University of Venda

Determination of factors that influence digit preference: A Case study of South African Census 2011 Age-Sex data

---

By

Netshiozwi Masala

Student No: 14001077

This Dissertation is submitted in fulfilment of the requirements for the degree of Master of Science in Statistics at the University of Venda, Thohoyandou

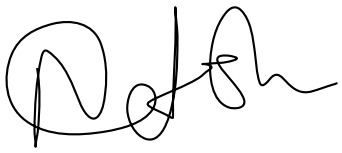
Supervisor : **Dr K.A Kyei**

Co-Supervisor : **Dr S Moyo**

January 2020

DECLARATION

I, Masala Netshiozwi [student Number: 14001077] hereby declare that the dissertation for the Masters of Science degree in Statistics (MSc e-Science) titled "**Determination of factors that influence digit preference: A Case study of South African Census 2011 Age-Sex data**", at the University of Venda, hereby, submitted by me, has not been previously submitted for a degree at this or any other university, that it is my own work in design and in execution, and that all reference material contained therein has been dully acknowledged.

Signature :  Date. 16/08/2020

ACKNOWLEDGEMENT

Glory be to God Almighty for all blessing, courage, strength and his unconditional love.

I am greatly indebted to my supervisors, Dr K.A Kyei and Dr S Moyo for their time, energy and support they invested in this project to be completed in time. To them all, I express my sincere appreciation for their untiring guidance and intellectual alertness throughout the course of the project.

The support of the DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP) towards this research is hereby acknowledged.

To my lovely mother Grace Netshiozwi and my sweet siblings Rendani, Rudzani, Anza and Neki, who always motivated me to pursue my studies and become the best I can be. Thank you for your prayers, constant support and encouragement you gave me through thick and thin.

My dearest friend Maduvhahafani Thanyani, special and heartfelt thanks goes to you. I am so greatful to have a friend like you, who always support, encourages and wishes me the best. We fought this journey together. No weapon fashioned against us shall prosper.

Lucky Daniel and Tendani Mutavhatsindi, your presence since the commencing of the course, is observed and extend my gratitude to you guys.

ABSTRACT

The age distribution of a population is one of the most important demographic factors that plays a major role in describing and making projections about the population. Age distribution determine life expectancy, fertility and migration. It suffers most of the difficulties with regard to its accuracy, due to age misstatement and other factors. The study sought to determine the factors that influence digit preference in Age data using the South African census 2011 Age-sex data. Various methods were applied to examine the objectives of the study. The Visual Inspection methods (Line graph and Population Pyramid), Statistical methods (Age Ratio and Sex Ratio) and multivariate methods (Generalized linear model, Principal Component analysis and Regression analysis) which have been reviewed in detail in the study. This study utilized a full age dataset in single years. Based on the United Nation Age-sex Accuracy Index which was found to be 18.3, it shows that the data collected was of good quality. Besides the results deduced from the analysis to determine the quality of data, the study found that education level, place of residence, gender and ethnic group are the factors that influence digit preference. This was provided as evidence by calculated p-values <0.05, showing a positive relationship for generalized linear model. Principal component analysis and Regression analysis confirm the findings by Generalized linear model.

Keywords: Age heaping, Age-Sex Accuracy Index, Digit preference and Education

Contents

List of Tables

List of Figures

ABBREVIATIONS

dev	Deviation
df	Degrees of freedom
GLM	Generalized Linear Model
KMO	Kaiser Meyer-Olkin
PCA	Principal Component Analysis
OLS	Ordinary Least Square
UN	United Nation

Chapter 1

INTRODUCTION

1.1 Introduction

Age distribution is one of the most important pieces of information that is derived from data, e.g census data [17], and provides the demographic make-up of a society [62]. Not only does it provide demographic make-up,but it also acts as a base for the projection on life expectancy, fertility and migration [15]. It is one of the most vital demographic variables that should be collected and recorded accurately [60]. Age is an important study variable in epidemiological studies and descriptive studies. Descriptive study is a statistical study for identifying a pattern or a trend in a situation. Therefore, age serves as a source-demographic variable related to the host in descriptive studies.

Epidemiology is defined as the study of the distribution and determinant of the occurrence of diseases in human population [61]. For example in health institution, a patient is asked to state his/her age by a doctor and that constitutes an epidemiological information. The accuracy of the reported age is important, as wrong information about age, for instance, could lead, to a doctor coming up with a wrong diagnosis of illness or prescribing wrong medication.

However, in spite of its importance, it is one of demography's most frustrating problems [18], as it is often subjected to errors reported and irregularities, which impact negatively

on its usage in the data. Demographers encounter challenges when analyzing the data; they tend to make wrong predictions due to errors and irregularities in the data. Digit preference forms part of errors in the collected data, due to age being misreported by respondents when reporting their ages [10].

Even though there is a tremendous variation in age reckoning among societies, it is possible to approximate age in one system to that in another, but this could create problems which could result in age heaping. Elderly people and care takers tend to "age heap" too.

Age preference is likely to be used as a result of uncertainty arising from lack of knowledge about date of birth or ignorance of exact age, leading to approximation by the respondents [11, 18, 23] or cognitive biases towards reporting landmark ages leading to 'heaping' of data [45, 53].

Some argue that, it is due to social-cultural (e.g.conscious or sub-conscious of certain ages) values that people in different societies attach to these numbers which consequently results in age misstatement. Others just choose not to reveal their actual age, like rounding up ages. In African countries , illiteracy is very high among the elderly and younger ages, age may be over and underestimated just for preference of certain number [60].

In some cases, intentional reporting biases may arise from efforts to meet age-eligibility cut-offs [21]. Age data plays an important demographic information analysis. This is because most national developmental plans for the provision of public services and goods such as education, employment, food, health, housing and manpower, depend on socio-economic statistics, usually disaggregated by age and sex [44].

Therefore it is crucial that age data is collected accurately, to avoid wrong and inaccurate inferences by demographers, planners, researchers and statisticians.

1.2 Background

Censuses in Africa are mostly characterized by various anomalies. However, there are some other countries which have effectively overcome these anomalies. In Kenya, the first census conducted was in 1948 when it was still a British Colony. Since 1969, censuses have been conducted every 10 years. The recent census in Kenya took place in 2009 and the country was the first in Africa to produce completely processed census within one year [57]. Uganda has also been conducting the population and housing censuses every 10 years since 1948. The latest census conducted in Uganda was in 2014, under the supervision of the Uganda Bureau and was the most complete census ever conducted in Uganda [57].

South Africa has a catalog of census irregularities. The First census conducted in the Union of South Africa was in 1911, wherein the Black population was not accurately counted. In 1950, the homelands population which resided on 13 percent of the land, excluded from the official census of South Africa, was assigned to the black population [29]. Due to the complexity of the laws concerning legal residency, the Black population avoided registering births, divorces, deaths and marriages because Blacks were confined into Informal settlements.

South Africa conducted its first official democratic census in 1996 which represented the first attempt by the government to obtain essential planning information and to do the proper counting in post-apartheid South Africa. Three important issues arose from the census; these were ensuring quality control, improving coverage and undercounts [51]. These issues affected the quality of the data collected meaning that the outcomes of the analysis of that data could not be reliable. It is imperative to evaluate age-sex data from censuses before they are used. This study was conducted in all Nine provinces of South Africa as shown in Figure 1.1.

1.3 Statement of the problem

The age distribution is one of the most vital basic quality characteristics of population composition. It is a crucial component in health and demographic analysis, as it provides fundamental contribution in the projections of the population growth, both at national and sub-national levels and is essential for the analysis of population dynamics.

Despite vital role age distribution plays in the discipline of population studies. It suffers most of the difficulties with regard to its accuracy, as it is often subjected to errors reported and irregularities. This impacts negatively on its usage. Digit preference forms part of errors in the collected data, due to age being misreported by respondents when reporting their ages, ending with certain digits, which results in age heaping.

Demographers, planners, researchers and statisticians encounter challenges when analyzing such data; they tend to make wrong predictions due to errors and irregularities in the data. Hence there is a need for researchers to focus extensively on the main causes of errors that are found mostly within the data. This study aims at finding factors that influences digit preference as a statistical component of the study.

1.4 Objectives /Aims of the Study

1.4.1 Main objective

The overall aim of the study was to determine and analyse factors that influence digit preference, using South African census 2011 Age-sex data.

1.4.2 Specific Objectives

The specific objectives of this study is :

1. To ascertains the level of accuracy of data on age and sex.
2. To examine sex differentials in age misreporting.
3. To examine the relationship between literacy and digit preference
4. To examine the relationship between rural-urban and digit preference
5. To examine the relationship between sex and digit preference
6. To examine the relationship between ethnic group and digit preference

1.5 Hypotheses

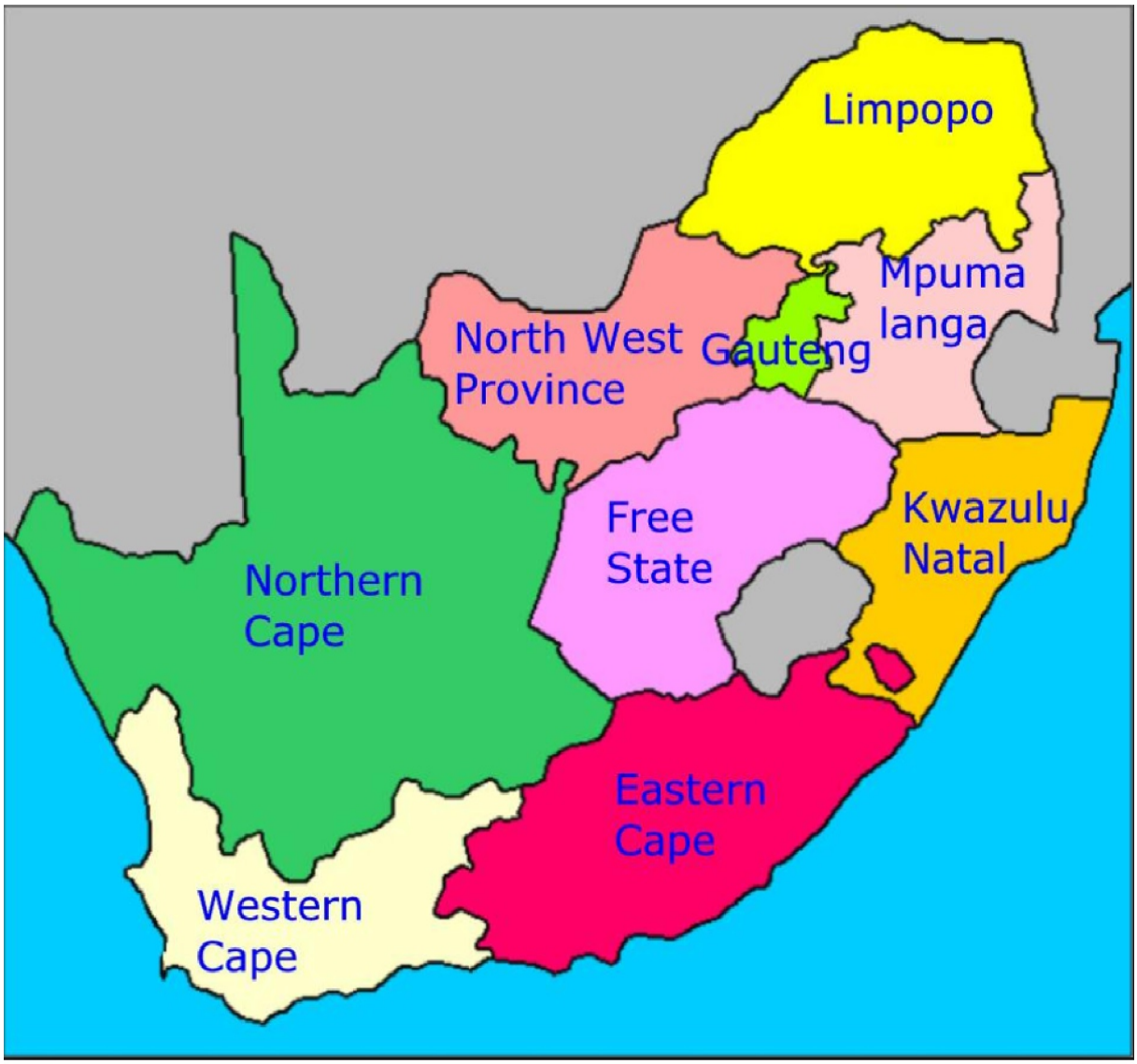
In line with the specific objectives above, the following hypotheses have suited the study at varying degree.

1. The quality of data on age and sex collected from censuses will be of good quality compared to other countries.
2. There is a relationship between gender and digit preference
3. There is a relationship between population group and digit preference
4. There is a relationship between education level and digit preference

1.6 Structure of the study

Chapter	Description of Chapters
Chapter 1	This chapter covers the introduction, Background, Statement of the problem, Objectives or Aims of the study and Hypothesis, and the outlay of the study
Chapter 2	This chapter gives a comprehensive review of the literature that examines problems with demographic data, application of digit preference in many dimension and the concept of age misstatement in different age groups.
Chapter 3	This chapter focuses on the research methodology, application of potentials methods that examines the objectives of this study.
Chapter 4	This chapter provides presentation and analysis of the research results. Data collected is recorded in a meaningful and presentable format.
Chapter 5	This chapter gives a detailed discussion of findings, make conclusion, and recommendations.

Figure 1.1: Provinces of South Africa



Chapter 2

LITERATURE REVIEW

2.1 Introduction

Reports from the analyses of censuses and surveys data, are the basis for reviewing a country's progress, the ongoing schemes of the government and most importantly, plans for the future. They provide valuable information for effective public administration , formulation of policies and planning. The data from this source are used by different parties for different purposes. They are used by business people, international agencies, private and public sectors, and many others.

The data from these sources are used for administrative purposes (for assessing conditions in human settlements), research, commercial and other activities. Age and sex are the basic demographic information collected about individuals in censuses. The study mainly focused on South African Census 2011.

2.2 Importance of Quality of Data on Age and Sex Distribution

Age and sex are the most essential type of demographic information in any survey operation or census. Many studies of population in Africa using census and survey

data reveal the unreliability of the data on age and sex. The various users of age data, recommend commonly the data on age to be gathered by gender. The quality of Age Data is commonly affected by age misreporting.

In every single demographic inquiry, data on age is one of the most vital items that are collected. This is because the age and sex structure of a population provides basic information necessary for providing key insights on social and economic characteristics, and planning. Vostrikova (1970) points out that age composition helps to identify populations for employment, retirement, schooling and voting [59]. Furthermore, Vostrikova (1970) says that sex distribution is important for identifying social characteristics, trends in community structure, and the population's economic potential [59].

Sex is the best fundamental differentiation in humans, because its distinction is based purely on biological organs. For example, It is useful to have information about the extent to which things are different for women and men. Considering factors such as population change in fertility, migration and mortality, to have separate data for females, males and age groups are very useful because they ease analysis according to gender.

Age data also play an important role in the demographic analysis, that cannot be overemphasized. This is because most national developmental plans for the provision of public services and goods, such as education, employment, food, health, housing, manpower and many more depend on socio-economic statistics usually disaggregated by age and sex [44].

Studies have revealed that age and sex structures are useful for making population projections at national and provincial levels, evaluating trends, and for assessing demographic dynamics and economic impacts on the living conditions of its people [3,22]. For example, there was a case of Philippines which is an example that age distribution for 1990 census, presents some distortions and further investigation reveals some forms of misreporting with strong preference for digits 0 and 1 [41].

This result will affect the planning of the population, since individuals misreported their ages, which lead to the wrong predictions that will be made when certain parties are doing their plannings.

Age and sex data are useful in allocating many instances. For example, funds from health programs, particularly those targeting specific age groups [28]. For instance, age data are used to calculate the proportion of school-aged children in each province in order to properly allocate funds for education [28]. Looking at the population of the elderly, data on age provides information such as measurement of people eligible for social security and Medicare beneficiaries [28]. These data can further be applied in the research field when analyzing trends related to mortality.

2.3 Techniques for evaluating and analyzing data on age and sex composition

This section highlights the methods that can be used to check preference amongst certain digits.

2.3.1 Whipple's Index

Whipple Index is a commonly used method in the determination of attractiveness or repulsiveness of particular ages using data from census or survey [25]. This method measures the tendency for individuals to report their actual age inaccurately and tends to prefer ages typically ending with digits '0' and '5' during a survey.

2.3.2 Myers Blended Index

This approach investigates preference for terminal digits ranging from 0 to 9. Myers checks the incidence of each of the digits 0, 1, 2 ...9 separately and compares it with the

sum of population in the total age range between 10 and 69. The calculated index shows preference for or the avoidance of, each of these digits, to determine preferences.

2.4 Review of some Examples in the Evaluation of Age Sex data

This section presents a review of the analyses of age sex data collected in different scenarios. It further shows that digit preference can be practiced in different dimensions.

2.4.1 Digit Preference in different censuses

An evaluation of Age Sex data of the Population Censuses of Sierra Leone : 1963-1985

Bailey and Makannah (1996) evaluated the quality of age and sex data for three consecutive censuses (1963, 1974 and 1985) in Sierra Leone [4]. They used Myer's index in the analysis of their data. In their study, they found that there was a concentration of ages ending with digit '0' and '5', but '0' was more preferred than '5'. The study furthermore found that digit '8' was popularly reported and the most avoided digit was '1'.

Assessing the reliability of the 1986 and 1996 Lesotho census data

Mba (2003) assessed the quality of reported age-sex distributions of Lesotho's 1986 and 1996 censuses using conventional demographic techniques [33]. He used Whipple's Index, Myers Index and Bachi's Index for his analysis. He found that there is a tendency to prefer even numbers and avoid odd numbers. The Whipple's index was 115 in 1986, decreasing to 106 in 1996. The Myers Index also declined from 11 in 1986 to 9 in 1996 and, Bachi's index, in the same manner, decreased from 7 to 6.

This study presents some inconsistencies in the reported age and sex distribution of Lesotho. A pattern of digit preference prevailed in both censuses of Lesotho. In the figures for 1986, digit 0 was most preferred by males and the numbers '8', '6' and '2', preferred by females. In 1996, digit 6 was most preferred by males but females preferred digit 0. Both sexes preferred digit '6'.

Digit preference in Nigerian censuses data of 1991 and 2006.

Dahiru and Dikko (2013) studied Nigerian Censuses for the year 1991 and 2006 [15]. Their study measured the extent of digit preference. The Whipple and Myers indices were used to analyze the data. They reported a Whipple Index of 293 for 1991 and 251 for 2006, while the Myers Index was 62.3 and 67.1 respectively. Those indices prevailed a level beyond expectancy. The study further found a strong preference for digits '0', '5', '8' respectively, and avoidance of digits '1' and '9'.

2.5 Digit Preference:

2.5.1 Health

Age misreporting is not only based on age-sex data, but can also be found in health. Crawford et al. (2000) assessed possible digit preference in the self-reported year at natural menopause [14]. The data was from the cross-sectional telephone interview from the Study of Women's Health Across the Nation (SWAN), a multi-site, multi-ethnic study of women aged 40-55. The study found that the highest frequency occurs for digits 4 and 5.

Thavarajah et al. (2003) examined the traditional hypertension blood pressure in 2003 [55]. They found that the traditional blood pressure (BP) is subjected to observer's error, such as single number preference and terminal digit preference, which leads to an inaccurate measurement. In their study, they quoted the general medical and hospital-based clinics report as a high percentage (60%-90%) of terminal BP.

Locker and Mason (2006) reported that digit preference is intended for health care and health service management [32]. The recording of the time the patients arrive and leave the emergency department, is affected by digit preference bias. Their results also revealed that some departments show the tendency of digit preference (0 and 5) bias in recording the time of departure from the emergency department. Such bias may cause difficulty in examining changes in the performance of the department.

2.5.2 Marriage

Digit preference for a particular number normally shows up as a measurement of errors found in some data. Mill(2000) stressed out that when respondents themselves report their timing of demographic events and age, there is a risk that they round off to a number ending with terminal digit 5 or 0 , or an even number [34].

Ghosh (1967) in his study of age at first marriage in India's West Bengal, established that unmarried women aged 15 years, tend to misreport their ages which affect age distribution of unmarried females [19]. Coale and Demeny (1966) studies [18], established a so-called African pattern (typical of populations in Africa and Southern Asia) in which females were grouped by a "surplus" of 5-9 years, and a deficit in the adolescent age intervals (10-14 and 15-19 years), followed by a surplus in the central ages of child bearing (25-34 years) [12].

The term "temporality" has surprising and vital influences. Mills (2000) stressed that temporality affects the actual demographic behavior. Some of the possible illustrations are seasonal cycles in weather (climate) conditions and social norms about the proper timing of life events. Mills (2000) hence, stated that everyday life is composed of institutional calendars and a certain date on calendar influences life course behavior [34].

Most of celebration in different countries are guided by ages. For example, in Sweden, turning an age that end with terminal digit 0, shows a larger celebration such as weddings,

which is a privilege to that particular country [34].

Coale and Demeny (1966) in their study, found that exaggeration of ages of girls (10-14) years is due to puberty stage which influences understatement or over reporting of those age group [12]. Furthermore, their study revealed that there is a tendency to exaggerate the ages of women 15-29 years, which is resulted by probability to make their ages consistent with age at fertility and marriage [12].

2.6 Problems experienced in collection, processing and reporting of data

This section reviews some challenges that impact age misstatement. It reviews how tradition, old age and uncertainty affects age-sex data.

2.6.1 The Problem of Proxy Reporting

In a household interview, there is frequently one respondent who supplies the majority of the information (usually the head of the household). In societies where one's own age is not important, the ages of others may seem even less important as well. Van de Walle (1968) has shown that all African demographic surveys or censuses share the problem of trying to record the ages of people who do not know their exact ages [58].

Interviewees generally have some knowledge about their age, despite the fact that it might be non-numeric. Most problems of age misreporting begin with the interviewee's ignoring to state exact numbers [18]. This takes two forms: (i) ignorance of one's own age, and (ii) ignorance about the ages of others about whom questions are asked [18].

This phenomenon of ignorance of exact age, results in distortions in the age data. The distortions range from omissions, due to understatement of certain ages to heaping on another terminal digit. Reasons given for the occurrence of these phenomenon in developing

countries includes, a high illiteracy rates, ignorance of age in the sense of the completed number of years, deliberate misstatement, and inability to understand the question asked [30, 38].

Most researchers agree that the age data incorporated by national population censuses may have some irregularities in age reporting. These errors are exceptionally consistent in numerous developing countries when contrasted with developed countries. Age distribution of the population is typically distorted by digit preference and digit avoidance, which are the irregularities in age reporting.

The studies conducted in Africa, Asia and Latin America, Coale and Demeny (1966) discovered two common patterns of age misreporting in these populations, namely as the African-South Asian pattern, which showed more distortions in male than in female populations, whilst in Latin American populations the converse is true [12].

Ueda (1976), stated that one of the major types of errors most commonly found in the sex-age data derived from censuses or similar surveys, is the false reporting of age [56]. In many cases, the erroneous reporting of age is attributable to the ignorance of the respondent. In most cases, ages are being reported on some particular digits, 0 and 5.

Shryock et al. (1973) indicates that studies on age reporting error need special attention since the errors in the age distribution, particularly in the censuses, are examined more intensively than any other information [50].

Mukherjee and Mukhopadhyay (1988) found age heaping in terminal digit 0 and 5 in their study using Turkish Census [35]. Kabir and Chowdhury (1981) analyzed census data of Bangladesh. They found the preference of digits 0 and 5 when reporting ages, with subsidiary heapings at ages ending at 8 and 2 respectively [27].

2.6.2 Uncertainty in First-Person Reports

In societies in which birth registration is uncommon and in which exact ages are not important, there is often a fundamental lack of knowledge about age. This was clearly expressed by one Moroccan woman in this exchange with a census enumerator [47]:

"What is your age?".
"Who is me?"
"Our generation was unrecorded".
"We didn't have any. No date of birth. Nothing."
"How many (years), how many? Estimate?".
"How am I going to estimate?"
"I have nothing to estimate with. I can tell you that I am 60 years, 70"
"I haven't yet reached. Have you reached 80?·
"I don't think so. Someone who is 80 is...?"
"You who still have energy, you are 70."
"Perhaps that, perhaps it is correct, sir."

Quandt (1974) found many of such expressions that indicated the respondents had been unaware of their own or ignorance of exact ages. Phrases such as "I don't know", "perhaps", and "he might be ...", were found in the responses of 69 percent of the households in the sample taken [47].

Gibril (1979) reports exactly the same types of expressions found by [47] in his study of recordings of census interviews from Gambia [20] . It appears that age misreporting in large part, is due to respondents' ignorance of exact ages and dates. This problem is compounded by third-party reports of age [18].

The information that respondents have about age, takes many different forms. In some cases, it depends on birth registration cards or other official documents that gives at least an estimated age for the person at some particular date in the past. This may incorporate

school registration documents, work permits, or village register books.

Such sources of information were very important in Morocco. The report by [47] shows that people often responded to the age questions by searching for official documents:

"How old is (your wife)?"
"Wait, I will look for her identity card".
"You don't have a paper for her?"
"Who do I work for, that I would be able to have papers?"
"Look, sir, in these papers if there is something..."

Documents that provide age estimates are most likely deceptive, because they do not guarantee the accuracy of information contained within them. The respondent is also likely to lack knowledge of the time in which the document was prepared, this also causes the enumerator to guess or use the uncertain information given by the respondent [18].

2.6.3 The Influence of Tradition and Beliefs

Tradition and beliefs also have impact on age misstatement in some parts of the world. In the method of age-reckoning for traditional Chinese or Muslim community, some authors pointed out that the Chinese traditional method of age-reckoning differs from the international method in a systematic manner [26, 49].

According to Shryock et al. (1973), "the causes and patterns of age or digit preference vary from one culture to culture. However, preference for ages ending in '0' and in '5' is quite widespread", particularly in the Western world [50].

On the day a child is born, he or she is already considered as one year old. If a child was born one week before the Chinese New Year, after two weeks the child will be two years old. In some Asian countries such as Korea and China, there is sometimes preference for ages ending in 3. This is because of the numeral 3 sounds like the word character for life

or word. However, they would avoid the number 4 because it has the same sound as the character meaning death or word.

Interviewers have some motivation to shift the ages of women who are within the boundaries of 15 to 59 interval to be below or above the minimum age of respondent eligibility. There may also be some shifting of births to be outside the maximum age of eligibility for the health questions [46].

2.6.4 Old Age and Age Misreporting

Some researchers have found that a higher tendency for age heaping occurs in the older age category of the population. Nagi et al. (1973) revealed that age heaping is a major source of inaccuracy in the age statistics in many of the developing nations on the African continent, particularly among Islamic populations [37]. This phenomenon was found to be more articulated among women than men, and it tends to increase with age.

Age misstatement is associated with literacy and low educational attainment. For example, Elderly groups misstate their ages if they are not literate. They tend to assume the age that well fit their description. Dechter and Preston (1991) concluded that misreporting is most severe at an older age [16]. They discovered evidence of very pervasive overstatement of age at advanced ages.

According to Hill et al. (2000), age inconsistencies tend to increase slightly with age among individuals aged 85 years and above [24]. Although it is apparent for both sexes, this age pattern is more pronounced in males. Rosenwaike and Logue (1983) additionally established that for extremely old persons, the older the age at death reported on the death certificate, the greater average error plotted against reported age at death which yields exponentially [48].

In addition to age misstatement, it seems to show that there are frequent errors due to exaggeration of the ages of older persons with evidence in U.S. censuses by [36].

2.6.5 Age Misreporting For Infants and Young Children

The ages of infants and children are often reported accurately than the ages of adults. The higher accuracy is due to the fact that ages of children are generally reported by parents or other adults who remember the birth of the child. However, some people tend to guess the age of a child, using his/her rapid physiological and psychological changes during childhood [18]. However, this is not accurate.

Bairagi et al. (1982) pointed out that misreporting occurs in the early age of the population especially in the rural area [5]. The misstatement for young children in rural Bangladesh increases monotonically with age and systematic errors in age misstatement display modest overstatement for the first years of life and more pronounced understatement for ages 4, 5 and 6 of the population are found. The census taken in Ghana also underwent a similar evaluation exercises. These can be seen in the work of Tekpli (2013) [54].

Age misreporting for children is commonly found within parents that tend to round off their child's ages, rather than truncating it to the number of completed years [18]. For example, in some cases, parents report their child as 2 years if the child is 2 years 6 months, due to the fact that they round of the child's age by 1 year

2.6.6 Data Processing Errors

Once the data have been collected, they must be processed before analysis can begin. At this stage there are several ways in which errors can be made. First, errors can occur any time the data are transferred from one form to another. For example, it can be during coding or keypunching. A common keypunching error, involves typing a '2' in place of a '5' or vice versa and can transform a 50-year-old mother of six into an apparently very young lady of 18-years-old.

In the 1960 U.S. census, Steinberg (1966) showed that about 10 percent of the undercount and 40 percent of the over-count were caused by data processing errors. Secondly, errors arise when missing values are replaced by statistical procedures, as in the attribution of

characteristics, or when data cleaning procedures reject individuals whose characteristics are in fact very different from the norm [52]. A third source of error is the loss of data or the specious inflation of data through double counting of questionnaires, which is not apt to affect the age distribution because the errors introduced are random. Finally, errors can arise during the tabulation of the results.

Although all four of these problems occur in almost every survey, there is little evidence of their documentation. Two studies on data processing errors in the U.S. censuses of 1950 [2, 13] have shown a similarity in how processing errors can significantly alter the analysis even when those errors are very rare.

Rare processing errors are the most important when they change the reported characteristics of an individual so that person moves from a large category to a small category [18]. For example, in the 1950 U.S. census a slight error in keypunching could alter a white child of the head of the household into a male Indian [18].

Another common error that arises during data processing is the confusion of blanks and zeros. Although this is a well-known problem, it still occurs. A similar problem can also arise when a code for "unknown" or "not reported" is accidentally included in the open-ended category. This can happen with age when a code of 88 is used to indicate persons whose age is unknown; a simple processing error can also include these observations in the last age group, for example, 75 and over.

All of these errors, attest to the fact that there are inherent errors in demographic data which are introduced during data collection and processing.

2.7 Conclusion

This chapter gave a literature review, to help the reader understand the various theories behind the proposed study. Importance of Age sex data was discussed, to give an overview of how age is important, because it projects aspects such as life expectancy, fertility and migration. This literature review, also discussed errors encountered in age-sex data and how it negatively impacts reporting in different scenarios, where digit preference was the most profound error individuals tend to practice. The review furthermore discussed methods that can be used to check errors in reported ages. Most of the studies showed a significant preference for digits 0 and 5 when individuals are reporting their ages.

Chapter 3

METHODOLOGY

3.1 Introduction

This chapter highlights the key methods that were applied in the analysis of the study. The study instrument was the South African Census 2011 Age-sex data. The Application of Visual Inspection/ Graphical, Statistical and Multivariate methods were used in the analysis. The Sex Ratio, Age Ratio and Age-Sex Accuracy Index method were used to evaluate the quality of data. The Multivariate methods such as Generalized Linear model, Principal Component analysis and Regression analysis were techniques used to determine factors that influence digit preference. All of these mentioned methods were administered to the study instrument.

3.2 Type and Perspective of the Study

The design of the current study on determining factors that influence digit preference in South Africa census 2011 Age-Sex data followed a quantitative perspective, and more specifically, the study made use of a normative comparison design. The aim of the study was not only to determine factors that influence digit preference, but also to recommend innovative ways of collecting reliable data. This study was based on demographic characteristics such as age, ethnic group, gender and place of residence.

The means of measurement of variables in the study were defined, and the statistical methods were used to evaluate the observed findings as well. Thus the significance of the study was descriptive analysis, which highlighted the different variations observed. By taking into consideration the demographic variables and other related factors, the study revealed the structural relationship between digit preference and other variables that were considered.

3.3 Variables that were Analyzed

The main variables which were explored for the research instrument were age, sex, place of residence and ethnic group. The following subsection gave details per variable, on how data for the instrument were collected.

Instrument: South Africa Census 2011 Age-Sex data was used as the instrument for the assessment of the variables involved.

3.3.1 Census Variables

The study selected variables according to the variables used in the Census when collecting the data. The following variables were used.

3.3.1.1 Age

Age was collected by asking the participant " What is your Age in completed years?". The intention was to obtain the exact age of the respondents. For children aged less than a year, the enumerator or respondent captured 000. Then, the ages were captured and recorded into groups as : 15-19, 20-24, 25-29, etc.

3.3.1.2 Sex

The question used to record the gender of the respondent was: "Is (person) a male or female?" and two options are available; namely, 1 = Male, 2 = Female. The sex of the respondent was not assumed by referring to physical build or name.

3.3.1.3 Provincial level

Places of residence were captured as follows. The question asked to gather information of the respondents' usual place of residence was: "Where does the person usually live?". This was recorded according to the nine South African provinces as follows: Western Cape (1), Eastern Cape (2), Northern Cape (3), Free State (4), KwaZulu-Natal (5), North West (6), Gauteng (7), Mpumalanga (8) and Limpopo (9).

3.3.1.4 Ethnic Group

Ethnic group plays an important role in obtaining a more meaningful picture of changes which have occurred in the population. This variable is attached to national picture. This enables the measurement of a self-defined racial identity. This study carried out some analysis on all population groups recorded: (1) African/Black; (2) Coloured, (3) Indian/Asian and (4) Whites.

3.3.1.5 Educational Level

Education of the respondent was collected by asking the history of education status to the respondent. The following were recorded: (1) No schooling, (2) Primary level, (3) Secondary level and (4) Higher education.

3.4 Source of Data

This study used a full dataset (Census 2011) which was acquired by requesting it from Statistics South Africa. This study was conducted on data from all the nine provinces of South Africa. Application of Graphical, statistical and multivariate methods were employed in the full dataset.

3.5 Visual Inspection/ Graphical methods

This section described the methods and procedures that were used for data analysis. The Graphical, Statistical and Multivariate methods were employed to the dataset. Single year age-sex data of Census 2011 was evaluated as described below.

3.5.1 Line Graph

The single age data was graphed in a line form, to check whether there exists some evidence of observed unusual concentrations at some digits. The unusual concentration was measured by checking erratic fluctuations depicted by peaks and troughs. The absence of a smooth descending curve with high ages in the graph could mean there were probable errors in the data. Further, inaccuracies detected were quantified by means of statistical methods.

3.5.2 Population Pyramid

Population pyramid by 5-year age group is the other way of examining irregularities in the reported age data. In a population with no extreme changes in fertility, migration, and mortality, where migration is either negligible or does not occur at selective ages, change in population in given particular ages is expected to decrease smoothly with a decrease in age. The base of the pyramid is mainly determined by the level of fertility in the population, while how fast it converges to the peak is determined by previous levels of mortality and fertility. Moreover, migration by age and sex also influences the shape of the pyramid. The population pyramid was used to check irregularities in the data.

3.6 Statistical techniques of measuring Accuracy of Age-Sex data

3.6.1 Sex Ratio

Sex ratio is defined as the number of males per hundred females in each group. It was used to compare the balance between the two sexes in different population groups, regardless of location, place, size and time. The sex ratio does not normally fluctuate from one period to another, unless there has been a major deviation from a smooth pattern of population growth. High or low sex ratio, indicates a precise numerical deficiency or dominance of males. This is typically driven by a high/low sex ratio at birth, net migration, and the effect of mortality on the sexes.

Sex ratio at birth is usually around 104 because of the biological fact that male births generally exceed female births. Then, it declines gradually with age, due to lower mortality of females. Sex Ratio was calculated as follow:

$$\text{Sex Ratio} = \frac{100 * M_{x,x+4}}{F_{x,x+4}}$$

Where:

- $M_{x,x+4}$:Male Population in a given age group
- $F_{x-5,x-1}$:Female Population in a given age group

The sex ratio was also employed in the accuarcy index to apprend the quality of data by multiplying it three times.

3.6.2 Age Ratio

Age ratio is usually defined as the ratio of the population in a given age group to one half of the population in the two adjacent age groups. Age ratio is equal to a hundred in the assumption of absence of errors in the age reporting. Age ratios serve as a primary measure of net misreporting. Nevertheless, age ratios should not be taken as errors for a

particular age group.

Age ratio defined for any group x to $x+4$ ($P_{x,x+4}$) is :

$$Age\ Ratio = \frac{100 * P_{x,x+4}}{{}^{1}/_{2}[P_{x-5,x-1} + P_{x+5,x+9}]}$$

$$= \frac{P_{x,x+4}}{P_{x-5,x-1} + P_{x+5,x+9}} \left[\frac{100}{{}^{1}/_{2}}\right]$$

$$= \frac{200 * P_{x,x+4}}{P_{x-5,x-1} + P_{x+5,x+9}}$$

Where:

- $P_{x,x+4}$: Population in a given age group
- $P_{x-5,x-1}$: Population in the preceding age group
- $P_{x+5,x+9}$: Population in the following age group

Age Ratios for 1^{st} and last groups cannot be calculated using this method, as there is only one adjacent age group corresponding to each one of them. The computed age ratios are then compared with the expected ratio, which is usually 100. The deviation at each age group is a measure of net age misreporting.

3.6.3 Age-Sex Accuracy Index

As an extension of evaluation of data, the United Nation proposed an age/sex accuracy index or Joint Score, to appraise data quality of the data collected. This technique which was proposed by the UN, employs age ratios and the sex ratio. The UN Joint Score method, also known as UN accuracy index will be used to check the quality of data for age-sex data.

The UN Joint Score is a combined or composite indicator of the overall age displacement in a five-year age group and differential errors in age misreporting between males and females. The computation is done for five-year group ages. The UN Joint Score or UN accuracy index is defined mathematically as follows:

UN Joint Score $= ARS_m + ARS_f + 3\text{SRS}$

Where:

- ARS_m : Age Ratio Score for males
- ARS_f : Age Ratio Score for females
- SRS : Sex Ratio Score

Table 3.1: The UN recommendations

Values of UNJS	Quality of data
Below 20	Accurate/reliable
20 to 40	Inaccurate; useable with adjustment
40 to 60	Inaccurate; useable with massive adjustment
Above 60	Highly inaccurate; risky to utilize

3.7 Multivariate Techniques for determining factors that influence digit preference

3.7.1 Generalized linear Model

The generalized linear model describes the pattern of associations and interactions. The parameters of the model measure the strength of associations. This model was used to examine the relationship between the dependent variable and the independent variables that answer the specific objectives of the study.

Dependent variable: Digit preference
Independent Variables: Education level, Province, Population Group and Sex

There are three components of GLM known as:

(a) Random component - refers to the probability distribution of the response variable (Y). This is also called a error model or noise model.

(b) Systematic component - specifies the explanatory variables $(X_1, X_2, ...X_k)$ in the model. The linear predictor is defined as: $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$

(c) Link function (η or $g(\mu)$) - is defined as the relationship of expected value of response to the linear predictor of explanatory variables. It is expressed as follows :

- $\eta = g[E(Y_i)] = E(Y_i)$ for linear regression
- $\eta = logit(\pi)$ for logistic regression

The model is developed based on the following assumptions:

- The variables $Y_1, Y_2, ...Y_n$ are independently distributed
- The dependent variable Yi assumes a distribution from an exponential family (e.g. Binomial, Normal and Poisson)

- GLM assumes linear relationship between the explanatory variables and the transformed response in terms of the link function : e.g. for binary logistic regression $logit(\pi) = \beta_0 + \beta X$.
- Independent variables are assumed to have power terms
- Errors are not normally distributed but need to be independent
- Maximum likelihood estimation (MLE) is preferred mostly over ordinary least squares (OLS) to estimates the parameters.
- Goodness of fit measures replies on sufficiently large samples. It follows a heuristic rule that not more than 20% of the expected cells are less than 5

3.7.2 Principal Component Analysis

Principal Component analysis is a dimensionality reduction technique used to reduce a large set of variables to smaller set that still contains most of information in the large set [31]. It explains the correlation structure of a set of predictor variable using the reduced set of linear combination [9].

PCA is a useful technique where explanatory variables are closely related (i.e multicollinearity). PCA transform into K uncorrelated variables, if there exist some evidence of k explanatory variables in the regression model [8].

The original explanatory variables are denoted by x_1, x_2, ... , x_k
The principal components are denoted by p_1, p_2, ..., p_k
The principal components are independent linear combinations of the original data:

$$p_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + ... + \alpha_{1k}X_k$$
$$p_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + ... + \alpha_{2k}X_k$$
.
.
.
$$p_k = \alpha_{k1}X_1 + \alpha_{k2}X_2 + ... + \alpha_{kk}X_k$$

α_{ij} represents the coefficient of the jth explanatory variable in the ith principal component. These coefficients are known as factor loadings [8]. The sum of squares of the coefficients for each component is required to be one [9]. The principal components are derived in descending order of importance.

The principal components is also expressed as eigenvalues of (X′X), where X is the matrix of observations on the original variables. If the ordered eigenvalues are denoted by λ_i(i = 1, ... k) :

$$\theta = \frac{\lambda_i}{\sum \lambda_i}$$

the ratio gives the proportion of the total variation in the original data explained by the principal component.

In order to compute PCA, some measurements have to be followed before any further analysis:

(a) Correlation between the variables must be greater than 0.3

(b) The Kaiser Meyer Olkin and Bartlett test must be conducted.

The Kaiser Meyer Olkin:
The test measures how suited the data is for factor analysis using a sampling adequacy for each variable in the model. Proportion of variance among variables that might have common variance is measured by the statistic. KMO return a values between 0 and 1 .
Bartlett's Test:
It tests the null hypothesis that the correlation matrix is an identity matrix; the variables

Table 3.2: A rule thumb for interpreting the statistic

Value	Meaning
KMO ≈ 0	There are large partial correlations compared to the sum correlation
KMO < 0.6	The sampling is not adequate and remedial action should be taken
$0.8 \leq$ KMO ≤ 1	The sampling is adequate

are uncorrelated. The statistic for this test is the p-value, which explains that a small value would indicate evidence against null hypothesis (i.e the variables are correlated). For p-values greater than 0.1, there is evidence that the variables are not correlated. Variables should be correlated so that principal component can be conducted.

3.7.3 Regression Analysis

The term "regression analysis" describes a collection of statistical techniques which serve as the basis for drawing inference as to whether or not a relationship exists between two or more quantities within a population or within a system.

In order to determine factors that influence digit preference, regression analysis was employed. The primary essence of regression analysis was to analyze the relationship between dependent and independent variable. This technique of analysis then exploits the association between these two variables to predict the values of the dependent variable from the independent variables.

The dependent variable is expressed as a function of the independent variables and its corresponding parameters plus a stochastic error term. The stochastic error term accounts for all unobserved independent variables that would have had a significant impact on the dependent variable.

There are variety of regression methods, but this study employed the ordinary least square method. The choice of OLS method was due to primary purpose of evaluating the relationship between a set of independent variables and a dependent variable.

OLS is a method that minimizes the responses predicted by the linear approximation and the sum of squared vertical distances between the observed responses in the dataset. The resulting estimator is expressed by a simple formula, especially in the case of a single regressor on the right-hand side. The OLS estimator is consistent when there is no multicollinearity, the regressors are exogenous and optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated.

OLS is based on the following assumption:

- Strict exogeneity : The errors in the regression should have conditional mean zero:
  $E[\epsilon|X] = 0$
- No linear dependence : The regressors in X must all be linearly independent.
  $Pr[rank(X)=p] = 1$
- Normality : Errors must have normal distribution conditional on the regressors.
  $\epsilon|X \sim N(0,\sigma^2)$
- Non autocorrelation: the errors are uncorrelated between observations:
  $E[\epsilon_i\epsilon_j|X] = 0$ for $i \neq j$

Chapter 4

DATA ANALYSIS

4.1 Introduction

One of the objectives of this study was to determine the factors that influenced digit preference in the South African census 2011 Age-Sex data. This chapter presents the results of this study in direct responses to stated objective. The results are organized into two sections. The first part, focuses on visual inspection of data which is illustrated by a line graph and a population pyramid. The second part, presents statistical findings for techniques that were applied to measure accuracy of data . The third part, presents the multivariate findings for factors that influences digit preference.

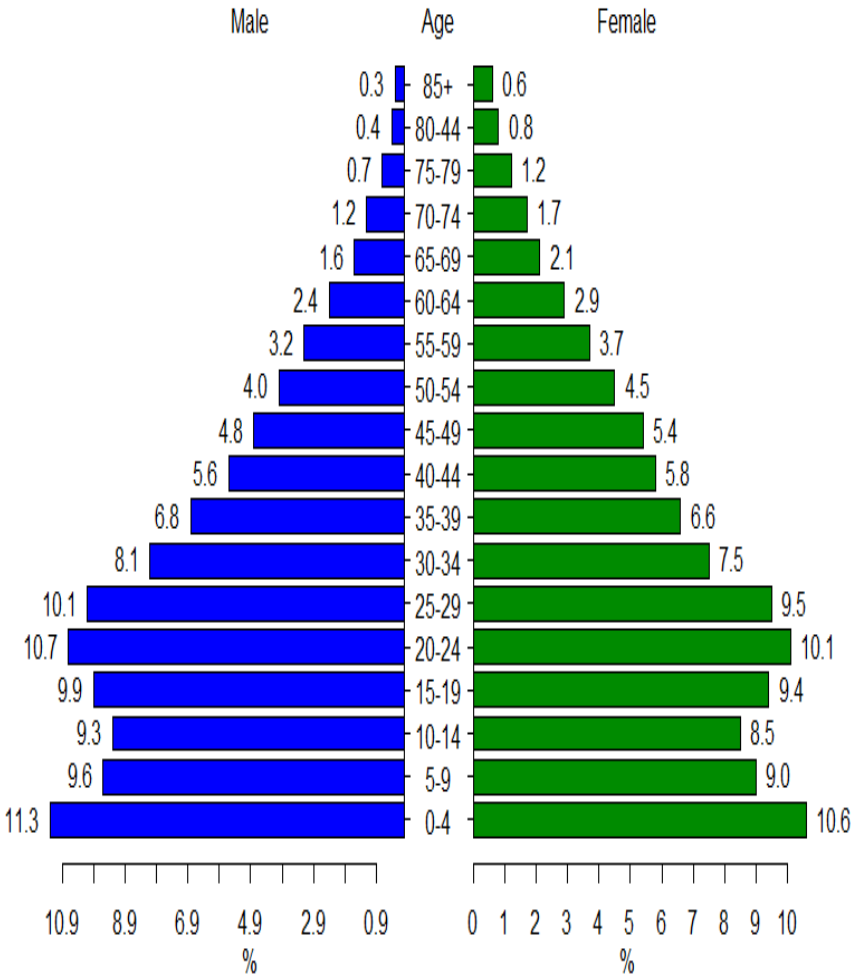
4.2 Source of Data

This study used a full dataset (Census 2011 Age-Sex data), which was obtained from Statistics South Africa, comprising information from all nine provinces of South Africa. The dataset was categorized according to education level, Population group, Place of residence and Sex. The total study population was 51,756,872, with 25,183,894 males and 26,572,978 females. Graphical, Statistical and Multivariate methods were employed in the analysis of the dataset.

4.3 Visual Inspections: Population Pyramid

The distribution of South Africa's population by sex is represented below on Figure 4.1. Out of a total of 51 756 872 people recorded, 25 183 894 were males and 26 572 978 were females.

Figure 4.1: Population Pyramid of South African Census 2011



The population distribution of South Africa depicts a youthful pattern with a broad base indicating high birth rate that signify large numbers of children and high fertility. For ages 25 and above, the number start to decrease, approaching 0.3% of the males population and 0.6% of females population. This might have been caused by mortality rate.

For the ages below 25 years, the changes with age are irregular, mostly due to variations in number of births over the last few decades and the impact of out-migration on the labour market.

The mortality rate is high amongst the males and females at advanced ages, especially at 60+, leading to decreased population by visual inspection. The population pyramid does not indicate severe age misreporting.

4.4 Statistical Results

Table 4.1: Age Specific Sex Ratio and Age Ratio of South African Census 2011

Age Group	Age Ratio		|Dev from 100|		Sex Ratio	Difference
	Male	Female	Male	Female		
0	-	-	-	-	101.6	-
1-4	-	-	-	-	101,8	0.2
5-9	93,1	94,5	6,9	5,5	101,3	0,5
10-14	95,2	91,9	4,8	8,1	104,2	2,9
15-19	99,2	101,6	0,8	1,6	99,7	4,4
20-24	106,9	106,7	6,9	6,7	100,6	0,8
25-29	107,5	107,7	7,5	7,7	101	0,5
30-34	95,8	93,2	4,2	6,8	102,2	1,1
35-39	99,4	99,4	0,6	0,6	97,2	5,0
40-44	96,5	97,2	3,5	2,8	90,7	6,5
45-49	99,1	103,5	0,9	3,5	83,9	6,8
50-54	100,7	100,2	0,7	0,2	83,8	0,1
55-59	100	99,5	0,0	0,5	82,4	1,4
60-64	100,9	100,3	0,9	0,3	79,2	3,2
65-69	88,7	90,6	11,3	9,4	72,2	7,0
70-74	-	-	-	-	64,5	7,7
75+	-	-	-	-	47.4	-
		Score	3.7	4.1		3.5

4.4.1 Age Sex Accuracy Index

Age-sex data plays an important role in the demographic analysis, where it's very important for individuals to report their ages accurately. In this study we initially checked the quality of data before further analysis, in order to ascertain whether the data could be used to make useful prediction. The result of the calculated index in Table 4.1 is interpreted as follows:

$$\begin{aligned}\text{Joint Score} &= ARS_m + ARS_f + 3\text{SRS} \\ &= 3.7 + 4.1 + 3(3.5) \\ &= 18.3\end{aligned}$$

The joint score of 18.3 reveals that the quality of the data was good. The joint score of 18.3 is below 20 of the UN recommendation, meaning that the data quality is very good and reliable, and can be used without any adjustment.

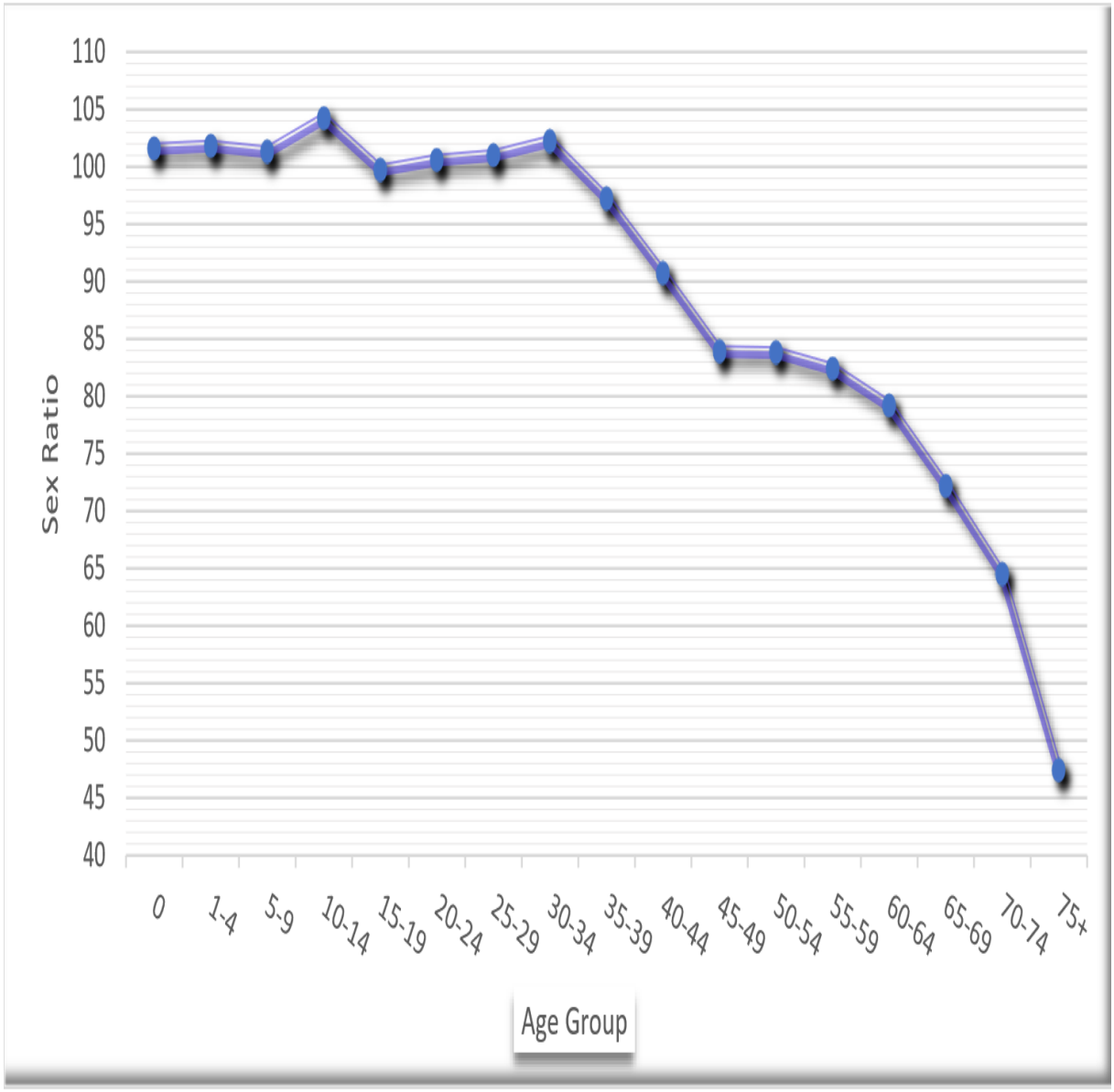
4.4.2 Sex Ratio

Table 4.1 above, presents calculated Sex ratio for different age groups (0 - 75+). The difference was calculated by subtracting Sex Ratio of a group from Sex ratio of the corresponding previous group. Sex ratio of 101.6 for age group 0 indicates an undesired number of males per 100 females in a young age group. This is because the sex ratio at birth is expected to range from 102 to 105 and the calculated sex ratio is falling out of the boundary. This shows less than expected males in age group 0, which might have been influenced by higher male infant mortality.

The pattern of sex ratio observed show a smooth pattern of gradual decline in sex ratio with age. The pattern of sex ratio with 5 year age groups is shown graphically in Figure 4.2 below.

The curve in Figure 4.2 below shows a smooth declining trend of sex ratio. The figure reveals an inversely proportional relationship. If Age increases, sex ratio decreases as well. The curve below cannot precisely tell misreporting of ages by individuals.

Figure 4.2: Sex Ratio of South Africa Census 2011



4.4.3 Age Ratio

Age ratio deviations from 100 indicate the sum of deviation (irrespective of sign) and the extent of misreporting in an age group which gives a measure of age misreporting or accuracy. An age ratio under or over 100 could imply that persons were misclassified to other age group, possibly that is adjacent to their actual age.

Table 4.1 presents calculated Age ratio for different age groups for both sexes. The female age structure appears to differ more than that of males. This assertion is confirmed by Age ratio score of 4.1 for females being higher than 3.7 age ratio for males observed. This might have been caused by plausibility of age misstatement among females than males.

For example, Age group 5-9 shows Age Ratio for both sexes below 100. This does not necessarily mean that young children of age group 0-4 were excluded when ages were reported, but they might have been a shift in that age group. In Age group 10-14, there might have been a likelihood of a shift in age reporting to 15-19 for females due to misstatements of age. The age ratio for 15-19 for the female suggests that teenage single girls are graded upwards in terms of their physical appearance, looking bigger than they are.

This leads to group 15-19 losing some proportion of its members to age group 20-29 resulting in upward shift of age, due to the physiological characteristics of married teenage girls. Van de Walle (1968) confirmed the observation about the characterisation of females which agrees with the patterns of age misreporting amongst women. He explained the phenomenon plausibly as a quasi-universal tendency to assume a higher typical age of marriage than actually prevails [58].

The pattern observed after the reproductive ages of women is virtually similar to that observed in the pattern for men at those same ages. As for men older than 50 years, it would appear like they preferred to be seen younger while on the other hand, some preferred to be seen older. These fluctuations result in peaks at age groups (50-54 and

60-64) and troughs at age group 50-59. The fluctuations are clearly shown in Figure 4.3.

The fluctuation in the sex ratio after the reproductive period (50+) could have resulted with older women losing memory and hence misreport their ages. Also, high parity women are assigned higher ages which account errors in age misreporting among females.

Figure 4.3: Age Ratio by Age and Sex

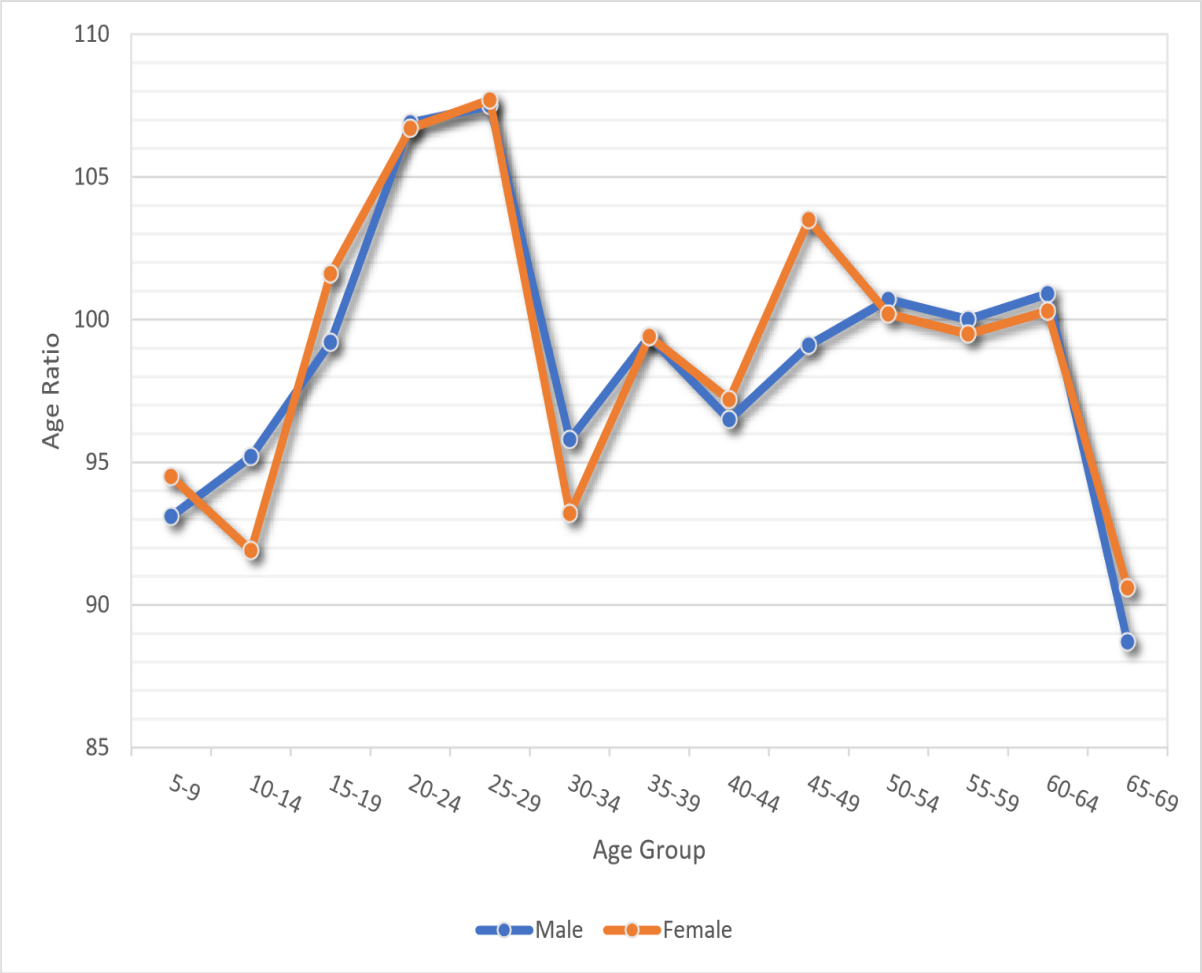


Figure 4.3: A graphical illustration of Age ratio deviations from 100 by Age group and sex. The Age ratios for each group are indicated by dots, which are then joined by straight lines resulting in peaks and troughs. This is an indication of marked age misreporting in particular groups.

4.5 Generalized Linear Model

This section presents the results of the investigation about the factors that influences digit preference in South Africa census 2011 Age-Sex data. The factors were sex, education level, place of residence and ethnic group. The relationship between an individual participant's digit preference and the factors indicated above, were examined using GLM.

4.5.1 Variable Analysis

The following table indicates variables used, with categories.

Table 4.2: Variables analysed with categories

Variables	Categories
Sex	Male, Female
Education Level	No schooling, Primary, Secondary, Higher
Place of Residence	Rural, Urban
Ethnic group	Black African, Coloured, Indian, White

4.5.2 Digit preference by sex

Table 4.3 below, presents the results for relationship between digit preference and gender. The test was conducted for both sexes to examine if there was a relationship between gender and preference of digits.

Table 4.3: Estimated coefficients, standard errors, two-tailed p-values and variables significant at the 5% for both sexes

Coefficients	Estimate	Std.Error	P-Value
Male	6.982e-05	3.815e-05	0.0702
Female	-2.208e-04	4.034e-05	3.4e-07***

Table 4.3 above shows an existence of relationship between female and digit preference with a p-value of (3.4e07***) < 0.05, showing a statistical significant. This means that female preferred certain terminal digits strongly when reporting their ages. Preference of those digits were over-selected by females compared to males. Hence the results above demonstrate a positive correlation between females and preference of terminal digits.

Furthermore, Table 4.3 above shows no existence of a relationship between Male and digit preference with a p-value of (0.0702) > 0.05. This shows no statistical significance. Hence, this shows that most of males ages were accurately stated compared to females. Males age reporting did not exhibit much of misreporting. Females played a major role in preference of digits which might have caused age heaping.

Table 4.4: Model Summary

Dependent Variable	Digit Preference
Independent Variables	Male, Female
Null deviance	85850.0 on 100 df
Residual deviance	3949.3 on 98 df
Akaike's Information Criterio(AIC)	664.91

Deviance is a measure of goodness or badness of fit statistic for a statistical model. GLM models do not present R^2, but R reports deviance in two forms (null deviance and residual deviance). The null deviance shows how the model performs, given the response variable as the predictor variable. This includes only the grand mean (Intercept).

Table 4.4 above, presents a null deviation value of 85850 on 100 df, which include independent variables (Male and Female). The deviance is decreased to 3949.3 on 98 df, this is a significant reduction. The Residual Deviance has reduced by 81900.7 = (85850.0 - 3949.3) with a loss of 2 df =(100 - 98) df. A high value of null deviance (85850) shows how bad the model is. This model does not fit well. The response variable is not well predicted by the model.

4.5.3 Digit preference by education level

Education is one of the other factors that this study proposed to investigate if it played a role in digit preference. Table 4.5 below presents the results of education level versus digit preference.

Table 4.5: Estimated coefficients, standard errors, two-tailed p-values and variables significant at the 5% for education level

Coefficients	Estimate	Std.Error	P-Value
No Schooling	2.273e-04	3.682e-05	2e-16 ***
Primary	-1.289e-04	9.828e-06	3.18e-07 ***
Secondary	-4.152e-05	2.111e-06	0.0985
Higher	-3.893e-05	7.177e-06	0.7113

Table 4.5 above shows an existence of relationship between education level and digit preference. There is a relationship between No schooling and digit preference, same applies to primary level. This relationship is attested by a p-value of (2e-16***) < 0.05 for primary level, for No schooling and a p-value of (3.18e-07***) < 0.05. These results show a statistical significance pattern.

Therefore, these results show that respondents who did not go to school had a higher propensity for terminal digit preference. Those who went to primary level also preferred certain terminal digits when stating their ages. The significant value of (3.18e-07***) for primary level of education is less compared to no schooling of (2e-16 ***).

The same results, Table 4.5, show that there was no relationship between preference of digits and the other two variables (secondary and Higher respectively). This is confirmed by a p-value of (0.0985) > 0.05 and a p-value of (0.7113) > 0.05, respectively. These all show that there were not statistical significant for these two variables.

This means that respondents who went to Secondary did not prefer any terminal digits. They reported their ages accurately probably because they had knowledge of their exact ages. And automatically those who attended Tertiary education. They reported their ages accurately. However, there might also be a small portion of participants who might have deliberately misreported their ages regardless of their level of education.

Table 4.6: Model Summary

Dependent Variable	Digit Preference
Independent Variables	No schooling, Primary, Secondary, Higher
Null deviation	2569 on 56 df
Residual deviation	121.6 on 3 df
Akaike's Information Criterio(AIC)	241.6

Table 4.6 presents a null deviation value of 2569 on 56 df which include independent variables (No schooling, Primary, Secondary, Higher). The deviance is decreased to 121.6 on 3 df, this is a highly significant reduction. The Residual Deviance has reduced by 2447.4 with a loss of 3 df. This model fit well the data points. The response variable is well predicted by the model.

4.5.4 Digit preference by rural-urban pattern

Table 4.7 below presents the results of digit preference by place of residence.

Table 4.7: Estimated coefficients, standard errors, two-tailed p-values and variables significant at the 5% for rural-urban pattern

Coefficients	Estimate	Std.Error	P-Value
Rural	-1.049e-04	8.150e-06	2e-16 ***
Urban	-3.420e-05	9.010e-06	0.000255 ***

The results on Table 4.7 above, shows that place of residence played a role in preference of digits by individuals when reporting their ages. The table above shows that there is a relationship between rural dwellers and digit preference with a p-value of (2e-16***) $<$ 0.05, the p-value shows a statistical significant pattern. Rural dwellers misreported their ages, they preferred certain terminal digits when reporting their ages. This might have been influenced by lack of education.

It is evident that Urban dwellers misreported their ages also. This is confirmed by a p-value of (0.000255***) $<$ 0.05, showing a relationship between urban and digit preference. Fewer Urban dwellers misreported their ages compared to rural dwellers. Further, the correlation for rural dwellers is more compared to Urban dwellers, which shows that a misreporting occurred in Rural dwellers.

Table 4.8: Model Summary

Dependent Variable	Digit Preference
Independent Variables	Rural, Urban
Null deviation	85850 on 100 df
Residual deviation	3817.7 on 98 df
Akaike's Information Criterio(AIC)	661.49

Table 4.8 presents a null deviation value of 85850 on 100 df which include independent variables (Rural and Urban). The deviance is decreased to 3817.7 on 98 df, this is a significant reduction. The Residual Deviance has reduced by 82032.3 with a loss of 2 df.

A high value of null deviance (85850) shows how bad the model is. This pattern is similar to the results of the model in Table 4.3. This model does not fit well. The response variable is not well predicted by the model.

4.5.5 Digit preference by ethnic group

Table 4.9 below presents the results for estimated coefficients, standard errors and p-values for different ethnic group.

Table 4.9: Estimated coefficients, standard errors, two-tailed p-values and variables significant at the 5% by population group

Coefficients	Estimate	Std.Error	P-Value
Black.African	-1.530e-06	8.372e-06	9.12e-11 ***
Coloured	-5.872e-04	7.224e-05	6.24e-08 ***
Indian	-1.026e-03	3.105e-04	0.00175 **
White	6.295e-04	1.443e-04	0.85569

Table 4.9 above shows that there is a relationship between population group and digit preference for three population groups (Black, Coloured and Indian). This is confirmed by their p-values < 0.05, which is statistical significant. This shows a relationship between population group and digit preference. As for Whites, these results shows that there was no relationship between digit preference and white population. This is confirmed by a p-value $(0.85569) > 0.05$. This means that Whites declare their ages accurately compared to other population group.

The Black population was the worst group which misreported their ages with a p-value (9.12e-11***). This might have attributed to by low level of education. Most of black respondents interviewed, might not have gone to school, which negatively impacts their responses. The Coloureds group was the second worst group that declared their ages inaccurately. This is confirmed by a p-value (6.24e-08***). The Indians misreported their ages better, but this was not much compared to the other two groups (Black and Coloured).

Table 4.10: Model Summary

Dependent Variable	Digit Preference
Independent Variables	Black African, Coloured, Indian, White
Null deviation	2650.0 on 55 df
Residual deviation	146.5 on 51
Akaike's Information Criterio(AIC)	250.94

Table 4.10 above, presents a null deviation value of 2650.0 on 55 df which include independent variables (Black African, Coloured, Indian, White). The deviance is decreased to 146.5 on 51 df, this is a significant reduction. The Residual Deviance has reduced by 2503.5 with a loss of 4 df. This model fit well. The response variable is well predicted by the model.

4.5.6 Model Diagnostics

Table 4.11 presents the results of goodness of fit test. The Goodness of fit test was used as a key principle, to test how a set of observations are well fitted by a statistical model. Table 4.11 presents the overall p-values, AIC variables evaluated accordingly. Deviation is divide into null and residual according to how its presented in R. AIC was used as decision criteria, to conclude which model best fit the data. The Rank of order is used to show which model best fit according to its AIC.

Table 4.11: Overall Goodness of fit

Model	Variables	Deviation		P-Values	AIC	Rank of Order
		Null deviation	Residual deviation			
1	Sex	85850.0	3949.3	2e-16 ***	664.91	4
2	Education level	2569.0	121.6	2e-16 ***	241.6	1
3	Place of residence	85850.0	3817.7	2e-16 ***	661.49	3
4	Ethnic Group	2650.0	146.5	2e-16 ***	250.94	2

It is sufficient to conclude that the model 2, fitted using Education level as reference , is the best model that measures how well the set of observations in a data is presented. This model has an AIC of 241.6, with a null deviation 2569. This model has the least null deviation showing how widely the model is useful to predict the response variable. This model best fit the data compared to the other three consecutive models with their reference (Sex, Place of residence and Ethnic Group).

The second best model is, model 4 with reference variable (Ethnic Group). This model has an AIC of 250.94, which is greater than the other two models with reference (Sex and Place of residence). The null deviation also gives the same results as the AIC, the model has the good null deviation (2650).

The least models that do not fit the set of observations properly, are models (3 and 1, respectively). These models perform least of prediction of the response variable. They do not fit well, and shows the unfitness of the models.

GLM models does not calculate R^2, but variation can be calculated using the reported deviation. The variation is calculated as follows:

$$\frac{\text{Null deviation - Residual deviation}}{\text{Null deviation}} \text{ x } 100$$

$$\Rightarrow \frac{2569 - 121.6}{2569} \text{ x } 100$$

$$\Rightarrow 95.26\%$$

$\therefore$ 95.26% explains all of the variation in the response variable around its mean.

4.6 Principal Component Analysis

This section presents result of PCA. Factors that influence digit preference were examined using PCA.

4.6.1 Appropriateness of PCA

In order for PCA to function appropriately some measures have to be met:

- PCA must have some correlations greater than 0.3 between the variables included in the analysis
- PCA carries out some test to assess whether the correlations are sufficient to proceed with the analysis. The Kaiser-Meyer-Olkin and Bartlett's test are carried out initially before any further analysis. The decision criteria of these tests are explained in detail in Chapter 3.

4.6.1.1 Correlation Matrix

Table 4.12: Correlation Matrix for 12 Variables

	No.school	Primary	Secondary	Tertiary	Male	Female	Black.African	Colour
No.school	1.00	0.24	-0.90	-0.40	-0.86	-0.84	-0.86	-0.77
Primary	0.24	1.00	-0.06	0.04	0.02	0.07	0.01	0.25
Secondary	-0.90	-0.06	1.00	0.53	0.99	0.98	0.99	0.93
Tertiary	-0.40	0.04	0.53	1.00	0.62	0.62	0.60	0.61
Male	-0.86	0.02	0.99	0.62	1.00	0.06	0.56	0.95
Female	-0.84	0.07	0.98	0.80	0.82	1.00	0.08	0.97
Black.African	-0.86	0.01	0.99	0.60	0.06	0.80	1.00	0.95
Coloured	-0.77	0.25	0.93	0.61	0.95	0.97	0.95	1.00
Indian	-0.71	0.17	0.86	0.84	0.92	0.92	0.90	0.93
White	-0.20	0.53	0.41	0.75	0.51	0.54	0.48	0.68
Rural	-0.87	0.05	0.99	0.44	0.98	0.98	0.98	0.93
Urban	-0.79	0.04	0.94	0.77	0.98	0.97	0.97	0.94

The correlation matrix on Table 4.12 below, satisfies the 1st requirement for further computation of PCA. For this set of variables, the values of correlation of some are greater than 0.3. This requirement is achieved. Some of them are marked with red.

4.6.1.2 KMO and Bartlett's Test

Table 4.13 below, presents the results of calculated measure of sampling adequacy. If the computed KMO value is greater than 0.5, this satifies the requirement to proceed with PCA.

Table 4.13: Test for Sampling Adequacy

KMO Test	0.687
Bartlett's Test	323.671
Df	26
P-value	0.0000123

From Table 4.13 above, the KMO for each individual variable included in the analysis is greater than 0.5. This means that further analysis of PCA can be carried.

Bartlett's test, tests the correlation matrix among variables. It makes conclusion based of 1% level of significance. Variables should be correlated to meet requirement of the test. Bartlet test is demonstrated as below :

H_0: Variables are not correlated
H_1 : Variables are correlated

Test statistic : Calculated p-value (0.0000123)

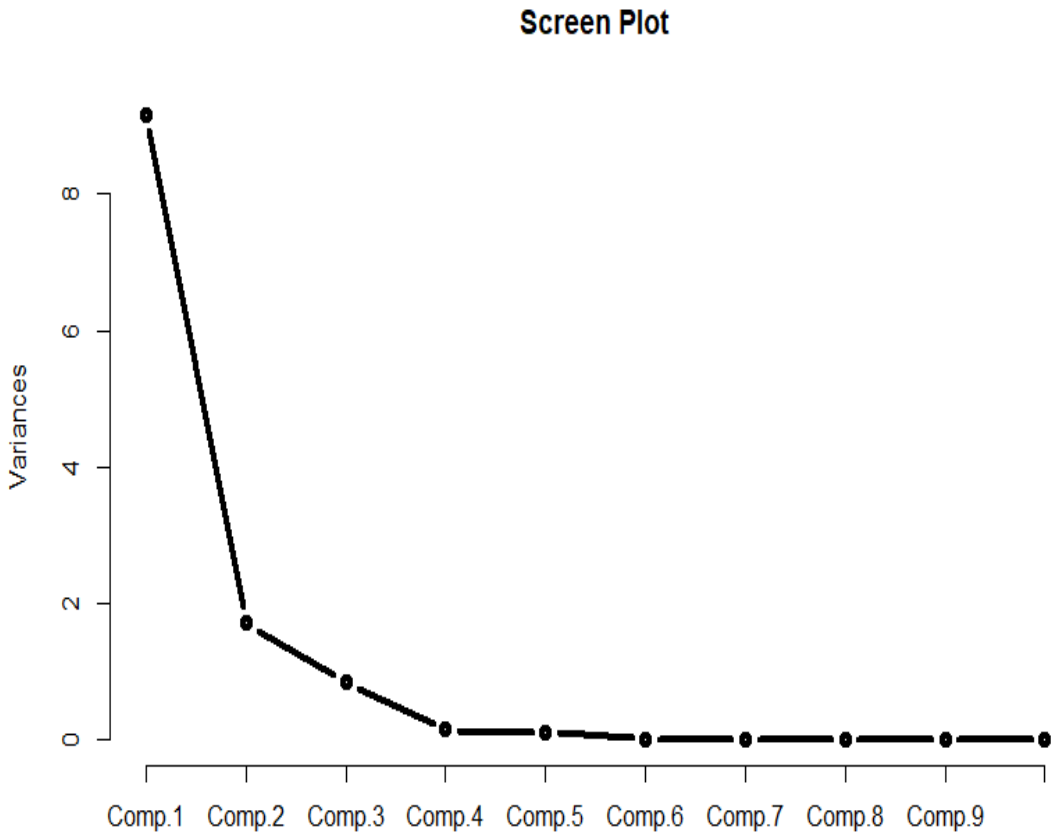
Decision : Reject H_0
Reason : P-value(0.0000123) $< \alpha$, α=1%
Conclusion : Variables are correlated

The KMO and Bartlett's Test are sufficient enough to proceed with PCA. They both meet the minimum requirement of PCA to be carried out.

4.6.2 Further PCA Computation

Figure 4.4 below presents a graphical demonstration of computed components for PCA.

Figure 4.4: Scree Plot



A scree plot is a graphical plot of component numbers and variances. Scree plot is used to find out how many components of variables should be retained. The number of factors that should be generated by the analysis is extracted by the point where the plot begins to straighten out (the elbow). From Figure 4.4, only three components are extracted. Therefore, it takes 3 components to explain most of the variance in 12 variables analysed.

Table 4.14: Rotated Component Matrix

Variable	Components		
	1	2	3
No.School	**0.945**	0.414	0.432
Primary	**0.928**	-0.312	0.316
Female	0.589	**0.968**	0.213
Black.African	**0.911**	0.432	0.478
Coloured	**0.865**	0.278	-0.342
Indian	0.341	0.387	**0.812**
Rural	**0.834**	0.429	0.425
Urban	**0.798**	0.567	0.456

PCA was carried out on 12 predictors of digit preference. However, only 8 predictor variables were chosen for the PCA. This was because, they were free of complex structure. Some factors had high loadings on more than one component which made it complex to derive analysis on such factors.

The components shown on Table 4.14, represent the partial correlations between a variable and a component. It ranges from -1 and 1. The predictors of digit preference is explained by three underlying principal components. The First component is usually regarded as the best explanation of correlation among the predictor.

Table 4.14 above, shows that education levels (No schooling and primary), Ethnic groups (Black, Coloured and Indian) and Place of residences (Rural and Urban) are highly correlated with the first principal component and vary simultaneous. This means that preference of digits is best explained by education level, ethnic group and place of residence.

Individuals who did not go to school and those who went to primary, have a likelihood of misreporting their ages. Same applies to individuals who stay in rural or urban areas regardless of their ethnicity. They misreport their ages in the same manner, resulting in them, showing certain traits of preference of terminal digits when stating their ages.

Table 4.14, furthermore reveals that Sex (Female) is highly correlated with the third principal component. This means that one's sex plays a role in determining factors that influences digit preference, but at a lesser extent.

This result resembles the same findings by GLM in the previous section.

4.7 Regression Analysis

This study used simple linear regression for analysis of determining factors that influences digit preference.

4.7.1 Regression coefficient

Table 4.15 below, presents results of calculated regression coefficients.

Table 4.15: Results for Regression Coefficients

Variable	Unstandardized Coefficients		Standardized Coefficients	t-value	Pr(>|t|)
	Estimate	Std Error	Estimate		
Intercept	-21.895	45.888		56.896	0.567
No Schooling	-13.678	-6.789	-8.431	-2.678	0.000*
Primary	-12.763	-8.112	-10.716	-4.890	0.000*
Secondary	1.987	6.360	2.181	1.116	0.089
Tertiary	4.679	2.181	0.798	1.234	0.123
Male	0.456	0.169	0.988	1.369	0.890
Female	-1.578	-2.810	-1.732	-1.689	0.000*
Black. African	-9.120	-5.169	-12.626	-6.213	0.000*
Coloured	-6.954	-6.181	-4.812	-6.333	0.000*
Indian	-5.892	-6.161	-0.862	-6.214	0.000*
White	-1.567	-4.170	-0.799	-6.213	0.000*
Rural	-5.895	-3.181	-3.130	-3.189	0.000*
Urban	-1.743	-2.765	-3.231	-3.761	0.000*

Estimated Linear Model : Unstandardized

$Y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$

Digit Preference = -21.895 - 13.678(No Schooling) - 12.763(Primary) + . . . +(-1.743)(Urban)

Education Level :

The negative sign for variables (No school and Primary) shows that there is an inverse relationship between Digit preference and those two variables. However, No Schooling and Primary are significant at all levels of significance (i.e 1%, 5% and 10%). There is a relationship between preference of digits and those variables. This variables influence digit preference. Population of those who did not go to school decreases the likelihood of declaring age accurately by 13.7 points. And the population of those who went to primary, decreases the chances of declaring accurate age by 12.8 points. This means that individuals were preferring certain digits when reporting their ages. From Table 4.15 above, we can deduce that those who did not go to school at all provided more wrong information about their ages compared to those that attended primary school .

The positive sign for variables (Secondary and Tertiary) shows no relationship between Digits preference and those variables. They are insignificant at all possible levels of significance (i.e 1%, 5% and 10%). The insignificance between those variables, shows that there is no relationship between digit preference and those digits. These two variables increase the likelihood of declaring accurate age. Individuals who went to secondary increase the percentage by 1.987 points, and those from Tertiary by 4.679 points. These figures show that those who went to Tertiary level declare their ages more accurately compared to those from Secondary schools.

These results show that Education level is crucial in reporting accurately ages related to data. In that way Education level is a factor that influences digit preference. Thus, this result agrees with those given by GLM and PCA.

Sex :

Digit preference varies with Sex. The results show that Sex of an individual matters when it comes to age reporting and this is a factor which influences digit preference as females and males respond differently when stating their ages. There exists an inverse relationship between females and digit preference. The relationship is observed with significance (000*). Females prefer certain terminal digit when reporting their ages. Thus,

Females decrease the percentage of reporting age accurately by 1.58 points.

In male population, there exists no relationship between digit preference. However, variable Male is insignificant at all levels of significance (i.e 1%, 5% and 10%). The significance value is 0.890 which is greater than (1%, 5% and 10%). This means that male reported their ages accurately compared to females, which might mean they had more knowledge about their ages. They increase a percentage of reporting accurately their ages by 0.46 point.

Ethnic Group :
The negative sign for variables (Black, Coloured, Indian and White) shows that there is an inverse relationship between Digit preference and those four variables. However those four variables are significant at all levels of significance (1%, 5% and 10%). Those variables influence preference of digits. This means that Black, Coloured, Indian and White population prefer certain terminal digit when reporting their ages. Their ages were not accurately reported.

Place of residence The results on Table 4.15, show that place of residence plays a role in factors that influences digit preference. The negative sign for variables (Rural and Urban), shows that there is an inverse relationship between digit preference and those variables. They are significant at all levels of significance (1%, 5% and 10%). Rural and Urban dwellers misreported their ages. They had a likelihood of preferring a certain digit when reporting their ages.

Overall, the results of Table 4.15 show the same findings as GLM and PCA. This is sufficient to predict that Education level, Sex, Ethnic group and Place of residence, are the factors that influences digit preference.

4.7.2 Model Summary

Table 4.16: Model Summary

Dependent Variable	Digit preference
Independent Variables	12
R^2 0.953	
Adjusted R^2	0.948

R^2, the coefficient of determination is the proportion of the variance in the dependent variable that is predictable from the independent variable. It ranges from 0 to 1; $0 \leq \mathbf{R} \leq 1$. The R^2 is positively biased, however an Adjusted R^2 was calculated in order to correct the biasness. The Adjusted R^2 for this multiple regression is 0.948. This means that 94.8% explains the variation in the dependent variable and only 5.2% is explained by the error term. Hence the model constructed for predicting factors that influences digit preference is good.

4.8 Summary of Results

This chapter presented the results of analysis carried out on South African Census 2011 Age-Sex data. The study initially checked the quality of data before making any analysis. The UN Joint score gave a value of 18.3, which shows that the data collected was of good quality. It was easy to commence with analysis, since there was no manipulation or adjustment to be carried out.

Henceforth, application of multivariate methods were applied to the data. Amongst the multivariate methods, were GLM, PCA and Regression analysis. GLM was used to examine and determine factors that influence digit preference. PCA and Regression analysis were introduced to check if they would yield the same outcome as GLM, and also used as decision criteria to make conclusion that certain factors influence digit preference, if they would give the same results.

From the results, GLM found a linear relationship between digit preference and variables (Education level, Ethnic group, Place of residence and Sex). P-values calculated was found to be less than 0.05. This showed that those variables are factors that influence digit preference. PCA also found the same results as GLM. PCA was carried out on 12 predictors of digit preference, 8 predictors variables were chosen. Those 8 predictors explained the correlation between digit preference and variables (Education level, Ethnic group, Place of residence and Sex). The correlation showed that those variables are factors that influences digit preference.

Regression analysis as well revealed the same results as the other two (GLM and PCA). It was sufficient to conclude that Education level, Ethnic group, Place of residence and Sex are factors that influence digit preference, because all of the methods gave almost the same outcome.

Chapter 5

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This study sought to assess and evaluate the reliability of data collected in South Africa about Age-Sex 2011 census, through the investigation of possible factors responsible for influencing digit preference. Several data evaluation techniques were used to address the specific objectives of this study. Among the techniques employed, were visual inspection, statistical and multivariate methods.

This chapter serves to discuss and interpret in detail, the findings obtained from the analysis. This chapter comprises of three sections. The first section gives a detailed discussion of the results obtained. The Second section gives a conclusion of this work based on the findings as presented in the results. The last section presents recommendation about what is required to address problems arising from the study. This could assist in policy formulation by various stakeholders.

5.2 Discussion of the Results

5.2.1 Quality of data

This study found out that the data collected was of good quality. The UN Joint Score method was applied to access the quality of data. The UN joint Score calculated was 18.3, which is less than 20. This result means that the age-sex data collected was accurate and is usable without any adjustments.

Most age data collected in different countries seem not to be of better quality than of this study. For example, Obisesan and Mojoyinola (2016) reveal the poor quality of data collected in Botswana. In their study, they found a United Nation Age-Sex Accuracy Index score of 68.29, which signifies a highly unreliable data and so risky to utilize such data [40].

Another study by Bwalya et al. (2015) reveals the poor quality of data collected in Zambia. In this study, authors found that age heaping in ages ending with the digit '0' and '5' exist in Zambia, with fewer females than males, having their age heaped at those two terminal digits [10]. However, besides a significant remarkable improvement from a score of 72.9 in 1969 to a score of 27.9 in 2010 as indicated by the United Nation Age-Sex Accuracy Index, age data still remained inaccurate due to a strong tendency of preferring those digits by individuals.

Another study by Bainame and Letamo (2015) showed the same pattern of poor quality of data. In their study, they found a United Nation Accuracy index of 21.0 for 2011 population census age data of Botswana [39]. That unreliable data had to be used to make certain projections, but only after some adjustment on it.

Furthermore, Bello (2012) evaluated the accuracy of age reporting by the outpatients in General Hospital Dutsin-ma, Katsina state, Nigeria using demographic techniques, namely Whipple's Index, Myer's Blended Index and UN Age-sex Accuracy Index. In his study, the age data observed was of bad quality, for both male and female outpatients [6].

The calculated Age-sex Accuracy Index was 156.8 that revealing highly inaccurate age data according to the United Nations scaling [6].

5.2.2 Assessment of Age ratio and Sex Ratio

This study analyzed the sex ratio for the population of South Africa Census 2011. From Table 4.1, we can see a low value of sex ratio at birth (101.6). Some literature reviewed that sex ratio at birth is usually around 104 because of the biological fact that male birth generally exceeds female birth [42]. This study found a value less than 104, which is abnormal. This might have been caused by high infant mortality rate or the decrease in fertility, or the under-enumeration of births.

In Central India, the ratio of males to females in different age groups are affected by the sex-selective migration of the working age population [49]. This pattern of the working age population is observed in Table 4.1, with sex ratio likely less than 100. The same results also show a shortage of males. This should have resulted from under reporting of males in those age groups.

In a normal population, Sex ratio decreases with age; males at every age have higher mortality rates than females [50]. Errors in age and sex data are commonly caused by a shift in men into older ages while women tend to understate their ages which pushes them into younger ages. In Table 4.1, we can see a pattern of under enumeration for sex ratios at ages below 10.

Sex ratios for populations of people aged at 60 years and above is expected to be low. Poston and Davis (2009) found that a decline in sex ratio levels around ages of 50 to 70 is as a result of low female mortality rate as compared to male mortality rate. A decline in sex ratio observed in Table 4.1 might have been a result of more males dying compared to females.

The results in Table 4.1 show that there was not much of major deviation from 100 since most of the Age Ratio values were so close to 100. This, as a result, shows no major changes in fertility, mortality or migration. Thus, the results show under enumeration at ages below 15 for males and below 10 for females, with an Age ratio less than 100. This indicate an under selection or errors in age reported. Not only was ages below 15 years under enumerated, but also age (30-49) for both sex were under enumerated.

The Age ratio of more than 100, suggests the opposite of under selection or error in age reported. This means that the Age ratios of more than 100 in Table 4.1 suggest an over selection of those ages. Age groups (25-29, 45-49) respectively, show a significant concentration of those ages by Females, with the highest age ratio value greater than 100 and age group 25-29 showing low concentration by males. This significant concentration is observed clearly in Figure 4.3 with fluctuations in the plot.

5.3 Assessment of Factors that influence digit preference

Educational level plays a major role in age declaration by individuals when reporting their ages. Most of the studies have revealed that age misstatement is mostly influenced by education level of the respondent. For example, a study by Agrawal1 and Khanduja (2015) revealed that literacy played an important role when indivuals were reporting their ages. In their study of "Influence of Literacy on India's Tendency for Age Misreporting", they found that illiterate people reported their ages inaccurately compared to literate people [1]. They concluded that "increased literacy will further improve the quality of age data in states and areas still lagging behind" [1].

Another study by Pal et al. (2015) revealed the same results as [1]. In their study of "Age Reporting Error and Effect of Education", they found that education level has impact in reporting of ages [43]. Those who completed primary level are slightly less prone to misreporting their ages since they are well informed with numbers compared to those who did not go to school at all [43].

This study found out that education level is directly related to digit preference in age reporting. Individuals who did not go to school, misreported their ages in large numbers and had a preference for certain digits, which turned out to be different from those preferred by those who atleast attended primary schools. This is confirmed by GLM, PCA and Regression analysis results.

Those who completed secondary level and enrolled for further education, did not show any pattern of preferring any digit when reporting their ages. Age misstatement of those who did not go to school might also have been influenced by illiterate elderly group. Dechter and Preston (1991) concluded that misreporting is most severe at old age for those that are illiterate [16]. This might have contributed to observed results of this study.

In this study, it was found that misreporting of age is severe at older ages (50+), with an age ratio above 100, as shown in Table 4.1. The observation is similar to the finding by Dechter and Preston (1991). A study by Dechter and Preston (1991), found that misreporting is more severe at an older age with evidence of a very pervasive overstatement of ages at an advanced ages [16]. However, their findings clearly show the low educational level and literacy is major agents of age misstatement.

Another study by Tekpli (2013) found that education level plays a major role in reporting of ages [54]. In his study, a relationship between digit preference and literacy existed in census of Ghana 2000 and 2010. He found that non-literate heads of household misreported more than their literate heads of household [54], which might have been also a case to this study.

The findings of this study are similar to what was mostly found by other studies like (Agrawal & Khanduja, 2015; Dechter & Preston, 1991; Pal et al. , 2015; Tekpli, 2013) [1, 16, 43, 54]. This study found that education level plays a major role in preference of digits, which result in individuals stating their ages inaccurately. This is evident as shown by computed P-values in Table 4.5.

The results in Table 4.14 of PCA and Table 4.15 of Regression analysis, revealed the same results as GLM. This agreement is evidently sufficient to lead us to conclude that education is a significant factor that influences digit preference. Individuals who did not go to school and those who went to primary school, prefer certain digits differently from each other when reporting their ages.

Many studies have revealed that place of residence influences preference of digits when individuals are reporting their ages. For example, a study by Agrawal1 and Khanduja (2015) found out that rural dwellers misreported their ages in large numbers compared to urban dwellers [1]. Henceforth, they found that urban people were more likely to give better quality age data compared to rural people, with a correlation coefficient of -0.90, which shows that there was no bias towards digit preference amongst urban dwellers when reporting [1]. They reported their ages accurately.

The results of this study are also in agreement with those of other studies like [1, 7, 62]. Preference of digits in rural places might have been influenced by lack of education amongst rural folks. They might have just reported their ages recklessly without any regard of the consequences. The rural people face a lot difficulities in their lives, ranging from lack of propers schools, clinics and other facilities rendering them susceptible to poverty and poor education .

Yazdanparast et al. (2012) found that age reporting is good and more accurate in urban than rural populations with index (111.58) for the whole population, which showed that the data was of good quality, but rural dwellers preferred certain digits when reporting their ages [62]. Borkotoky and Unisa (2014) revealed the same pattern as [62]. In their

study, they pointed out that digit preference has a relationship with place of residence [7]. They found out that rural people were reporting their ages inaccurately in large numbers and they were preferring certain digits [7].

Not only was education level and place of residence the factors that influences digit preference this study found. This study found that ethnic group also influences preference of digits. This was shown in the results of GLM in Table 4.5. The results show that Black, Coloured and Indian respondent had a likelihood of preferring certain terminal digit when reporting their ages. As a results, their ages where not reported accurately. The findings of GLM, was confirmed by PCA and Regression analysis. Based on agreement between PCA, Regression analysis and GLM results, it is sufficient to conclude that ethnic group also influences digit preference.

5.4 Conclusion

Theoretically speaking, collection of information on age should be a very simple and easy task and yet errors inevitably occur in the collection of census data. It is essential to detect and quantify those errors, so that the users are aware of the quality of the data.

Based on the data from Census 2011 of South Africa, the accuracy of the quality of age-sex data was examined using the proposed method (Age-sex Accuracy Index), which was discussed in previous chapters. The study furthermore examined the specific objectives proposed, by the use of Visual Inspection methods (Line graph and Population Pyramid), statistical methods (Age Ratio and Sex Ratio) and multivariate methods (GLM, PCA and Regression analysis) which were also discussed in detail in previous chapters.

This study found out that the quality of data collected, was good, reliable and can be used for analysis without any adjustment. The calculated UN Joint Score was 18.3, shows that the data was of good quality.

The calculated UN Joint Score of this study based on 2011 South Africa Census is significantly different from other studies with regard to the evaluation of quality of data. This study compared findings of different countries such as Botswana, Nigeria and Zambia. All of them, provided a poor quality of data collected. Thus, this makes this study unique.

This study did not only dwell on the quality of data, but it also examined possible factors that influence the preference of certain digits. This study found that the education level, population group, place of residence and sex are all factors that influence digit preference.

Educational level was one of the factors that this study proposed to examine. It was found that education level has a relationship with preference of digits. This was shown in calculated p-value (2e-16***) < 0.05 for those who did not go to school. This result showed that illiterate respondents were preferring certain digits when declaring their ages, which causes age heaping. Those who went to school were unlikely to misreport their ages since they are knowledgeable enough about their ages.

Education level played a major role in other factors such as sex, rural-urban and ethnic group. All of those other factors showed that digit preference has a relationship with them. Rural urban factor showed that rural dwellers were preferring digits when stating their ages which might have been influenced by education level at that area. Ethnic group reveled that Black population preferred digits when reporting their ages, which as well might have been attributed by black respondent who did not go to school.

Overall, this study is sufficient to conclude that education level, population group, place of residence and sex, influence digit preference and accept the hypothesis of this study. This is served with evidence of calculated results of GLM, PCA and Regreesion analysis, figures in the study, and referring to what other studies prevails.

5.5 Recommendation

In spite of the progress that have been made to ensure that age sex data is collected accurately, there is still much room for improvement. There are urgent policy issues that need to be addressed, to try to minimize errors that might be encountered. The following are some of the recommendations this study propose. This can be monitored, implemented and evaluated through any relevant policy interventions:

1. Preference of certain terminal digits its a major concern of quality of data. This can be utilized by creating a workshop for digit preference. Statistics South Africa should take into account, organizing a workshop for their field staff on digit preference and how it impacts data quality. Henceforth, they can do a pilot study to check validity and reliability of the instruments through pre-testing interview questions.

2. Not only should Statistics South Africa organise a workshop for field staff, they should advocate the mass awareness about the importance of accurate ages in our daily life. This will limit misreporting of ages, since respondents will be well informed why its important for the them report their ages accurately.

3. Statistics South Africa should partner with IEC South Africa to get the exact respondent age, since the IEC South Africa have the record of the voters ID. This will minimize errors that might be encountered while recording age.

4. Enumerators should be trained a recall technique in obtaining exact ages (e.g. the use of life history and community calendars) which relates the birth of an individual to some notable event. This will assist elderly people who do not recall their exact ages.

Bibliography

[1] Agrawal, G., and Khanduja, P. Influence of literacy on india's tendency for age misreporting: Evidence from census 2011. *Journal of Population and Social Studies 23*, 1 (2015), 47–56.

[2] Akers, D., and Larmon, E. A. Indians and smudges on census schedule. In *Journal of the American Statistical Association* (1968), vol. 63, pp. 745–745.

[3] Ansari, M. Economic growth, rural poverty and human resource in bhutan: a description of the linkages in the eastern himalayas. *SAARC J Hum Resour Dev vol 4*, 1 (2008), pg 87.

[4] Bailey, M., and Makannah, T. J. An evaluation of age and sex data of the population censuses of sierra leone: 1963-1985. *Genus vol 52*, 1/2 (1996), pg 191–199.

[5] Bairagi, R., Aziz, K., Chowdhury, M., and Edmonston, B. Age misstatement for young children in rural bangladesh. *Demography 19*, 4 (1982), pg 447–458.

[6] Bello, Y. Error detection in outpatients' age data using demographic techniques. *International Journal of Pure and Applied Sciences and Technology 10*, 1 (2012), 27.

[7] Borkotoky, K., and Unisa, S. Indicators to examine quality of large scale survey data: an example through district level household and facility survey. *PLoS One 9*, 3 (2014), e90113.

[8] Botha, I. A comparative analysis of the synchronisation of business cycles for developed and developing economies with the world business cycle. *South African Journal of Economics 78*, 2 (2010), 192–207.

[9] Brooks, C. *Introductory econometrics for finance*. Cambridge university press, 2019.

[10] Bwalya, B. B., Phiri, M., and Mwansa, C. Digit preference and its implications on population projections in zambia: Evidence from the census data.

[11] Coale, A., and Li, S. The effect of age misreporting in china on the calculation of mortality rates at very high ages. *Demography 28*, 2 (1991), 293–301.

[12] Coale, A. J., and Demeny, P. G. *Regional model life tables and stable populations*. Princeton University Press, 1966.

[13] Coale, A. J., and Stephan, F. F. The case of the indians and the teen-age widows. *Journal of the American Statistical Association 57*, 298 (1962), 338–347.

[14] Crawford, S., Johannes, C., Bradsher, J., Stellato, R., Sherman, S., and Samuels, S. Digit preference in year at menopause: Data from the study of women's health across the nation. *Annals of epidemiology vol 10*, 7 (2000), pg 457.

[15] Dahiru, T., and Dikko, H. G. Digit preference in nigerian censuses data of 1991 and 2006. *Epidemiology, Biostatistics and Public Health vol 10*, 2 (2013).

[16] Dechter, A. R., and Preston, S. H. Age misreporting and its effects on adult mortality estimates in latin america. *Population Bulletin of the United Nations*, 31-32 (1991), pg 1–16.

[17] Economic, and for Western Asia (ESCWA), S. C. *A study of Age Reporting in Some Arab Censuses of Population*. 2013.

[18] Ewbank, D. C. Age misreporting and age-selective underenumeration: sources patterns and consequences for demographic analysis. . (1981).

[19] Ghosh, A. *Demographic trends in West Bengal during 1901-1950*, vol. 5. Office of the Registrar General, India, Ministry of Home Affairs, 1967.

[20] Gibril, M. *Evaluating census response errors: a case study for the Gambia.* Organisation for Economic Co-operation and Development. Development Centre Paris France., 1979.

[21] Harling, G., Tanser, F., Mutevedzi, T., and Barnighausen, T. Assessing the validity of respondents reports of their partners' ages in a rural south african population-based cohort. *BMJ open vol 5*, 3 (2015), e005638.

[22] Haslett, S., and Jones, G. Potential for small area estimation and poverty mapping at constituency and at gewog/town level in bhutan. *World Food Programme, New Zealand* (2008).

[23] Hill, M., Preston, S., Elo, I., and Rosenwaike, I. Age-linked institutions and age reporting among older african americans. *Social Forces 75*, 3 (1997), 1007–1030.

[24] Hill, M. E., Preston, S. H., and Rosenwaike, I. Age reporting among white americans aged 85+: Results of a record linkage study. *Demography vol 37*, 2 (2000), pg 175–186.

[25] Hobbs, F. *Age and sex composition in Methods and materials of demography*, 2 ed. Elsevier Academic Press, San Diego, CA, 2004.

[26] Hock, C. M. *The Early Chinese Newspapers of Singapore, 1881-1912.* University of Malaya P; Oxford UP, 1967.

[27] Kabir, M., and Chowdhury, K. The pattern of age reporting errors in the districts of bangladesh. *Rural demography vol 8*, 2 (1981), pg 33–54.

[28] Kamleu, G. *Assessing the quality of demographic data on age and sex collected from census 2001, General Household surveys (2004-2007), Labour Force surveys (2005-2007) and Community survey 2007 in South Africa.* PhD thesis, University of the Western Cape, 2012.

[29] Khalfani, A. K., Zuberi, T., Bah, S., and Lehohla, P. J. Population statistics. *The Demography of South Africa* (2005), pp 3.

[30] Kpedekpo, G. *Essentials of demographic analysis for Africa.* London England/Exeter NH Heinemann Educational Books 1982., 1982.

[31] Larose, D. T., and Larose, D. T. *Data mining methods and models*, vol. 12. Wiley Online Library, 2006.

[32] Locker, T. E., and Mason, S. M. Digit preference bias in the recording of emergency department times. *European Journal of Emergency Medicine vol 13*, 2 (2006), pg 99–101.

[33] Mba, C. In *Assessing the reliability of the 1986 and 1996 Lesotho census data* (2003), vol. 18, School of Social Work, University of Zimbabwe.

[34] Mills, M. In *Providing space for time: The impact of temporality on life course research* (2000), vol. 9, Sage Publications, pp. 91–127.

[35] Mukherjee, B. N., and Mukhopadhyay, B. K. A study of digit preference and quality of age data in turkish censuses. *Genus* (1988), pg 201–227.

[36] Myers, R. J. Validity of centenarian data in the 1960 census. *Demography vol 3*, 2 (1966), pg 470–476.

[37] Nagi, M. H., Stockwell, E. G., and Snavley, L. Digit preference and avoidance in the age statistics of some recent african censuses: Some patterns and correlates. *International Statistical Review/Revue Internationale de Statistique* (1973), pg 165–174.

[38] Newell, C. *Methods and models in demography.* Guilford Press, 1990.

[39] Obisesan, K. O., and Mojoyinola, M. O. Error detection in census data age reporting. *Botswana Notes and Records vol 46* (2015).

[40] Obisesan, K. O., and Mojoyinola, M. O. Error detection in census data age reporting. *Science, Engineering and Innovative Research vol 7* (2016).

[41] Opiniano, J. M. *Our future beside the exodus: Migration and development issues in the Philippines*. Friedrich-Ebert-Stiftung, 2004.

[42] Organisation, W. H. Sex ratio. `http://www.searo.who.int/entity/health_situation_trends/data/chi/sex-ratio/en/`. Accessed April 19, 2019.

[43] Pal, J. K., Mukhopadhyay, B. K., and Tewari, H. Age reporting error and effect of education: A village study. *American Journal of Social Science Research 1*, 3 (2015), 158–162.

[44] Palamuleni, M. E. Age misreporting in malawian censuses and sample surveys: An application of the united nations' joint age and sex score. *Southern African Journal of Demography* (1995), pg 11–17.

[45] Pardeshi, G. Age heaping and accuracy of age data collected during a community survey in the yavatmal district, maharashtra. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine vol 35*, 3 (2010), 391.

[46] Pullum, T. A statistical reformulation of demographic methods to assess the quality of age and date reporting, with application to the demographic and health surveys. In *Annual Meeting of the Population Association of America, Philadelphia, March* (2005).

[47] Quandt, A. S. The social production of census data: interviews from the 1971 moroccan census. (1974).

[48] Rosenwaike, I., and Logue, B. In *Accuracy of death certificate ages for the extreme aged* (1983), vol. 20, Springer, pp. 569–585.

[49] Seng, Y. P. In *Errors in age reporting in statistically underdeveloped countries: With special reference to the Chinese population of Singapore* (1959), vol. 13, Taylor & Francis, pp. 164–182.

[50] Shryock, H. S., Siegel, J. S., and Larmon, E. A. *The methods and materials of demography*. US Bureau of the Census, 1973.

[51] Statistics, S. A. Census 2001: Census in brief. *Statistics South Africa.* (2003).

[52] Steinberg, J. *In Proceedings of the Social Statistics Section.* American Statistical Associatio, washington, D.C, 1966.

[53] Stockwell, E., and Wicks, J. Age heaping in recent national censuses. *Social Biology vol 21*, 2 (1974), 163–167.

[54] Tekpli, G. *Evaluation and Adjustment of Age Sex Data of the Population and Housing Census of Ghana 2000 and 2010.* PhD thesis, University of Ghana, 2013.

[55] Thavarajah, S., White, W., and Mansoor, G. In *Terminal digit bias in a specialty hypertension faculty practice* (2003), vol. 17, Nature Publishing Group, p. 819.

[56] Ueda, K. *Accuracy of census sex-age data in Asia and the Pacific countries, Institute of Developing Economies Statistics Division Tokyo Japan.* Institute of Developing Economies Statistics Division Tokyo Japan 1976., 1976.

[57] Uganda. Uganda bureau of statistics. `http//www.ubos.org/`. Accessed March 14, 2019.

[58] Van de Walle, E. Marriage and marital fertility. *Daedalus* (1968), pg 486–501.

[59] Vostrikova, A. The importance of data on the age and sex structure of the population. *Problems in Economics vol 12*, 11 (1970), pg 27–39.

[60] West, K. K., Robinson, J. G., and Bentley, M. Did proxy respondents cause age heaping in the census 2000? *ASA Section on Survey Research Methods* (2005), pg 3658–3665.

[61] Woodward, M. *Epidemiology: study design and data analysis.* CRC press, 2013.

[62] Yazdanparast, A., Pourhoseingholi, M. A., and Abadi, A. Digit preference in iranian age data. *Italian Journal of Public Health vol 9*, 1 (2012).