of Reliable Vital Statistics. Technical Papers. edited by N.P. Powell: International Institute for Vital Registration and Statistics.
International Monetary Fund. 2007. The Special Data Dissemination Standard: Guide for Subscribers and Users. Washington, D.C. 20431, U.S.A: International Monetary Fund.
Jensen, E. 2012. A Review of Methods for Estimating Emigration. U.S Census Bureau.
Joubert, J., C. Rao, D. Bradshaw, R.E. Dorrington, T. Vos, and A.D. Lopez. 2012. "Characteristics, availability and uses of vital registration and other mortality data sources in post-democracy South Africa." *Global Health Action* no. 5 (19263). doi: http://dx.doi.org/10.3402/gha.v5i0.19263.
Joubert, J., C. Rao, D. Bradshaw, T. Vos, and A.D. Lopez. 2013. "Evaluating the Quality of National Mortality Statistics from Civil Registration in South Africa, 1997–2007." *PLoS ONE* no. 8 (5). doi: 10.1371/journal.pone.0064592.
Keyfitz, N. 1971. "On the momentum of population growth." *Demography* no. 8 (1):71-80.
Khalfani, A.K., T. Zuberi, S. Bah, and P. Lehohla. 2005. Population Statistics. The Demography of South Africa. Statistics South Africa. edited by Zuberi Tukufu, Sibanda Amson and Udjo Eric. New York: M.E. Sharp Inc.
Kiwanuka, M., and T. Monson. 2009. Zimbabwean migration into Southern Africa: new trends and responses. Forced Migration Studies Programme Wits University. edited by D. Vigneswaran, T. Polzer and J. Vearey: University of Witwatersrand.
Landau, L.B., and A.W Wa Kabwe-Segatti. 2009. Human Development Impacts of Migration: South Africa Case Study. Human Development Reports. United Nations Development Programme.
Long, J.F. 1993. Population Estimates: States, Counties and Places. edited by USCB. United States of America.
Lopez, A., L. Mikkelsen, R. Rampatige, S. Upham, C. AbouZahr, S. Gamage, D. de Savigny, and A. Schmider. 2012. Strengthening civil registration and vital statistics for births, deaths and causes of death: Resource Kit. World Health Organisation.
Machemedze, T. 2009. *Old age mortality in South Africa*, MPhil thesis, University of Cape Town, Cape Town.
Mbananga, N., R. Madale, and P. Becker. 2002. Evaluation of Hospital Information System in the Northern Province in South Africa. Pretoria: Medical Research Council.
McKibben, J., and D. Swanson. 1997. "Linking Substance and Practice: A Case Study of the Relationship between Socio-economic Structure and Population Estimation." *Journal of Economic and Social Measurement* no. 24 (2):135-147.
Moultrie, T., and R. Dorrington. 2004. Estimation of fertility from the 2001 South Africa Census data. Cape Town, South Africa: Centre for Actuarial Research.
Moultrie, T., and I. Timæus. 2003. "The South African fertility decline: Evidence from two censuses and a demographic health survey." *Population Studies* no. 57 (3):265-283.

Moultrie, T., and I.M. Timæus. 2002. Trends in South African fetility between 1970 and 1998. An analysis of the 1996 Census and the 1998 Demographic and Health Survey. Medical Research Council.
Nannan, N., R. Dorrington, R. Laubscher, N. Zinyakatira, M. Prinsloo, T. Darikwa, R. Matzopoulos, and D. Bradshaw. 2012. Under-5 mortality statistics in South Africa : Shedding some light on the trend and causes 1997–2007. Cape Town: South African Medical Research Council.
National Institute of Statistics and Economic Studies. 2014. *Definitions and Methods - Statistical Operation: Population Estimates* 2009 [cited 06 March 2014 2014]. Available from http://www.insee.fr/en/methodes/default.asp?page=sources/ope-adm-esti.
Nordbotten, S. 2010. "The Use of Administrative Data in Official Statistics - Past, Present, and Future - With Special Reference to the Nordic Countries." In *Official Statistics - Methodology and Applications in Honour of Daniel Thorburn*, edited by M. Carlson, H. Nyquist and M. Villani.
Office for National Statistics. 2001. A Short Guide to Population Esimates. edited by Population Estimates Unit. United Kingdom.
Office for National Statistics. 2007. Making a population estimate in England and Wales. National Statistics Methodological Series. edited by J. Jefferies and R. Fulton: Office for National Statistics.
Ormiston-Smith, N., J. Smith, and A. Whitworth. 2006. An international comparative study on the use of the Cohort Component Method for estimating national populations. United Kingdom.
Penneck, S. 2007. Using Administrative Data for Statistical Purposes. In *ICES-III*. Montreal, Quebec, Canada.
Phillips, H., E. Phoshoko, and M. Cronje. 2004. Fertility levels and trend in South Africa: evidence from the 2001 census of population. Pretoria: Statistics South Africa.
Ravenstein, E.G. 1985. "The Laws of Migration." *Journal of the Statistical Society of London* no. 48 (2):167-235.
Republic of South Africa. 1992. Births and Deaths Registration Act, 1992 (ACT NO. 51 OF 1992). In *35346*, edited by Department of Home Affairs. Pretoria, South Africa.
Rogers, A., and LJ. Castro. 1981. Model Migration Schedules. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Ruotsalainen, K. 2004. Use of Administrative Data in Population Censuses in Finland. TACIS Seminar. edited by Statistics Finland. Paris, France.
Sadie, J.L. 1993. A projection of the South African population, 1991-2011. Research report 196. Pretoria, South Africa: Unisa, Bureau of Market Research.
Schmitt, R.C., and A.H Crosetti. 1954. "Accuracy of the Ratio-Correlation Method for Estimating Postcensal Population." *Land Economics* no. 30 (3):279-281.
Skeldon, R. 2013. Global Migration: Demographic Aspects and Its Relevance for Development. Technical Paper No. 2013/6. New York, USA: United Nations Department of Economic and Social Affairs.
Snow, E.C. 1911. "The application of the method of multiple correlation to the estimation of post-censal populations." *Journal of the Royal Statistical Society* no. 74 (part 6):575-629.
Statistics Canada. 2012. Population and Family Estimation Methods at Statistics Canada. edited by Demography Division. Ottawa, Canada: Statistics Canada.
Statistics South Africa. 1998. Mid-year estimates P0302. Statistics South Africa.

Statistics South Africa. 2000. South African life tables 1985-1994 and 1996. Pretoria: Statistics South Africa.
Statistics South Africa. 2002. Mid-year estimates 2002. Statistical release P0302. Pretoria: Statistics South Africa.
Statistics South Africa. 2003. Mid-year estimates 2003. Statistical release P0302. Pretoria: Statistics South Africa.
Statistics South Africa. 2004. Mid-year population estimates, South Africa 2004. Statistical release P0302. Pretoria: Statistics South Africa.
Statistics South Africa. 2005. Mid-year population estimates, South Africa 2005. Statistical release P0302. Pretoria: Statistics South Africa.
Statistics South Africa. 2009a. Assessment of the Health Information System in South Africa. Pretoria: Statistics South Africa.
Statistics South Africa. 2009b. South African Statistics, 2009. South African Statistics. Pretoria: Statistics South Africa.
Statistics South Africa. 2010. Mid-year population estimates 2010. Statistical release P0302. Pretoria: Statistics South Africa.
Statistics South Africa. 2011. Mid-year population estimates 2011. Statistical release P0302. Pretoria: Statistics South Africa.
Statistics South Africa. 2012a. Documented immigrants in South Africa 2011. Pretoria: Statistics South Africa.
Statistics South Africa. 2012b. Mortality and causes of death in South Africa, 2010: Findings from death notification. edited by Statistics South Africa.
Statistics South Africa. 2012c. Recorded live births 2011. Pretoria, South Africa: Statistics South Africa.
Statistics South Africa. 2012d. Recorded live births 2011. Statistical release P0305. Pretoria: Statistics South Africa.
Statistics South Africa. 2013. Documented immigrants in South Africa, 2012. Statistical release P0351.4. Pretoria: Statistics South Africa.
Statistics South Africa. 2014. A methodology for population estimation at the national and provincial levels: The approach used by Statistics South Africa. Pretoria, South Africa: Statistics South Africa.
Sterns, F.H. 1935. "Methods of Estimating Post-Censal Population for Individual Communities." *American Marketing Jounal* no. 2 (4):224-235.
Stoto, M.A. 1983. "The accuracy of population projections." *Journal of the American Statistical Association* no. 78:13-20.
Stover, J. 2003. AIM version 4. A computer program for HIV/AIDS projections and examining the social and economic impacts of AIDS. Spectrum system of Policy Models. The Futures Group International.
Stover, J., and S. Kirmeyer. 2008. DemProj A Computer Program for Making Population Projections. U.S. Agency for International Development (USAID).
Swanson, D. 1978b. Preliminary Results of an Evaluation of the Utility of Ridge Regression for Making County Population Estimates. Annual Meeting of the Pacific Sociological Association. Spokane, WA.
Swanson, D., and D. Beck. 1994. "A New Short-term County Population Projection Mwthod." *Journal of Economic and Social Measurement* no. 21:25-50.
Swanson, D., and J. Tayman. 2011. On the Ratio-Correlation Method.
Tayman, J., and E. Schafer. 1985. "The Impact of Coefficient Drift and Measurement Error on the Accuracy of Ratio-Correlation Population Estimates." *The Review of Regional Studies* no. 15 (2):1-3.

Timæus, IA. 1993. Adult Mortality. In *Demographic change in sub-Saharan Africa*, edited by KA. Foote, KH. Hill and LG. Martin. Washington, DC: National Academy Press.
U.S. Census Bureau. 1994. Population Analysis with Microcomputers. edited by E. Arriaga, P. Johnson and E. Jamison. Washington, DC: U.S Census Bureau.
Udjo, E.O. 1997. Fertility and mortality trends in South Africa: The evidence from the 1995 October Household Survey, and implications on population projections. Pretoria: Statistics South Africa.
Udjo, E.O. 1998. Additional evidence regarding fertility and mortality in South Africa and implications for population projections. Pretoria: Statistics South Africa.
Udjo, E.O. 1999. A four-"race" model estimating the population of South Africa. Workshop on Phase 2 of Census 1996 Review. Johannesburg.
Udjo, E.O. 2003. Modelling the population of South Africa within the context of HIV/AIDS as a means of evaluating the 2001 census. Monograph of the Epidemiology and Demographic Unit. HSRC.
Udjo, E.O. 2004a. "An examination of recent census and survey data on mortality in South Africa within the context of HIV/AIDS in South Africa." In *After Robben Island: The demography of South Africa*, edited by T. Zuberi, A. Simbanda and E.O. Udjo. USA: M. E. Sharpe Inc.
Udjo, E.O. 2004b. Is fertility information from the 2001 South African population census useable? Monograph of the Epidemiology and Demographic Unit, HSRC.
United Nations Department of Economic and Social Affairs. 2010. Status of Civil Registration and Vital Statistics in the SADC Region. Technical Report. United Nations.
United Nations Population Division. 1952. Methods of Estimating Total Population for Current Dates. edited by United Nations Population Division. New York, U.S.A: United Nations.
United States Census Bureau. 2013. Methodology for the United States Population Estimates by Age, Sex, Race, and Hispanic Origin (Vintage 2013): April 1, 2010 to July 1, 2013. United States Census Bureau.
Van Aardt, C.J., and J.L. Van Tonder. 1999. A projection of the South African Population, 1996-2021. In *Research report no. 270*. Pretoria: Bureau of Market Research.
Wilmoth, J.R., K. Andreev, D. Jdanov, and D.A Glei. 2007. Methods Protocol for the Human Mortality Database. University of California (United States) and the Max Planck Institute for Demographic Research (Germany)

Appendix A: Demographic analysis and adjustment of the base population

Figure A-1: 2001 Projected census distribution and 2001 census distribution, females

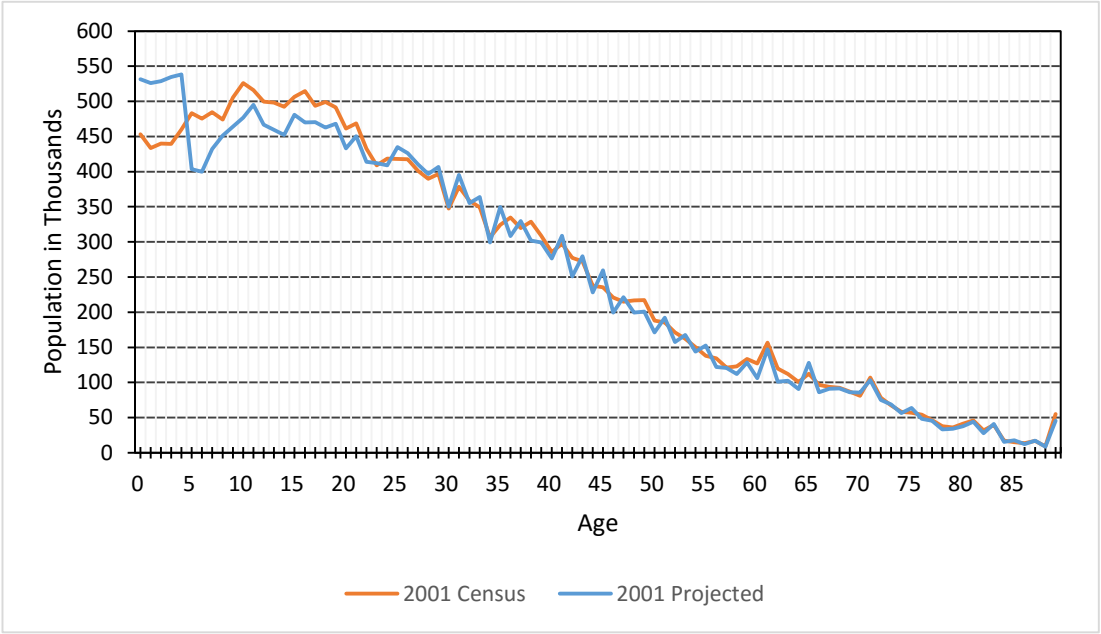
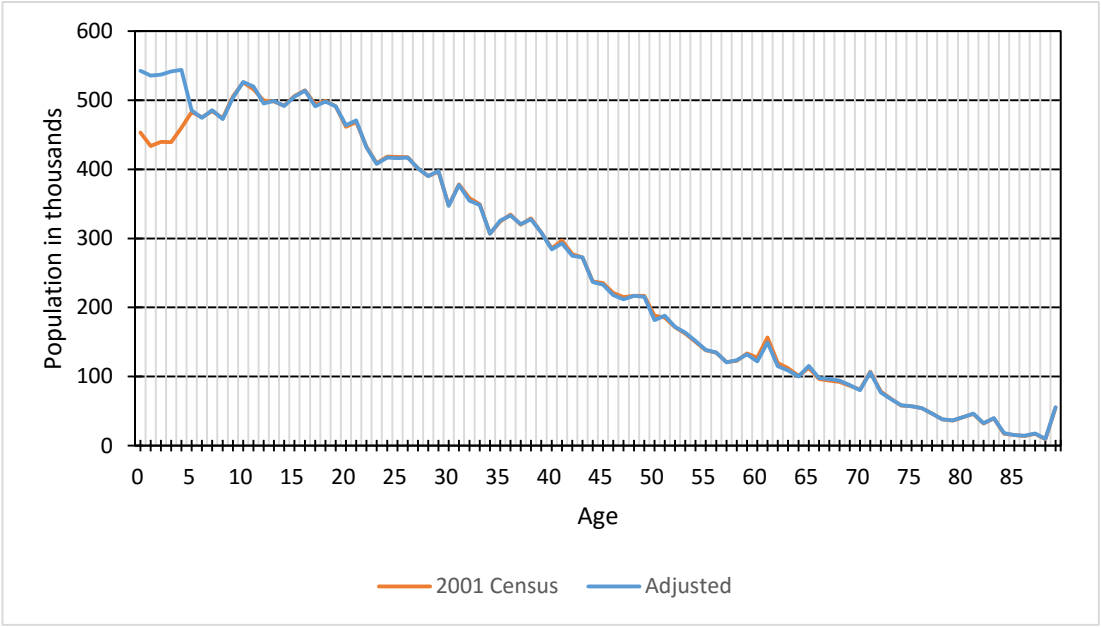


Figure A-2: Comparison of the census female population distribution to the adjusted female population distribution



Appendix B: Mid-year estimates using administrative data

Table B-1: Mid-year estimates using administrative data, 2002-2011

Age group	2002			2003			2004			2005			2006		
	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
0-4	2755476	2719916	5475391	2810868	2775380	5586248	2873309	2838599	5711909	2955697	2921322	5877019	3034450	3001758	6036208
5-9	2467029	2461305	4928334	2540627	2528400	5069028	2595168	2576524	5171692	2654195	2632123	5286318	2712312	2680393	5392706
10-14	2523796	2543811	5067607	2506208	2516008	5022217	2498090	2505306	5003396	2456501	2458231	4914732	2411117	2415007	4826124
15-19	2441991	2497078	4939069	2444933	2495276	4940209	2453705	2496604	4950310	2463612	2500753	4964365	2493637	2520696	5014333
20-24	2164919	2255476	4420395	2260200	2336322	4596522	2324313	2386865	4711179	2376279	2422444	4798723	2420775	2458655	4879430
25-29	1902436	2020267	3922703	1914421	2014015	3928436	1944466	2021665	3966131	1993500	2055499	4048999	2043585	2087666	4131251
30-34	1664376	1803445	3467821	1681470	1818362	3499832	1704701	1833634	3538335	1720360	1841633	3561993	1769562	1879694	3649256
35-39	1408207	1596218	3004425	1406768	1595258	3002026	1419847	1603790	3023637	1446426	1618952	3065378	1451637	1609369	3061006
40-44	1267679	1418859	2686538	1288663	1456536	2745199	1301060	1481526	2782586	1297976	1498224	2796200	1294035	1515129	2809163
45-49	971664	1104780	2076444	1006934	1145999	2152932	1031682	1192571	2224254	1080672	1248855	2329527	1100273	1279276	2379550
50-54	799820	909487	1709307	818990	950456	1769446	832113	975908	1808021	830973	989876	1820848	850269	1025534	1875803
55-59	553520	658384	1211903	576467	687704	1264171	604431	727386	1331817	635307	767112	1402419	655004	793856	1448860
60-64	481882	613777	1095658	481823	615006	1096829	478295	607464	1085760	459400	580844	1040243	458684	588121	1046804
65-69	300196	488933	789129	311468	489119	800586	327556	493498	821054	363447	530681	894129	372829	519697	892526
70-74	237548	401313	638861	236952	411010	647961	229542	413459	643001	217153	389888	607041	225275	413240	638514
75-79	138355	240090	378445	278887	549970	828857	149720	269090	418810	162361	303086	465447	163697	305134	468831
80+	133959	293157	427117				138928	304525	443453	146524	319820	466344	152266	332209	484475
Total	22212852	24026296	46239148	22565679	24384819	46950498	22906928	24728415	47635343	23260384	25079341	48339725	23609408	25425434	49034842

Table B-1: Mid-year estimates using administrative data, 2002-2011 (continued)

Age group	2007			2008			2009			2010			2011		
	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
0-4	3075632	3043354	6118986	3070818	3038182	6109001	3037591	3004080	6041671	2956222	2922670	5878892	2855290	2821181	5676471
5-9	2728318	2696483	5424802	2782624	2750702	5533326	2846165	2815000	5661165	2931407	2900485	5831893	3014207	2984366	5998573
10-14	2456760	2453167	4909927	2529919	2519738	5049657	2584305	2567524	5151829	2643422	2623102	5266524	2701850	2671624	5373474
15-19	2511575	2532072	5043647	2493880	2504551	4998431	2485790	2494064	4979853	2444536	2447243	4891780	2399854	2404496	4804350
20-24	2427605	2455350	4882955	2430691	2455341	4886032	2440535	2460555	4901089	2451931	2468639	4920570	2483993	2493206	4977200
25-29	2119326	2151764	4271090	2215873	2234654	4450527	2280810	2288900	4569710	2334806	2331098	4665904	2383072	2376707	4759779
30-34	1778411	1869097	3647508	1791870	1862895	3654765	1828303	1878630	3706934	1884994	1924027	3809022	1944321	1969773	3914094
35-39	1514436	1670962	3185398	1527408	1681820	3209227	1551184	1698154	3249338	1572282	1713201	3285483	1630149	1760827	3390976
40-44	1261067	1486852	2747919	1258654	1483348	2742002	1274365	1493875	2768240	1305453	1512882	2818335	1316271	1509133	2825404
45-49	1129756	1330347	2460103	1147013	1363956	2510969	1158141	1386859	2545000	1156010	1404310	2560320	1156501	1424641	2581142
50-54	847383	1029623	1877007	879532	1067043	1946575	902631	1111999	2014629	951545	1167591	2119136	972635	1198374	2171009
55-59	688634	842033	1530667	703323	878859	1582182	713582	900811	1614393	711835	913949	1625783	733100	951122	1684223
60-64	462526	596648	1059173	483782	623947	1107729	508884	661315	1170200	534901	697751	1232653	549791	721294	1271085
65-69	390718	536055	926773	388441	536244	924684	383612	527716	911328	365302	502534	867836	367401	513969	881370
70-74	228714	411197	639911	238255	410241	648496	252287	414703	666990	284605	450608	735212	289034	436200	725235
75-79	167591	314363	481954	167807	323151	490958	161019	324274	485293	150050	301213	451263	161337	329078	490415
80+	156239	341421	497659	162422	355825	518247	170447	373639	544086	187282	413249	600531	191039	417613	608652
Total	23944691	25760788	49705479	24272311	26090496	50362808	24579651	26402098	50981749	24866582	26694553	51561136	25149846	26983607	52133453

Appendix C: Comparison of administrative approach with Stats SA's approach

Figure C-1: Comparison of the age distributions produced using vital registration (Administrative) data and those produced by Stats SA; 2003-2011

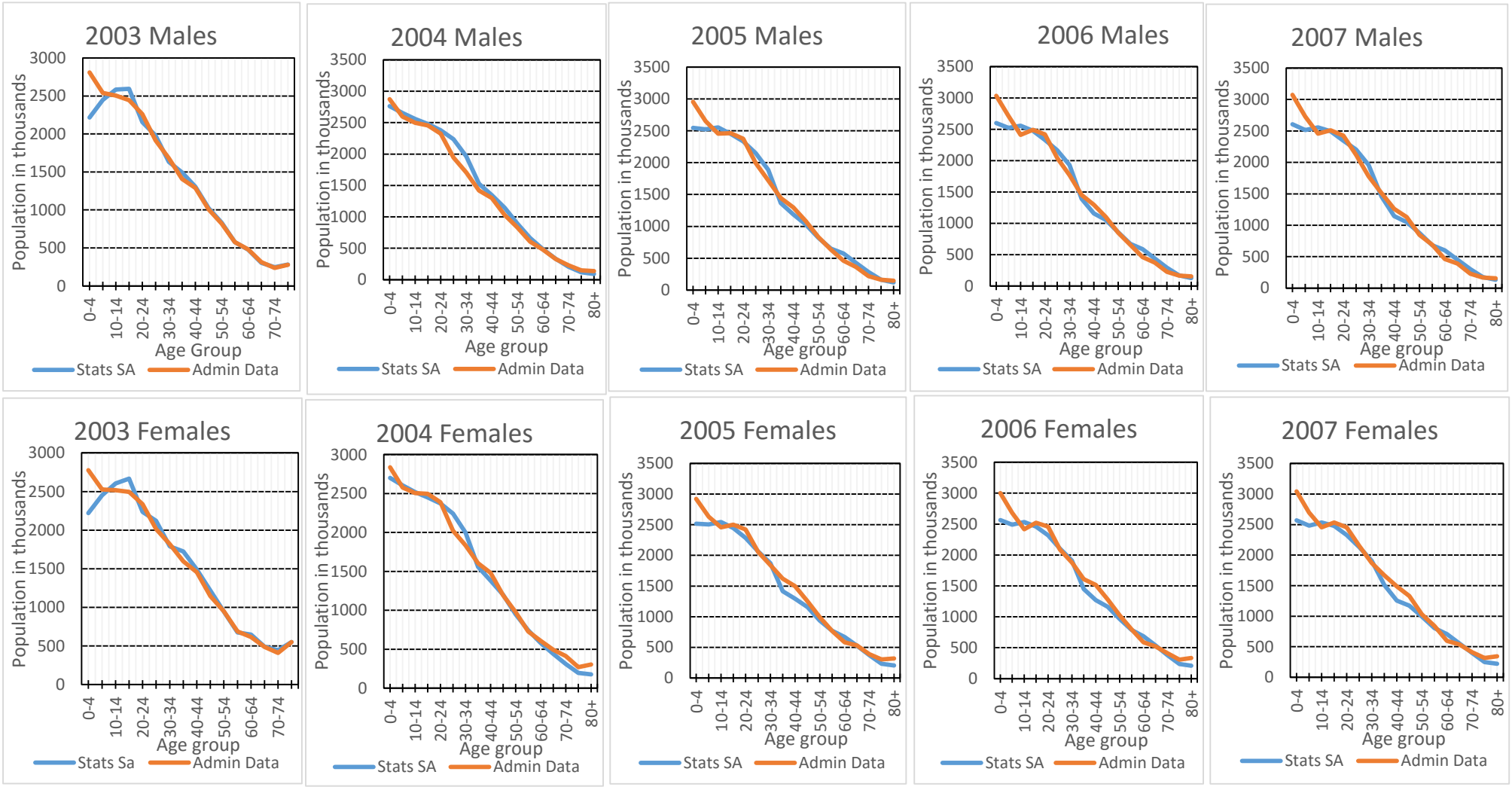


Figure C-1: Comparison of the age distributions produced using vital registration (Administrative) data and those produced by Stats SA; 2003-2011 (continued)

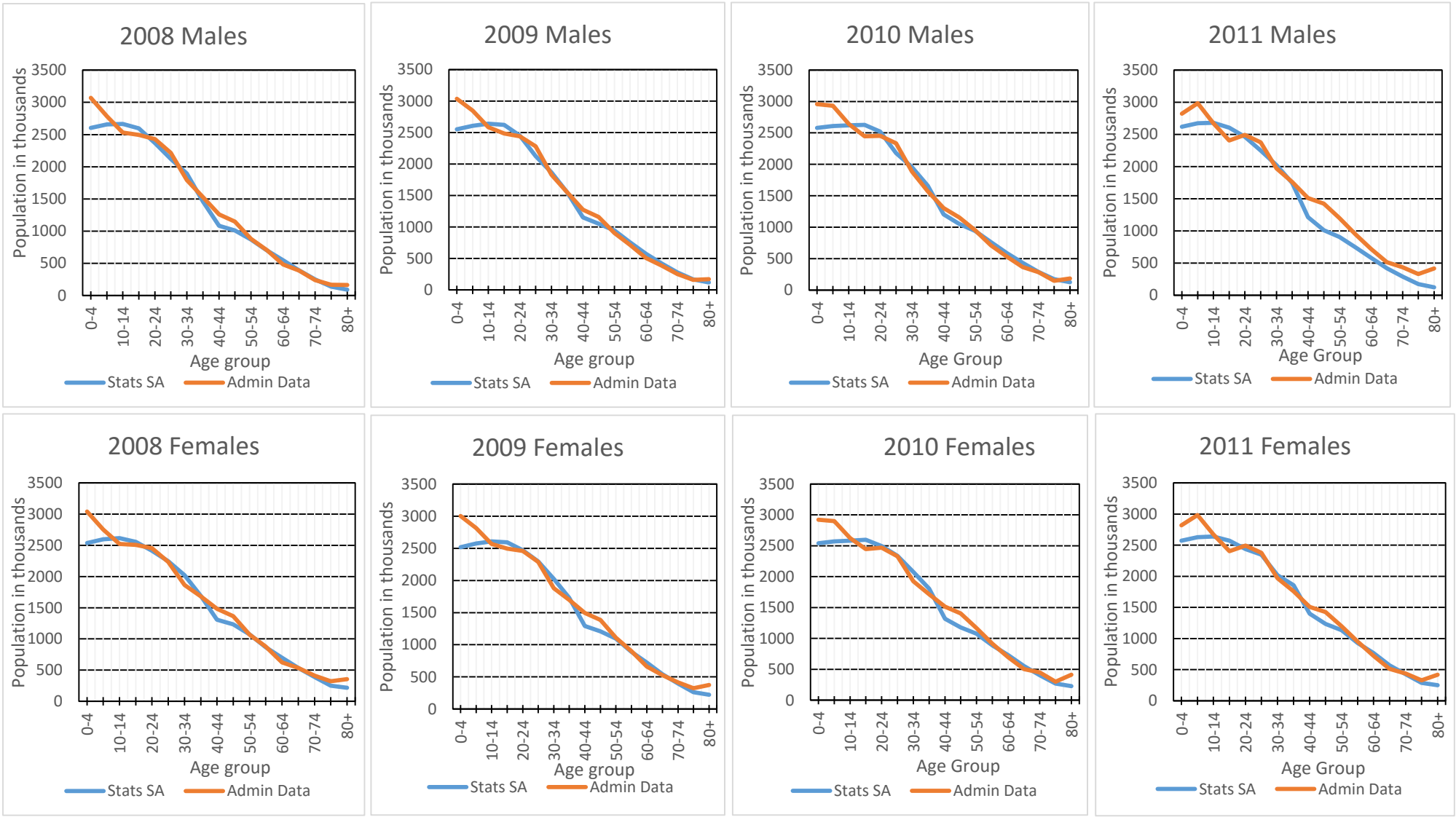
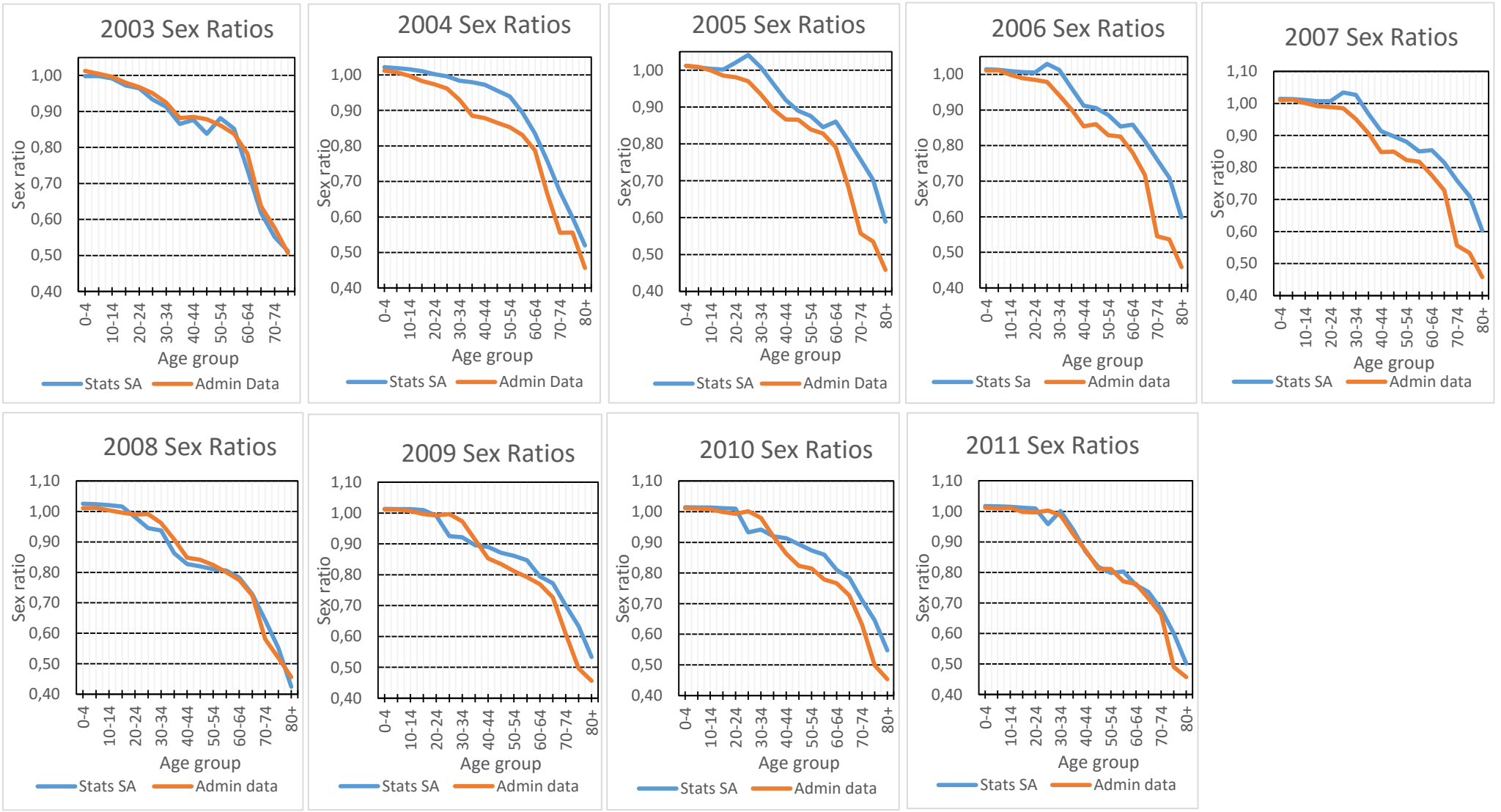


Figure C-2: Comparison of sex ratios produced using administrative data and those produced by Stats SA, 2003-2011



Appendix D: Completeness of birth and death registration

Table D-1: Fitted completeness of birth registration, 2002-2011

Year	Completeness (%) by	
	Year 0	Year 1
2002	39.68	67.06
2003	43.92	73.54
2004	48.38	79.30
2005	52.98	84.18
2006	57.62	88.14
2007	62.20	91.24
2008	66.64	93.61
2009	70.85	95.38
2010	74.76	96.68
2011	78.33	97.63

Table D-2: Fitted completeness of adult death registration, 2001-2011

Year	Males	Females
2001	87.08%	88.88%
2002	88.16%	90.01%
2003	89.09%	90.95%
2004	89.91%	91.72%
2005	90.62%	92.34%
2006	91.23%	92.85%
2007	91.77%	93.27%
2008	92.22%	93.60%
2009	92.62%	93.87%
2010	92.96%	94.09%
2011	93.25%	94.27%

Table D-3: Final estimates of completeness of infant and child death registration, 2001-2011

	Fitted completeness		
Year	Infants	1-4	Under 5
2001	54.06%	42.93%	50.91%
2002	56.17%	42.96%	52.34%
2003	58.25%	42.99%	53.74%
2004	60.28%	43.02%	55.12%
2005	62.26%	43.05%	56.47%
2006	64.19%	43.08%	57.80%
2007	89.33%	58.76%	77.77%
2008	90.78%	60.43%	79.74%
2009	91.87%	62.00%	81.43%
2010	92.69%	63.48%	82.87%
2011	93.30%	64.87%	84.08%