DEVELOPMENT OF CENSUS OUTPUT AREAS IN SOUTH AFRICA

By

THOLANG ALFRED MOKHELE

SUPERVISED BY PROF ONISIMO MUTANGA AND PROF FETHI AHMED

Submitted in fulfilment of the academic requirements for the degree of Doctor of Philosophy
in the School of Agricultural, Earth and Environmental Sciences
University of KwaZulu-Natal
Westville Campus
South Africa

December 2015

ABSTRACT

The use of the same geographical unit for both census data collection and dissemination is common in many countries across the world, especially in developing countries. This poses some serious concerns. Firstly, this practice has caused various difficulties for the census data users as the ideal characteristics of an area to facilitate efficient census data collection differ considerably from those which aid analysis and interpretation of the published data. Secondly, some Enumeration Area (EA) populations fell below the census confidentiality limits, requiring the data to be combined with those of a nearby EA. Thirdly, the design of EAs before census data collection does not take into account local social divisions in boundary placement. Lastly, the shape compactness of areas is often ignored. In order to address these four concerns, the advanced techniques of automated zone design methods, such as Automated Zone-design Tool (AZTool), are required for the development of suitable output areas in South Africa that would address the four concerns as much as possible. Therefore, the overall aim of this study was to develop optimized census output areas using AZTool program in South Africa. In order to achieve this aim, among others, the following research objectives had been developed; firstly, the creation of output areas using AZTool program with the 2001 census EAs as building blocks in South Africa. Subsequently, the determination of the statistical qualities of the AZTool generated output areas with regard to population target mean, minimum population threshold, social homogeneity and shape compactness was explored. In addition, the comparison of the newly created output areas with existing census small areas was also considered. The study area comprised of two of the nine provinces of South Africa. These included the Free State (representing rural settings) and Gauteng (representing urban areas). This study employed EAs from the 2001 census estimates (HSRC, 2005) as building blocks for creating new census output areas in South Africa. The 2001 census SubPlaces, the 2001 census Small Area Layers (SALs) and 2011 census SALs data were also explored for further evaluation the AZTool program. In order to validate results from the AZTool program, some analyses such Analysis of Variance (ANOVA), Shapiro-wilk test, paired t-test, and Kolmogorov sminov test were performed using Statistical Package for Social Sciences (SPSS). Results showed that the primary criterion of minimum population threshold of 500 people (which is the official minimum population threshold used by Statistics South Africa) was kept and not breached throughout all the AZTool newly created output areas at different geographical levels as well as in both rural and urban areas.

Furthermore, the Intra-Area Correlation (IAC) of 0.62 for the two provinces (Free State and Gauteng) combined indicated that the selected homogeneity variables (geotype and dwelling type) were good indicators of social homogeneity for creating optimised output areas in South Africa. It was also found that the newly AZTool generated census output areas out-performed the existing official SALs and SubPlaces, non-zone design developed geographies. This was proven by the fact that AZTool output areas effectively satisfied minimum and target population thresholds, while the population distributions were much narrower in range than those of the existing SALs and SubPlaces. However, the AZTool created output areas were less compact in shape than the SALs and SubPlaces in all geographical regions. In general, there was statistically significant ($p < 0.05$) difference in Perimeter Squared per Area (P2A) means between the output areas and the SALs. The LSD post-hoc test revealed that difference between the P2A means for the AZTool output areas and the SALs was not statistically significant ($p > 0.05$). Therefore, it was concluded that there is potential in application of automated zone design methods, particularly AZTool program, in the creation of optimized census output areas in South Africa. It was also concluded that findings from this study contribute to the research in general and to the potential applications of automated zone design methods in developing countries. One of the main recommendations is that further research and general work should evaluate the application of automated zone design methods, such as AZTool computer program, in the creation of census output areas across the entire country. In addition, data should be made accessible at lower geographical level such as EA or household levels even if it is under secure conditions to allow robust developments of optimized census output areas using automated zone design techniques.

PREFACE

This study represents original work by the author and has not been submitted in any form for any degree or diploma to any other tertiary institution. Where use has been made of the work of others it is duly acknowledged in the text.

DECLARATION 1 – PLAGIARISM

I, Tholang Alfred Mokhele, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
    a. Their words have been re-written but general information attributed to them has been referenced
    b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References section.

Signed:____________________________

DECLARATION 2 – PUBLICATIONS

1. **Mokhele TA**., Mutanga O. and Ahmed A. The use of GIS in census operations: An overview. *South African Journal of Science. (Submitted).*

2. **Mokhele TA**., Ahmed A. and Mutanga O. Development of census output areas with Automated Zone Tool in South Africa. *South African Journal of Science.* ***(Accepted/In press)***.

3. **Mokhele TA**., Ahmed A. and Mutanga O. The statistical qualities of the AZTool census output areas. *(In Preparation).*

4. **Mokhele TA**., Mutanga O. and Ahmed A. Effects of different building blocks designs on the statistical characteristics of AZTool output areas. *International Journal of Geographical Information Science. (Submitted).*

5. **Mokhele TA**., Mutanga O and Ahmed A. Comparison of AZTool output areas with existing official census dissemination areas in South Africa. *International Journal of Geographical Information Science. (Submitted).*

All the conceptual and experimental work, analysis of data, writing and preparation of above publications were accomplished by the candidate, Tholang Alfred Mokhele, under the supervision of Prof Onisimo Mutanga and Prof Fethi Ahmed.

Signed: ____________________________

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Above all, I would like thank God for giving me strength which allowed me to achieve this milestone.

I would like to thank Prof David Martin and Dr Samantha Cockings for permission to use AZTool program and their technical advice, especially Dr Cockings. I would also like to thank the following people who proof-read different papers that are part of this thesis; Prof Zitha Mokomane, Dr Elfatih Abdel-Rahman, Dr Gina Weir-Smith, Dr Mokhantso Makoae, Ms Leana Meiring and Dr Konosoang Sobane.

Great thanks go to the University of KwaZulu-Natal and the Human Sciences Research Council for their funding assistance which made this research study successful.

Kind regards are owed to Prof Fethi Ahmed and Prof Onisimo Mutanga for their remarkable supervision and support throughout this research. Indeed I learnt a lot from them.

Great regards are also due to my family, friends and colleagues. Those words of "*when are you finishing?*" provided a motivation indeed. In addition, their support has not gone without suffering for them, especially my wife and my mother, Malemohang Mokhele and Malisemelo Mokhele. Their valuable support throughout this milestone was incomparable, unconditional and unbelievable. To my three daughters, those weekends when I had to leave you and mom are not gone unnoticed, you will understand as you grow up that it was worth it.

DEDICATION

This Thesis is dedicated to my second and third daughters, Limpho Grace Mokhele and Liteboho Precious Mokhele, who were born during this milestone.

CHAPTER 1

INTRODUCTION

1.1 Background

Many countries use the same geographical area for both census data collection and dissemination. There are some concerns related to the use of the same geographical unit for both census data collection and dissemination. Firstly, this had caused various difficulties for the census data users as the ideal characteristics of an area to facilitate efficient data collection vary considerably from those aimed at analyzing census data as well as interpreting the published data (Martin, 1998a). Secondly, some EA populations fell below the census confidentiality limit, forcing the data to be combined with those of a nearby EA or ED (Martin, 1998a; 1998b; Verhoef and Grobbelaar, 2005; Grobbelaar, n.d.). Thirdly, the design of EAs before census data collection does not take local social divisions in boundary placement into account (Martin, 2004; Verhoef and Grobbelaar, 2005). Lastly, the shape compactness of areas is often not prioritised (Martin, 1998a; 1998b). In addition to this, the fact that census data is collected at household or individual level but disseminated at higher level leads to the scale and zoning problem called the Modifiable Areal Unit Problem (MAUP). The Geographic Information Systems (GIS), together with its related technologies such as automated zone design tools, is growing exponentially to address the above conceptual issues. It is worth noting that the capacities and capabilities in developing countries are not as much as they are in developed countries. The conceptual underpinning of the automated zone design tools emanate from the above mentioned issues. For instance, the first automated zone design program, Automated Zoning Procedure (AZP), developed by Openshaw (1977) was an attempt to solve the MAUP problem. The current AZTool program is derived from the AZP program.

A population census involves counting the number of people at a given point in time, collecting, compiling, evaluating, analyzing and disseminating information about their demographic, social and economic characteristics (Margeot and Ramjith, 2001; Stats SA and HSRC, 2007).Census data is the foundation for a wide range of analyses essential to improve the standard of living of people in any country (Schwabe, 2003). Census data is a powerful tool for world development and poverty reduction (Margeot and Ramjith, 2001; Stats SA and HSRC, 2007; Owiti, 2008). It also provides a sampling framework for surveys that provide further insights into demographic and socio-economic trends, with the aim of assessing, monitoring and evaluating the implementation of policies and programs (Margeot and Ramjith, 2001; Stats SA and HSRC, 2007; Owiti, 2008).

1.2 Census geography

Spatial analysis and GIS mapping in the census process are determined by the geography on which the census data is collected and disseminated (Stats SA and HSRC, 2007). Census geography is often defined as the division of the country into geographical areas for the purpose of census process (Ralphs and Ang, 2009). An appropriate census geography system should facilitate the management of census itself as well as the publication of small area statistics which meet the needs of census data users (Martin, 2002). Ideally, before any census can take place, the country usually gets divided into census EAs (Stats SA and HSRC, 2007). An EA, also known as Enumeration District in England, Wales, and Northern Ireland (Martin *et al.,* 2001), is normally the smallest geographical area into which the country is demarcated for census data collection. This small area should be of a size to be covered by one enumerator in a given time.

In South Africa, EAs normally contain between 100 and 250 households (Stats SA, 2003; Verhoef and Grobbelaar, 2005). The most important criteria for the design of EAs are as follows: they should not overlap, should be compact without disjointed sections and should cover the entire country. Their boundaries should be physically identifiable and they should be of approximately equal population size to enable an enumerator to cover each one in the allocated time period (Stats SA and HSRC, 2007). It is important to note that it is often not possible that the EAs have equal populations.

1.3 Study motivation

As indicated in section 1.1, most countries use the same demarcation areas for both data collection and dissemination of their census data. In cases where enumeration units are not used for census dissemination, they are normally used as building blocks for developing output areas or zones or they are aggregated to higher spatial levels (Cockings *et al.*, 2013). This aggregation is often done on the basis of geographical location and usually data are made available at two or more spatial levels (Flowerdew, 2011). It is important to indicate that small areas or building blocks would always be of high importance for the dissemination of national population statistics even if census is replaced by other systems such as registers or any administrative datasets, like in Denmark and Finland (Valente, 2010). The reason behind this is that data has to be released at aggregated level to avoid disclosure of personal information of individuals, households, or organisations (Cockings *et al.*, 2011; Flowerdew, 2011). For instance, other countries such as UK use output areas for census dissemination. The automated zone design methods were used for creation of these output areas for both 2001 and 2011 censuses. Nepal also publishes census data in the form of area-based aggregated population distribution for a range of geographical boundaries. Their smallest population enumeration unit is the ward administrative boundary (Dhonju *et al.*, 2015). For Serbia, census data are commonly presented on the level of census designation places which correspond to census blocks in the USA and enumeration districts in UK (Bajat *et al.*, 2013).

For South Africa, in the 1991 and 1996 censuses, the EAs were used for the dissemination of census data. For the 2001 census, it was decided that census data should be disseminated on a geographical level higher than an EA level as a substantial number of EAs had very low populations, posing a threat to personal security disclosure. Therefore, the two names (SubPlaces and MainPlaces) were attached to each EA and a geographic layer was generated from the name attributes by Statistics South Africa (Stats SA) (Verhoef and Grobbelaar, 2005; Grobbelaar, n.d.). The areas created were considered to be too large by majority of the users of census data. Therefore, in 2005 the Small Area Layer (SAL) was developed to address this concern by the census data user needs. The overall goal of the SAL was to have a geographic layer that corresponds as much as possible to the EA layer, but lies within confidentiality thresholds (Verhoef and Grobbelaar, 2005; Grobbelaar, n.d.). The SAL mainly focused on the first two concerns, the census confidentiality thresholds and population size, while the issues of social homogeneity and output shape were not directly addressed. It is worth noting that the confidentiality limit was not addressed as a substantial number of SALs breached this criterion. Therefore, the advanced techniques of automated zone design methods, such as the AZTool program, are required for the development of suitable output areas that would address the four concerns as much as possible in South Africa. Briefly, the AZTool program takes a set of building blocks and iteratively aggregates them into a number of larger areas optimised based on criteria set by the user. The AZTool program does not require ArcInfo for data preparation which makes it an ideal automated zone design program for developing countries.

There are other automated zone design programs such as Automated Zoning Procedure (AZP), Automated Zone Matching (AZM) and automated zone design program (A2Z). The AZP does not allow intersection of two zonal geographical systems while AZM program requires ArcInfo which is not affordable by most users in developing countries. To support this, Klosterman (1995) once suggested that appropriate technology for the developing world does not have to be old fashioned or unsophisticated; instead it has to be cheap, effective, reliable, and easy to use. The A2Z has not been used for any national census dissemination and was not readily available for this study. Therefore, AZTool program was employed for this study. Avenell *et al.* (2009) tried similar automated zone design technique when creating datazones for the analysis of deprivation in South Africa at small area level. They indicated that there were big challenges when dealing with former homelands as empty shells were experienced.

In addition, spatial statistics analysis often requires the aggregation of geographic areas into larger areas to preserve confidentiality, to minimize population differences, to reduce the effects of outliers in the data, or just to facilitate the visualization and interpretation of information in maps (Duque *et al.,* 2007). The investigation of different methods in the creation of census output areas is of high importance worldwide; hence South Africa is no exception. Therefore, this study explored the creation of census output areas in South Africa using the 2001 census data as the baseline input. The 2011 census data was also explored to some extent. The study further compared the newly AZTool created census output areas with existing South African official census output areas.

1.4 Study aim and objectives

This section highlights the main aim of this study. The specific objectives that helped in achieving this aim are also highlighted in this section.

1.4.1 Aim

The overall aim of this study was to develop an optimized census output areas using AZTool program in South Africa. The specific objectives are as follows:

1.4.2 Specific objectives

1. To develop optimised output areas using AZTool with 2001 census EAs as building blocks.
2. To determine the statistical qualities of the AZTool created output areas.
3. To determine the effects of building blocks designs on the statistical characteristics of AZTool output areas.
4. To compare the AZTool developed output areas with existing official census dissemination areas in South Africa.
5. To evaluate the AZTool application in South Africa.

1.5 Methods

This section summarizes the study area, overall methods and data used as well as analysis of data. The study area comprised of two of the nine provinces of South Africa. These included the Free State (representing rural settings) and Gauteng, the most populated province but the smallest in area (representing urban areas). This study employed EAs from the 2001 census estimates (HSRC, 2005) as building blocks for creating new census output areas in South Africa. The 2001 census data based on SupPlaces, here after termed “2001 SubPlaces” data, the 2001 census data based on SAL (Verhoef and Grobbelaar, 2005), here after termed “2001 SALs” and 2011 census SALs were also explored. Table 1.1 shows different census datasets that were used in this study as well as the homogeneity variables that were analysed from these datasets. The reason for using EAs from 2001 census estimates instead of 2011 census EAs was that the accessibility of recent data at lower geographical levels such as household and EA levels from the national statistics office (Stats SA) was not successful.

The AZTool software version 1.0.3 (Cockings *et al.*, 2011) was used for the creation of optimal output areas in this study. This software was derived from the AZP which was developed by Openshaw (1977). The user may also set various options as to how the AZTool would operate for example: how many iterations and swaps the AZTool should run; whether donuts are allowed or not (that is, one output area within another); setting minimum boundary length; and allowing the output areas to be wholly contained within higher geographical levels or regions (Ralphs and Ang, 2009).

In this study, the minimum population threshold, population target, shape and homogeneity criteria were pre-defined in the creation of these optimised output areas. In this study, *census dissemination* is the process of releasing and publishing census information for wide use by the public. *Census confidentiality limit* is defined as the minimum population that is used for disseminating or publishing census data in order to avoid personal information disclosure. *Social homogeneity* is defined as the state of all areas being the same based on particular variables. *Shape compactness of the area* represents the degree to which the area has a compact (rather than linear) shape. For instance, a minimum population of 500 (Verhoef and Grobbelaar, 2005) and a population target of 1000 were set. For homogeneity, this study employed the Intra-Area Correlation (IAC), which is a direct measure of within-area homogeneity and between-area heterogeneity (Tranmer and Steel, 1998; 2001; Martin *et al.*, 2001; Flowerdew, 2011; Cockings *et al.*, 2013). The higher the IAC values, the higher the degree of homogeneity. With regard to compactness of the shape of output areas, the study used the overall Perimeter Squared per Area (P2A) (MacEachren, 1985; Cockings and Martin, 2005; Haynes *et al.*, 2007) as a measure of shape compactness. In contrast to the IAC values, the lower the shape (P2A) means values or P2A scores, the more compact are shapes of the output areas. ESRI’s ArcGIS 10.2 and Microsoft excel were employed for data preparation to be used by the AZTool software and for displaying AZTool output results.

Table 1.1: Different census datasets and homogeneity variables used

2001 EAs and SubPlaces	
Dwelling Type	**Geotype (Geography type)**
1=House or brick structure on a separate stand or yard	1=Formal Urban
2=Traditional dwelling/hut/structure made of traditional materials	2=Informal Urban
3=Flat in block of flats;	3=Informal Rural (Tribal areas)
4=Town/cluster/semi-detached house (simplex, duplex, triplex)	4=Formal Rural (Farms)
5 =House/flat/room in back yard	
6=Informal dwelling/shack in back yard	
7=Informal dwelling/shack NOT in back yard, e.g. in an informal/squatter settlement	
8=Room/flatlet not in back yard but on a shared property	
9=Other dwelling	
2001 SALs	
Dwelling Type	**Geotype (Geography type)**
No dwelling type variable	1=Urban
	2=Rural
	3=Mixed
2011 SALs	
Dwelling Type	**Geotype (Geography type)**
1=House or brick/concrete block structure on a separate stand or yard or on a farm	1=Urban
2=Traditional dwelling/hut/structure made of traditional materials	2=Tribal
3=Flat or apartment in a block of flats	3=Farm
4=Cluster house in a complex	
5=Townhouse (semi-detached house in a complex)	
6=Semi-detached house	
7=House/flat/room in back yard	
8=Informal dwelling (shack in back yard)	
9=Informal dwelling (shack not in back yard, e.g. in an informal/squatter settlement/on a farm)	
10=Room/flatlet on a property or a larger dwelling/servants' quarters/granny flat	
11=Caravan/tent	
12=Other dwelling	

Statistical Package for Social Sciences (SPSS) was used for inferential statistical analyses. These included Analysis of Variance (ANOVA), Shapiro-wilk test, paired t-test, and Kolmogorov sminov test. The ANOVA was used to test the statistical significance of the difference between the P2A means of two geographic areas. The Least Significant Difference (LSD) post-hoc test was performed for comparing means for more than two groups. A paired t-test was perfomed to see if the means from the two datasets were the same. The Shapiro-wilk test and Kolmogorov sminov test were employed for normalcy testing of population distributions.

1.6 Thesis outline

This thesis is structured in a form of papers or articles which are either accepted, submitted or in preparation. These papers are in line with the above mentioned objectives of the study. Chapter 1 entails the background as well as the motivation as to why the study was conducted. The aim and objectives of the study are also presented in this chapter. This chapter further highlights the overall methodology and data used in this study as well as the outline of this thesis. Chapter 2 provides an overview of the use of GIS in census operations. Further details about the GIS applications for census use in South Africa are also dealt with.

Chapter 3 contains development of census output areas using AZTool in South Africa. The 2001 census EAs were used as building blocks in different spatial levels in both rural and urban settings in South Africa. Chapter 4 explores the statistical quality of the AZTool created census output areas with regard to population target mean, minimum population threshold, social homogeneity and shape compactness.

Chapter 5 evaluates the effects of different building blocks on the statistical characteristics of output areas generated using the AZTool program. Different spatial layers from the 2001 census data were used as building blocks for generation of census output areas in both rural and urban settings of South Africa. Chapter 6 provides the comparisons of the newly AZTool developed output areas with existing official census output areas in South Africa. Chapter 7 contains the summary of findings, study limitations, conclusions and recommendations gathered from the preceding chapters of this thesis.

CHAPTER 2

THE USE OF GIS IN CENSUS OPERATIONS: AN OVERVIEW

This chapter is based on

Mokhele TA., Mutanga O. and Ahmed A. The use of GIS in census operations: An overview. *South African Journal of Science. (Submitted).*

Abstract

The application of GIS in census processes is rapidly growing. Among the reasons for this rapid growth is that there is a relationship between census and geography. The use of GIS in censuses dates back to the mid-20th century. Even though the use of GIS in census is still a new technology in developing countries, especially in Africa, its growth seems to be promising. However, there is a research gap on the potential of GIS in censuses, especially in the analytical mode in these developing countries, including South Africa. Therefore this chapter aimed to provide an overview of the application of GIS in census operations, as well as to unpack the research gaps in the science of GIS in census operations. Among the critical observations made from this overview that require attention in developing countries is limited peer-reviewed literature on the applications of GIS on the census data in developing countries, especially Africa. Furthermore, most countries use the same areas for both census data collection and dissemination. Therefore, further investigation on the creation of optimized census output areas using automated zone design methods such as AZTool computer program in developing countries is required. In order to address this, the following areas should be investigated: the statistical characteristics of the census output areas with regard to minimum population threshold, social homogeneity and shape compactness; the effects of building blocks designs on the statistical characteristics of census output areas and; the comparison of optimised census output areas with existing official geographies that are used for census dissemination in these developing countries.

Keywords: Census, Census geography, Enumeration Areas, GIS, South Africa.

2.1 Introduction

The application of GIS and related technology in census operations is growing. This is proven by the fact that, among 109 country representatives that were interviewed regarding the use of new technologies in their 2010 round censuses, GIS scored the highest (58%), followed by computer-assisted coding (42%) while other scanning methods were 37% (Mbogoni, 2012). Among the reasons for this rapid growth is that there is a relationship between statistics (census) and geography. Lehohla (2005) mentioned that geography needs statistics and statistics needs geography. Hence, geography, with its spatial technologies and data such as

GIS, is indeed an equal partner in the production of statistics (Laldaparsad, 2007). The integration of census and geographic information can not only enhance management of census data but further widens the applications of census data in other potential research areas such as demography, decision-making of population control, sustainable development in society and economy (Wang *et al.,* 2001). This chapter aimed to explore the application of GIS in census operations, primarily in developing countries. It also aimed to identify the obstacles and challenges to the use of GIS in census operations and to highlight areas of need for research into the use of GIS in census operations.

2.2 Importance of population census

As indicated earlier in Chapter 1, a population census involves counting the number of people at a given point in time, collecting, compiling, evaluating, analyzing and disseminating information about their demographic, social and economic characteristics (Margeot and Ramjith, 2001; Stats SA and HSRC, 2007). Census data is the foundation for a wide range of research and analyses required to improve the standard of living of people in any country in the world. One of the most important analytical outputs based on census information are population projections (Schwabe, 2003; Stats SA, 2012). The characteristics of all individuals within a given area are recorded simultaneously during census data collection. Then, this data is used to inform policy making, planning and administration, to facilitate sustainable development in the country, and for research to inform business, industry, labour and the public (Margeot and Ramjith, 2001; Stats SA and HSRC, 2007; Owiti, 2008; Stats SA, 2012; Wall, n.d.). Census data is also used by the private sector for market analyses (Stats SA, 2012). In addition, census data provides a sampling framework for surveys that provide regular insights into demographic and socio-economic trends in order to assess, monitor and evaluate the implementation of policies and programs (Margeot and Ramjith, 2001; Stats SA and HSRC, 2007; Stats SA, 2012)

Most countries conduct their censuses at regular intervals of five or ten years i.e. every 10 years in Lesotho, Botswana and Nigeria and every 5 years in New Zealand. In South Africa, Stats SA is mandated by the Statistics Act No. 6 of 1999 to undertake a census on a five-year cycle. In 2004, the South African Cabinet however pronounced that population censuses would be conducted every ten years (Stats SA and HSRC, 2007). However, statistical data is needed for planning on a yearly basis hence there are also alternatives to census such as registers, surveys, rolling censuses, and linked administrative data in between the census periods. It is very important to have these various data sources geo-referenced or have uniform geographic frames especially small area estimation.

2.3 Census geographies

GIS mapping operation in the census process is primarily determined by the geography on which the census data is collected and disseminated (Stats SA and HSRC, 2007). Census geography is defined as the division of the country into spatial units for census process (Ralphs and Ang, 2009). Ideally, an appropriate census geography system should facilitate the management of census process as well as the publication of small area statistics which meet the needs of the census data users (Martin, 2002). Before any census can take place, the country usually gets demarcated into census EAs (Stats SA and HSRC, 2007). An EA, also known as Enumeration District in UK, is the smallest spatial area into which the country is demarcated for census enumeration, of a size enough to be covered by one census enumerator in a given time period. In South Africa, EAs usually comprise of between 100 and 250 households (Stats SA, 2003; Verhoef and Grobbelaar, 2005). This is in close proximity with other African countries such as Botswana, Mozambique, Ethiopia, Tanzania, Kenya, Zambia and Mauritius with EA sizes ranging between 70 and 200 households (Ndubi, 2007; UN, 2007).

Among the criteria for the design of EAs in South Africa are that: they should not overlap; they should be compact without disjointed sections; they should cover the entire country; should have identifiable boundaries on the ground; and they should be of approximately equal population size to enable an fieldworker to cover each one in the allocated census period (Stats SA and HSRC, 2007).

Most countries use the same EAs for collecting and disseminating their census data. Some exceptions exist such as in the UK where they use Output Areas (OAs) for their census disseminations (Duke-Williams and Rees, 1998; Martin *et al.*, 2001; Martin, 2004; Cockings *et al.*, 2011; Martin *et al.*, 2013). The practice of using the same EA for both census data collection and dissemination raises some concerns. Firstly, this had caused various challenges for the census data users as the ideal characteristics of an area to facilitate efficient data collection differ considerably from those which are aimed at assisting in the analysis and interpretation of the disseminated census data (Martin, 1998a). Secondly, some EA populations fell below the census confidentiality thresholds, requiring the data to be combined with those of a nearby EA or other statistical disclosure control measures (Martin, 1998a; 1998b; Verhoef and Grobbelaar, 2005; Grobbelaar, n.d.). Thirdly, the design of EAs before census data collection does not take local social divisions in boundary placement into account (Martin, 2004; Verhoef and Grobbelaar, 2005). Finally, the issue of shape compactness of these areas is somehow not prioritised (Martin, 1998a; 1998b). Compactness of the zone shape is crucial especially for accuracy of EA values as well as urban morphology.

For South Africa, in the 1991 and 1996 censuses, the EAs were used for census data dissemination. During the 2001 census, a decision was taken by Stats SA that census information must be disseminated on an area larger than an EA because of issues around personal security disclosure. In order to distinguish between the different settlements types, EAs within the same locality with the same EA type were merged to create a SubPlace name (Dube, 2005). In many cases, most census data users felt that the SubPlaces were too big for, hence the Small Area Layer (SAL) was created in 2005. The automated (non-zone design) spatial creation of the SAL was based on the principle of merging individual EAs if they are within the same SubPlace; they have the same EA geography type; population of each EA is less than 500 and; the SALs should have a total population of 500 and more. The overall aim of the SAL was to have a geographic level that corresponds as much as possible to the EA layer, but lies within confidentiality thresholds (Verhoef and Grobbelaar, 2005).

The SAL mainly focused on the first two concerns, the confidentiality thresholds and population size, leaving the problems of social homogeneity and output shape not directly addressed. It is important to note that even the confidentiality threshold limit was not fully addressed as there was a substantial number of SALs which breached confidentiality limit. Therefore the advanced techniques of GIS, such as automated zone design methods, are required for the development of optimised census output areas in South Africa that would address the four concerns as much as possible. Avenell *et al.* (2009) tried a similar method when creating datazones for the analysis of deprivation at small area level in South Africa. They highlighted that were big challenges when dealing with former homelands as empty shells were experienced. Therefore, there is a need for thorough research on these two methods in order to develop the appropriate approaches in South Africa with regard to creation of small area statistics or output areas.

2.4 The use of GIS in census operations

Globally, it is believed that GIS provides significant benefits to the census process in relation to reducing cost and time for pre-census activity (Martin, 1998b; Mbogoni, 2012). Most of the emphasis to-date using GIS in census had focused on the pre-census activity of mapping and defining areas where enumeration could take place (Tye, 2009). GIS is one of the main technologies used to represent census information in electronic format and facilitates further spatial analysis of the information (Schwabe, 2003). The census database generally contains attribute information which can be linked to spatial units by geo-referencing. Therefore, relating the spatial component with the non-spatial attributes of the existing corporate information enhances the understanding of the user and gives new insights into the patterns and relationships in the data that would not be found (Owiti, 2008).

The use of GIS in census processes is crucial as it plays a major role in efficient time use. For instance, census data capturing activity of the first 1984 census of Ethiopia needed about two years using a main frame system while the second census took more than a year using a stand-alone Personal Computer (PC) system involving 180 data capturers with 90 PCs (Seid and Gutu, 2009). In addition, the use of scanning technology for data capturing had shortened the data capturing period into four months, hence the census data was disseminated within a relatively shorter period of time compared to previous censuses (Seid and Gutu, 2009).

Traditionally, the role of maps in the census process had been to support production of enumerator maps and to present aggregate census results in cartographic form. In addition to these, GIS currently plays a key role in the dissemination and analysis of population and household census data (UN, 2000; 2009; Loots, 2005; Rain, 2008; Seid and Gutu, 2009; Mbogoni, 2012). The purpose for which the map would be used is the one that normally determines the geographic level, the layout, the colours and text fonts, the size and quality of the paper, and the degree of effort that must be put into making the map graphically satisfying (Stats SA and HSRC, 2007).

GIS mapping has been an integral part of census taking place for a long time. Very few enumerations during the last several census rounds were executed without the help of detailed maps, globally. In general, GIS mapping plays a critical role in all three stages in the census process: pre-enumeration; during enumeration and post-enumeration (UN, 2000; 2009; Rain, 2008). Furthermore, the data integration functions provided by GIS have led to a much wider use of census (statistical) information. This has exerted pressure on National Statistical Offices (NSOs) to produce high-quality geo-referenced information for small spatial areas. Types of applications for this data are almost limitless (UN, 2000; 2009). These include: Poverty analysis - small area census information, together with geo-referenced information on infrastructure and agro-ecological conditions, can be used to estimate poverty levels and the location of poor communities. Utility service planning - private and public water, electricity and telecommunications utilities also use geo-referenced data to assess current and future demand for services. Emergency planning – geo-referenced census data, in combination with digital elevation and transportation maps, are essential tools in the identification of highly populated areas that are difficult to evacuate during emergencies (UN, 2000; 2009; Maantay *et al.*, 2007). The commonality of these illustrations is that they rely on the availability of small area demographic and social data. Therefore, the only reliable sources of such information are censuses or population registration systems in some countries (UN, 2000; 2009).

Most developed countries as well as some developing countries such as, United Kingdom, United States of America, Canada, Japan, France, Australia, China, Israel, India, Nepal, Bangladesh, South Africa, Uganda, Malawi, and many others, have successfully applied GIS in census operations in the past decades (Mobbs, 1998; Wang *et al.*, 2001; Prasad, 2006; Dhonju *et al.*, 2015; Khatun *et al.*, 2015). For instance, the United States of America has played a distinguished role in the census GIS from the 1980s (Wang *et al.*, 2001). The US Census Bureau developed Topologically Integrated Geographic Encoding and References (TIGER) system which was applied in the 1990 census (Wang *et al.*, 2001). In Japan, the Census Mapping System started during the years 1991 – 93. Furthermore, the Israel 1995 census linked the census data to spatial elements as small as the buildings while in 2004, Uganda Bureau of Statistics (UBOS) also adopted the GIS policy stating that it was the key to improve their services (Prasad, 2006). The growth of GIS through the 1990's led to a dramatic increase in the utilisation of digital geographic information (Martin, 2000). In some countries such as Trinidad and Tobago, spatial information forms an essential component of government held information. Therefore the government had to collect and maintain large amounts of spatial information on a continuous basis (Wall, n.d.).

In most African countries, the application of GIS technologies in census is still a new phenomenon. Few countries have applied GIS in their censuses and some of them had not made information about the use of GIS in their census available; hence it is not easy to find peer-reviewed literature on the applications of GIS on census data across the continent. Since 1997 NSOs in more than 15 African countries had benefited from GeoSpace International in terms of census mapping solutions, technical support and training. For instance, it provided census mapping solutions to the censuses of Namibia (2001 and 2011); South Africa (2001 and 2011); Tanzania (2002 and 2012); Lesotho (2006) and Southern Sudan (2008). In addition, it provided technical support and training to Botswana, Kazakhstan, Ethiopia, Ghana, Malawi, Nigeria, Seychelles, Swaziland, The Gambia and Zambia (GeoSpace, n.d.). GIS was also used in Zimbabwe 2012 census. Although, the provincial boundaries followed the 2008 boundaries set up by the Zimbabwe Electoral Commission (ZEC), the provinces were further divided into administrative districts and wards. It is important to note that in each province designated urban areas were treated separately from the administrative districts as areas which include Municipalities, Town Councils and Local Boards were given separate codes in the district block of the geocode system (ZimStat, 2013). Other than in national censuses, the GIS and satellite imagery have been applied in a sub-census in Nigeria, where all households in Minna were sampled for assessment of the relationship between housing conditions and rental value (Ajayi *et al.*, 2015).

2.5 The history of census development in South Africa

In South Africa, it is necessary that the history of census taken is investigated before one can start dealing with the applications of GIS technologies in census process. For census taking before 1960, history illustrated that censuses in South Africa have been fragmented, covering only parts of the country and sections of the population. For instance, for the first 1798 census, every head of a household in the Cape Colony had to submit a return stating the size of the family and the number of slaves and cattle that were owned (Stats SA and HSRC, 2007). Population censuses for all races were conducted in the Cape in 1865 and 1875, in the Free State in 1880 and in Natal in 1891 (Stats SA and HSRC, 2007; Christopher, 2009; 2010).

In 1890, only members of the white population took part in the first census undertaken in the Transvaal (Stats SA and HSRC, 2007). The 1904 censuses were conducted in the Cape, Natal, the Free State and the Transvaal. Although they were separate, they did cover the whole country within the same year (Khalfani and Zuberi, 2001; Stats SA and HSRC, 2007; Christopher, 2009). In 1910, when the Union of South Africa was formed, the government decreed in terms of the South African Act of 1909 that a census for all races be conducted in 1910 and thereafter whenever the government felt it was necessary. The censuses that enumerated all population groups were conducted in 1911, 1921, 1936 and 1951. Additional censuses, in which only white people were covered took place in 1918, 1926, 1931 and 1941 (Khalfani and Zuberi, 2001; Stats SA and HSRC, 2007; Christopher, 2009).

Since 1960, all of these censuses were *de facto* censuses, meaning that the participant had to report on all persons, whether a usual resident or not, who spent the night at the particular household. The reference nights of these censuses were 6 May 1970, 6 May 1980, 5 March 1985, 7 March 1991, 9 October 1996, 9 October 2001 and 9 October 2011 (Stats SA and HSRC, 2007; Christopher, 2009; Stats SA, 2012). It is important to note that the 1980, 1985 and 1991 intervening national censuses omitted the areas of the nominally independent homelands of Bophuthatswana, Ciskei, Transkei and Venda. Therefore, these states conducted their own censuses based on their own time-schedules and needs (Khalfani and Zuberi, 2001; Christopher, 2001; 2009).

South Africa is one of the few countries in Africa that has progressed to incorporate the EA boundaries and census attributes into GIS for future census planning and dissemination of the results (Stats SA and HSRC, 2007). However, there is still a research gap with regard to science of GIS applications in spatial analysis of census data in South Africa, as it is the case in other countries across the continent. In South Africa, the Municipal Demarcation Board (MDB) is responsible for maintenance and update of municipality and provincial boundaries while Stats SA is responsible for other geographical hierarchy such as place names and EAs (Dube, 2005). Note should be taken that any changes to municipality boundaries affect the census geography frame because the South African geographical area hierarchy model ensures that a spatial layer at any level is a grouping of smaller areas at lower levels. This frame ensures that each level of geographic layer fits within a hierarchy (Dube, 2005).

As stated earlier in Chapter 1, South Africa had remained one of the few countries in the continent that had not only conducted regular censuses but had progressed to incorporate the EA boundaries and census attributes into GIS for the census planning and dissemination of the results (Stats SA and HSRC, 2007). KwaZulu-Natal was the first province to develop an electronic database that covered the entire province for both the 1985 and 1991 censuses. The success of capturing census boundaries into GIS for KwaZulu-Natal and the identified need by several major service providers and government departments led to the 1991 census boundaries being captured into GIS for most of South Africa (Schwabe, 2003).

Prior to 1996, EA boundaries were hand-drawn, which is traditional demarcation (Laldaparsad, 2007). The 1996 census showed a transition from traditional demarcation and mapping methods towards an electronic spatial database. Up to the 1996 census, administrative maps served as the basis for demarcation. For large new towns and developments, town planning maps were employed. Furthermore, aerial photographs were utilised to pinpoint new residential units. Most maps were insufficient and out-dated. It was difficult to produce map-based publications of census results. The Human Science Research Council (HSRC) took an initiative to address this situation and captured the EAs of the 1991 census digitally (Dube, 2005). A national GIS database for the entire country had been available only since the release of the 1996 census GIS database information in 1999 (Stats SA and HSRC, 2007). Furthermore a decision was taken in 1999 to replace the old census methodology of using Photostat copies of 1:50 000 topographical and municipal maps, hard copy aerial photographs or sketches of the EAs as a base for enumeration with a GIS as the framework for producing high quality, accurate enumeration maps (Margeot and Ramjith, 2001; Laldaparsad, 2007).

For the 2001 census, for the first time, GIS technology was utilized to demarcate EAs (80787 EAs) and for map production instead of the traditional methods of using analogue and sketch maps. For instance, 80% of the 2001 EA demarcation was done in the office on a GIS using aerial photography and digital topographical maps. For the other 20%, field inspection was conducted (Stats SA, 2003; Laldaparsad, 2007). Thus, a comprehensive digital geographic database was developed from several data sets acquired from government departments and private sector companies. These data sets were then integrated into one common geographic frame (Stats SA, 2003).

For the 2011 census, it was expected that the EA demarcation process would be less of a challenge than it was for the previous census (Laldaparsad, 2007). There was awareness that census data should be collected in a structured manner that enables statistical integrity. The MDB had already started this in 2005 (Dube, 2005). That was the movement towards standard municipalities and electoral wards, hence a similar endeavour was needed for other geographies like place names and addresses (Dube, 2005; Gregory and Ell, 2005; Lehohla, 2005). GIS technology was indeed highly incorporated into the census processing. This was shown by the employment of Satellite Pour l'Observation de la Terre (SPOT) 5 2008 imagery in preparation of EA boundaries/demarcation for the 2011 census as well as in the development of Dwelling Unit Frame for the whole country. Each dwelling point was geo-referenced and 18 attributes describing it and its location were collected (Basson, 2007; Laldaparsad, 2007). This process yielded in the country being divided into 103576 EAs (Stats SA, 2012). The similar modern approach had also been used in Namibia 2001, Tanzania 2002, and Lesotho 2006 censuses (UN, 2007).

In summary, the role played by the Stats SA in GIS applications in population censuses is very crucial. This, together with thorough research would lead to reliable and consistent statistics about people of South Africa which would also comply to the UN census guidelines as South Africa is a member state. Therefore both public and private entities would benefit from reliable official statistical information that would lead to economic growth as well as poverty reduction. The use of GIS technology had already benefited Stats SA in terms of cost reductions and reliable data as well as production of re-usable data in comparison to manual methods.

2.6 Challenges of the GIS applications in census and their research

Literature indicates that one of the challenges of GIS applications on census data and their research is the Modifiable Area Unit Problem (MAUP) (Openshaw, 1977; Reynolds, 1998; Ratcli¨e and McCullagh, 1999; Ralphs and Ang, 2009; Weir-Smith and Ahmed 2013; Weir-Smith, 2014). This is because census information is collected at an individual household level (higher resolution) but the data is disseminated at aggregated level (lower resolution, such as EAs /output areas) up to larger levels for dissemination due to confidentiality and data manageability problems (Openshaw, 1984; Yuan *et al.*, 1997). The MAUP is defined as a problem emanating from the imposition of artificial units of spatial reporting on continuous geographic phenomena which results in the generation of artificial spatial patterns (Openshaw, 1984; Heywood *et al.,* 2002). The MAUP consists of two main components, namely *scale problem* (which is the variation in the results arising from the progressive aggregation of smaller zones into larger zones) and *zoning problem* (which is the variation in results arising from different arrangements of a set of zones) (Openshaw, 1977; 1984; Reynolds, 1998; Kitchin and Tate, 2000; Ralphs and Ang, 2009). The MAUP is often associated with the problem called Ecological Fallacy, which occurs when it is inferred that inferences from data for areas under study can be applied to the individuals within those areas (Heywood *et al.*, 2002). The two problems can be avoided if homogeneous data are aggregated into zonal data. However, this is not practical as geographical data are rarely homogeneous. Reynolds (1998) also explored various statistical analyses to solve this MAUP but the results showed that there was a degree of regularity in the behaviour of aggregated statistics data depending on the spatial autocorrelation and configuration of the variable values. Weir-Smith and Ahmed (2013) proportionally aggregated data from the 1991 and 1996 magisterial districts to 2005 municipal boundaries in an attempt to overcome MAUP.

In many countries, administrative boundaries keep on changing for every census period. This makes it difficult to use GIS technology for trend analysis and making comparisons (Dube, 2005; Gregory and Ell, 2005; Lehohla, 2005). Therefore, there is need to maintain existing geographical frames as much as possible to make comparison over certain time possible (Dube, 2005; Gregory and Ell, 2005; Lehohla, 2005). What makes it difficult to maintain these geographic frames is that settlement status, household size as well as population size and composition change with time and space (Dube, 2005; Gregory and Ell, 2005; Lehohla, 2005). These settlements changes normally result in changing of EA boundaries especially if the number of houses and size of the population are to be kept almost uniform for each EA. Exeter *et al.* (2005) hinted this problem can be managed by creating consistent local geographies for recent censuses. Some researchers have explored GIS techniques such as areal interpolation and dasymetric mapping which can be used to minimise this challenge (Gregory, 2002; Maantay *et al.*, 2007; Bajat *et al.*, 2013; Dhonju *et al.*, 2015; Stevens *et al.*, 2015).

Furthermore, the availability of up-to-date maps in developing countries, particularly Africa, makes GIS application in census information challenging. This normally affects pre-census preparation and planning, hence it is not easy to avoid skewed and misleading results (Schlossberg, 2003; Loots, 2005; Wall, n.d.). There is also a huge challenge, especially in developing countries, for GIS research as well as other research communities as census data are not available at lower level due to confidentiality. This limits the possible use of census data by public and private sectors for research purposes (Martin *et al.*, 2001; Wu and Murray, 2007).

In developing countries, census mapping normally relies on a very small permanent staff and a huge temporary workforce, funded directly from the census budget, to execute the census mapping process. This is because most statistical agencies in developing countries, especially in Africa, are usually understaffed and GIS staff is often the worst case (Loots, 2005; Mbogoni, 2012). Furthermore, most newly established GIS offices often close immediately upon completion of a census operation (Loots, 2005; Mbogoni, 2012). The matters are even worse when it comes to GIS research as this is rarely budgeted for by the national statistics agencies (Loots, 2005; Mbogoni, 2012). This exacerbates the limited amount of GIS research in censuses in these developing countries. The limited amount of GIS research in census in developing countries especially in Africa is proven by limited peer-reviewed work in this field. It is important to note that some of work on use of GIS and Remote Sensing in socio-economic data in developing countries especially Africa has been reported, these include but not limited to (Klosterman, 1995; Eze, 2009; Mansour *et al.*, 2012; Linard *et al.*, 2012; Tabatabai *et al.*, 2014; Weir-Smith, 2014; Ajayi *et al.*, 2015; Stevens *et al.*, 2015).

Lastly, GIS is widely perceived to be prohibitively expensive, hence it is not widely utilised in developing countries (UN, 2004; Loots, 2005; Mbogoni, 2012). This is despite the growing awareness of the economic importance of spatial data across the globe (UN, 2004; Loots, 2005). More work still has to be done in order to make developing countries aware of GIS applications in census operations (UN, 2004; Loots, 2005; Mbogoni, 2012). UN (2004) recommended integration of geospatial technologies with census mapping for better decisions, workshops and technical advisory services, as well as the establishment of a group of experts to reflect on these important issues of GIS applications in census operations.

2.7 Opportunities in GIS applications in census and their research

GIS has made a significant contribution to the development of geographical research in general, especially in developed countries (Murayama, 2001; Murray, 2010a; b). GIS technology is proving valuable in the formulation of solutions to the problems at hand. Generally, GIS is utilised in combination with high resolution satellite imagery (such as SPOT 5) to confirm and define administrative boundaries in the development of the EAs (Tye, 2009). As indicated earlier in this chapter, some developing countries believe GIS application in census operation is a costly exercise. Contrary to this, when appropriately applied, GIS technologies could allow much saving in the census operation (Loots, 2005). For instance, findings from Tanzania (2002) and Lesotho (2006) indicated that Geo-information technology-based solutions, which include GIS and remote sensing, could reduce the cost for demarcating EAs by up to 80% (Loots, 2005). Moreover, the integration and utilization of GIS and remote sensing technologies are also becoming easier and more cost effective. There is no doubt that the right tools (GIS and remote sensing technologies) exist and it had never been easier to employ these tools in geospatial applications. Therefore, it was argued that, if implemented as part of a population and housing census process, the technology was within the financial reach and technical capabilities of most statistical agencies in Africa (Loots, 2005).

The role of UN in promoting the utilisation of Global Positioning System (GPS), remote sensing and GIS on census processes is unbelievable, hence it was expected that more countries would adopt GIS applications in census operations in the 2010 round of censuses (Loots, 2005). Indeed it was found that among the new technologies in the 2010 round of census, GIS was the highest with 58% across 109 countries (Mbogoni, 2012). This awareness is done through series of UN continental and worldwide workshops and conferences on census GIS as well as the formation of group experts in this area. This is also an opportunity for GIS research community as they can benefit by joining these expert groups and attending conferences for their ideas to be heard.

One of the lessons learnt from this overview is that there would be a point or time where all UN countries should have standard use of GIS in their census in order for their results to be recognized by the UN. This is also an opportunity for GIS research as researchers would be able to do comparative studies within continents and across the world on UN countries. This would also allow census data to be linked with several datasets within and across countries.

2.8 Conclusions

Based on this overview, it can be concluded that GIS and related technologies on the census are growing in developing countries, especially Africa. Among the critical observations made from this overview that require attention in developing countries are; firstly, limited peer-reviewed literature on the applications of GIS on the census operations in Africa. This does not mean that GIS and related technologies are not used, but documentation of how these technologies were utilised for census preparation and analysis is limited, even from the website of National Statistical Offices. Secondly, most countries use the same areas for both collection and dissemination of the census data; therefore, there is a need for further investigation on the creation of optimized census output areas using automated zone design methods such as AZTool computer program in developing countries. In order to address this, the following areas should be explored: the statistical characteristics of the census output areas with regard to minimum population threshold, social homogeneity and shape compactness; the effects of building blocks designs on the statistical characteristics of census output areas and; the comparison of optimised census output areas with existing official geographies that are used for census dissemination in these developing countries.

CHAPTER 3

DEVELOPMENT OF CENSUS OUTPUT AREAS WITH AZTOOL IN SOUTH AFRICA

This chapter is based on

Mokhele TA., Ahmed A. and Mutanga O. Development of census output areas with Automated Zone Tool in South Africa. *South African Journal of Science.* **(Accepted/In press)**.

Abstract

The use of the same geographical unit for collecting and disseminating census data is common in many countries across the world, especially in developing countries. This poses some challenges such as there possible disclosure of persons, households or organisations' information. The other challenge is the design of small geographic units EAs to facilitate efficient data collection differs considerably from those that aid data analysis and interpretation. For instance, in South Africa confidentiality limit of 500 persons has to be respected for census dissemination. This chapter aimed to create optimised census output areas using the AZTool program with the 2001 census EAs as building blocks in different spatial levels in both rural and urban settings within two South African provinces. Results were consistent and stable because the primary criterion of the confidentiality limit of 500 was respected at all geographical levels or regions as well as in both urban and rural settings for newly created optimised output areas. For the second criterion, the lower IAC values at the lower geographical levels in both rural and urban areas showed that higher geographical levels produced more homogeneous output areas than lower geographical levels or regions. The IAC of 0.62 for the two provinces combined indicated that the selected homogeneity variables were good indicators of social homogeneity for creating optimised output areas in South Africa. It was therefore concluded that the AZTool program could be used to effectively and objectively create optimized output areas in South Africa. Further research on the comparison of the newly created output areas with existing output areas in South Africa should be explored.

Keywords: AZTool; Automated zone design; Census; Enumeration areas; Output areas.

3.1 Introduction

Many countries use the same geographical layer for both census data collection and dissemination. This was also the case in South Africa prior to the 2001 census. This tendency had caused challenges for census data users as the ideal characteristics of an area to facilitate efficient data collection are not the same as those which aid data analysis (Martin, 1998a;

Hofstee and Islam, 2004). Secondly, some EAs populations fall below the confidentiality limits, resulting in these EAs being merged to nearby EAs (Martin, 1998a; 1998b; Verhoef and Grobbelaar, 2005). Thirdly, the design of EAs prior to census collection does not consider social homogeneity such as dwelling or housing type and tenure (Martin, 2004; Verhoef and Grobbelaar, 2005). Lastly, the shape compactness is also not directly considered (Martin, 1998a; b). Some exceptions exist such as in the United Kingdom (UK) where Output Areas (OAs) were used for census disseminations (Duke-Williams and Rees, 1998; Martin *et al.*, 2001; Martin, 2002; 2004; Cockings *et al.*, 2011; Martin *et al.*, 2013).

The fact that census data is collected at household level and disseminated at higher geographical levels or regions such as EAs raises some concerns. This results in a problem called the Modifiable Areal Unit Problem (MAUP), a term first used by Openshaw (1977) but originally established by Gehlke and Biehl (1934). The MAUP comprises of the following: a) *scale problem* – which is the variation in the results caused by the progressive aggregation of smaller areas into larger areas and b) *zoning problem* – which is the variation in results caused by different arrangements of a set of zones (Openshaw, 1977; 1984; Reynolds, 1998; Ratcli¨e and McCullagh, 1999; Kitchin and Tate, 2000; Heywood *et al.*, 2002; Duque *et al.*, 2007; Dumedah *et al.*, 2008; Ralphs and Ang, 2009). Openshaw (1977) developed Automated Zoning Procedure (AZP) in an attempt to solve the MAUP problem. The AZP algorithm works by iteratively combining and recombining sets of building blocks to create output areas which optimise a set of pre-specified design criteria (Martin, 2003; Cockings *et al.*, 2011; Sabel *et al.*, 2013). This AZP was further enhanced by Openshaw and Rao (1995). The AZP was further reviewed and extended to Automated Zone Matching (AZM) software by Martin in 1998 and in 2003 to permit its application to the intersection of two zonal geographical systems. In 2006, Cockings, Martin and Harfoot from University of Southampton developed the AZTool software from AZM. This tool was further enhanced to the current version (AZTool 1.0.3) which does not require Arc Info for preparing .pat and .aat files.

Among many studies on the automated zone design applications, in 2002, Martin and the Office for National Statistics created output areas for the 2001 Census for England and Wales using automated zone design methods. These output areas were designed to respect minimum population and household threshold sizes of 100 and 40 respectively, as well as a compact shape and with a degree of homogeneity in terms of housing tenure and type. In addition, these output areas had to be nested within higher geographical regions. This project was seen as a success even though there were some concerns to the resulting abstract nature of output area boundaries.

The applications of automated zone design techniques were further employed in the health research environment by Cockings and Martin (2005) and Flowerdew *et al.* (2008). For instance, Flowerdew *et al.* (2008) used the 1991 limiting long-term illness (LLTI) data in Great Britain with Enumeration Districts (EDs) as building blocks to construct alternative zonal systems with the AZTool zone design algorithm in order to determine if neighbourhoods defined in various ways would have similar implications for health. Their results showed that, for sets of pseudo-wards that made sense in terms of population equality and shape, the zonation effect was real. Hence, they concluded that it did matter where boundaries are drawn.

Haynes *et al.* (2007) compared automated zone design program ''A2Z'' zones, developed by Daras (2006), with areal units identified subjectively by local government officers as communities in the city of Bristol, UK. They found that automated zone design was close to replicating the subjective communities when the balance of objectives and boundary constraints was adjusted. In 2009, Ralphs and Ang developed new geographies in New Zealand using the AZTool. Their results indicated that the newly created geographies substantially outperformed the current geographies across almost all of their optimisation criteria. Ralphs and Ang (2009) argued that the algorithm they used was stable and consistent hence it could repeatedly generate high-quality solutions in a timely manner. In France, Sabel *et al.* (2013) used the AZTool program (using the 250 x 250 m cells as building blocks) to create new zones to explore relationships between asthma and deprivation in Strasbourg. They found that their newly produced synthetic neighbourhood solution performed better than the then existing IRIS census areas, measured by improved statistical relationships between asthma and deprivation.

In South Africa, for the 1991 and 1996 censuses, the same EAs were used for both census enumeration and dissemination. For the 2001 census, it was decided that census data must be released on an area larger than an EA due to confidentiality (Verhoef and Grobbelaar, 2005). It was for that purpose that two names were attached to each EA and a spatial layer was created from the name attributes (SubPlaces and MainPlaces). In many instances, the areas created were too large for most census data users. In 2005, a non-automated zone design approach was employed to create the SAL for dissemination of the 2001 census in an effort to meet user needs. A similar non-automated zone design approach was also employed in the creation of SAL for the 2011 census data. This was mainly to have a spatial area layer that corresponded as much as possible to the EA layer while adhering to the confidentiality limit of 500 (Verhoef and Grobbelaar, 2005). For instance, the following criteria were set and adhered to as far as possible for the creation of the SAL: firstly, EAs could only be merged if they are within the same SubPlace; secondly, EAs could only be merged if they have the same EA geography type; thirdly, an EA could only be merged if its population is less than 500; and lastly that the resulting small area polygons must have a population total of 500 and more (Verhoef and Grobbelaar, 2005). This resulted into 56255 SALs from 80787 EAs as shown in Table 3.1, which highlights the South African geographical levels or regions that were used for the 2001 census. It is important to note that the maintenance and update of provincial and municipality boundaries is the responsibility of the Municipal Demarcation Board (MDB), while the National Statistics Office (Stats SA) is responsible for the creation and maintenance of MainPlaces, SubPlaces, SALs and EAs.

Table 3.1: South African geographical levels or regions for the 2001 census

Regions	**Number**	**Population Mean**
Provinces	9	4979997
District Municipalities*	52	861923
Local Municipalities*1	257	174397
MainPlaces	3109	14416
SubPlaces	21243	2110
Small Area Layers (SALs)	56255	797
Enumeration Areas (EAs)	80787	555

*Include 6 Metropolitans which are both District and Local Municipalities
[1]Include 20 District Management Areas (DMAs)

In the creation of the SAL, only the census confidentiality limits and population size were addressed while social homogeneity and output shape were not. It is also worth noting that out of 56225 SALs, 13.5% of the SALs breached the confidentiality limit (Verhoef and Grobbelaar, 2005). Although the issue of census output areas being too large from South African census data users’ perspective had been addressed by the creation of the SAL, the issue of confidentiality remains a concern. Policies for census output areas vary from country to country, but confidentiality requirements are strictly enforced in almost all countries. Therefore, the advanced techniques of automated zone design methods such as the AZTool worth exploring for the creation of optimal output areas in South Africa. This chapter attempted to address this by creating census output areas using AZTool software with the 2001 census EAs as building blocks and prioritising the confidentiality limit (minimum population threshold of 500), homogeneity, population mean target and shape compactness. In addition, the chapter examined the performance of the AZTool program for both urban and rural areas in South Africa at different geographical levels or regions. This was to give a general picture as to how the program was likely to perform when the entire country was analysed.

3.2 Methods

This section entails the area where the study was conducted, the AZTool program and how the data was prepared. In addition, the AZTool design criteria and how the results were displayed are also discussed in this section of methods.

3.2.1 Study area

The study area comprised of two of the nine provinces of South Africa (Figure 3.1). These included the Free State (representing rural settings) and Gauteng, the most populated province but the smallest in area (representing urban areas). The uniqueness of the two provinces is that the Free State province has former homelands of Phuthaditjhaba and Botshabelo and it is one of the two provinces which did not experience any provincial boundary change for the 2001 and 2011 censuses. It experienced less than 1.5% population increase within this period. Gauteng on the other hand is the most populated and the most developed province in South Africa, with the highest population growth from 2001 to 2011. Therefore, the analysis of both provinces provides examples of both rural and urban settings in South Africa. Analyses were done at provincial, district, municipality and mainplace levels in each province to gain a better understanding of the performance of the AZTool at each geographical level in both rural and urban settings. Therefore, in the Free State province, Thabo Mofutsanyane district and Maluti-a-Phofung municipality were selected. In addition, Phuthaditjhaba mainplace, (a former homeland) was analysed to get a deeper understanding of the behaviour of the AZTool at lower geographical levels in a rural setting. For Gauteng province, City of Tshwane Metropolitan (which is both a district and a metropolitan municipality) was analysed. Pretoria mainplace was selected from this district/metro in order to explore the potential challenges that might occur in urbanised settings at lower geographical levels or regions.

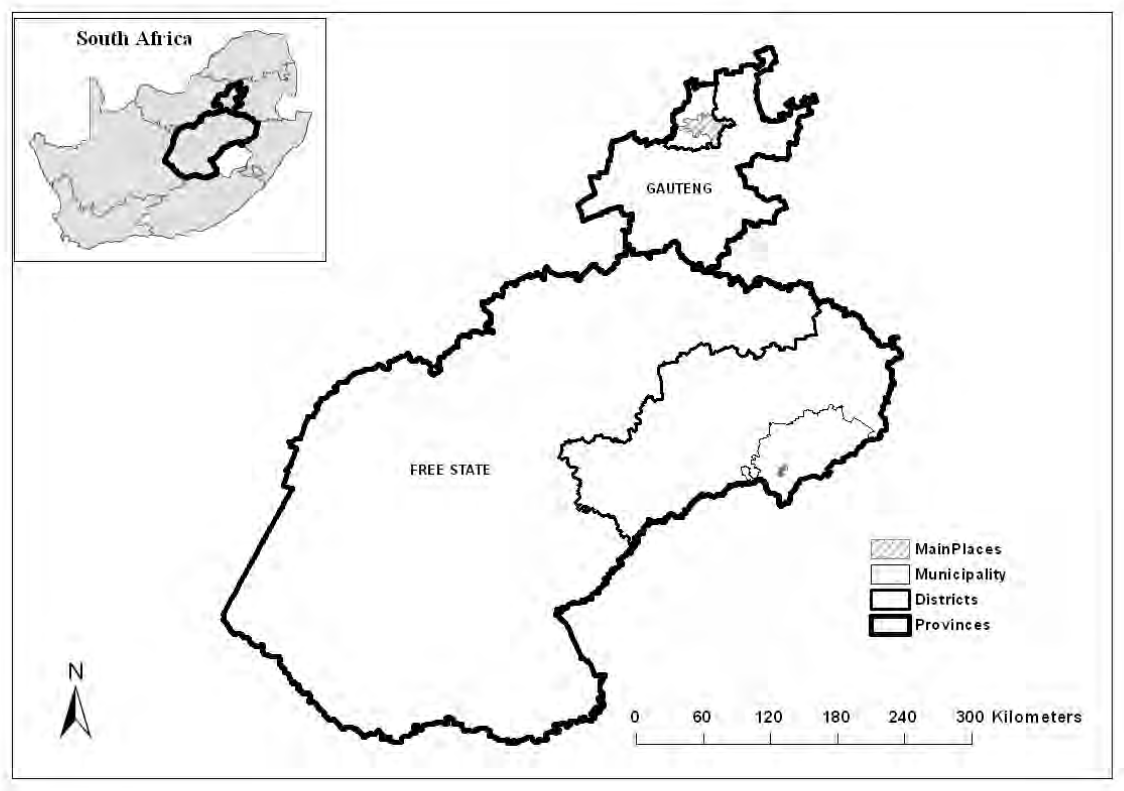


Figure 3.1: Selected study areas

3.2.2 AZTool program

The AZTool software version 1.0.3 (Cockings *et al.*, 2011) was used for the creation of optimal output areas in this chapter. This software was derived from the AZP which was developed by Openshaw (1977). The automated zone design tools normally take input building block geographies and iteratively aggregate them into larger output areas or geographies from an initial random aggregation (IRA) by checking the effect of swapping individual building blocks between output areas based on the criteria set by the user, such as mean population target, minimum population threshold, homogeneity and compactness of the shape. Improved swaps are therefore kept as part of the resultant solution and the IRA is thus refined to give an optimal boundary configuration, based on a particular set of design constraints (Martin, 2003; Sabel *et al.*, 2013). Furthermore, the user may also set various options as to how the AZTool would operate for example: how many iterations and swaps the AZTool should run; whether donuts are allowed or not (that is, one output area within another); setting minimum boundary length; and allowing the output areas to be wholly contained within higher geographical levels or regions (Ralphs and Ang, 2009).

3.2.3 Data preparation

This study employed EAs from the 2001 census estimates (HSRC, 2005) as building blocks for creating new census output areas in South Africa. The accessibility of data at lower geographical levels such as household and EA levels from the national statistics office (Stats SA) was not successful. ESRI's ArcGIS 10.2 was used to prepare data to be used by the AZTool software. The variables employed were total population, dwelling type and geotype as well as higher geographical levels or regions. The geotype variable was the geography type of the EA which was divided into the following: Geotype1=Formal Urban; Geotype2=Informal Urban; Geotype3=Informal Rural (Tribal areas); and Geotype4=Formal Rural (Farms) (see Table 1.1). The AZTool expects the IAC variables to be provided as counts, therefore, the geotype variable, which was categorical with four categories, was further expanded into four attributes i.e. one for each category with a count of 0 or 1. The AZTImporter, which is part of the AZTool software download, was used to convert the building block shapefile (geospatial vector data format) to polygon attribute table (.pat) and arc attribute table (.aat) files which are the format required by the AZTool software.

3.2.4 Zone design criteria

The criteria or rules for the AZTool runs were set in the .xml parameter file. This file specifies the location of both .aat and .pat files as well as defining the parameters, rules, constraints, criteria and column position of variables in the .pat file to be used in the AZTool run. The following criteria were considered for developing optimised output areas:

- Minimum threshold population size, 500 (minimum used by Statistics South Africa (Verhoef and Grobbelaar, 2005))
- Homogeneity – IAC measure of dwelling type and geotype variables
- Shape compactness – Perimeter squared per area (P2A)
- Mean target population – 1000

The minimum threshold population size is a hard constraint, as are the higher geographical regions. Others are soft constraints which are traded off in the objective criteria as in previous studies which also indicated that it is not possible to satisfy all four criteria (Ralphs and Ang, 2009; Cockings and Martin, 2005; Drackley *et al.*, 2011). The weights for population target, homogeneity (IAC score) and shape compactness were left at default weight of 100% indicating that all were weighted equally. The same design criteria were applied to all geographic levels in both rural and urban settings.

Confidentiality limit

The population variable from the 2001 census was used for respecting the confidentiality limit, with a minimum population of 500 set for output areas. This is minimum population threshold that is used by Statistics South Africa, the National Statistical Office in South Africa (Verhoef and Grobbelaar, 2005). Generally, statistical spatial data analysis requires the aggregation of basic spatial areas into larger areas to preserve confidentiality, to minimize population differences, and to reduce inaccuracies in the data (Duque *et al.*, 2007). Therefore, the population target mean was also set to 1000 in this study in order to minimize population differences.

Degree of homogeneity

In order to measure the degree of homogeneity within the created output areas, IAC was employed. The IAC is a direct measure of within-area homogeneity, which is the correlation for a given variable between different people living in the same areal unit (Tranmer and Steel, 1998; 2001; Martin *et al.*, 2001; Flowerdew, 2011). The higher values indicate a higher degree of homogeneity within-area and a higher degree of heterogeneity between areas (Tranmer and Steel, 1998; Martin *et al.*, 2001; Cockings *et al.*, 2013). The homogeneity variables that were selected from the 2001 census data included dwelling type and geotype.

The dwelling type or housing type is the commonly used variable as a proxy for social built environment homogeneity measure, as it had been identified as one of the variables that tend to experience the greatest degree of homogeneity (Martin *et al.*, 2001; Ralphs and Ang, 2009). It was therefore also applied in this paper. The EA geographic type (Geotype) was also used as one of the rules for creating SAL which was used to disseminate the 2001 census data in South Africa (Verhoef and Grobbelaar, 2005).

Shape compactness

Shape compactness, adapted from Cockings and Martin (2005) and Haynes *et al.* (2007), was used in an effort to produce more compact (circular, rather than linear) output areas. The overall perimeter squared per area (P2A) was used as a measure of shape compactness. The lower P2A mean values indicate that output shapes are more compact, whereas higher P2A mean values indicate that output areas are less compact.

Number of AZTool iteration runs

Furthermore, number of AZTool iteration runs was also explored. The intention was to determine if increasing the number of iteration runs would improve the statistical characteristics in terms of population targets, social homogeneity and shape compactness. The following number of iteration runs were employed; 10, 20, 50, 100, 500 and 1000 runs.

3.2.5 AZTool output results

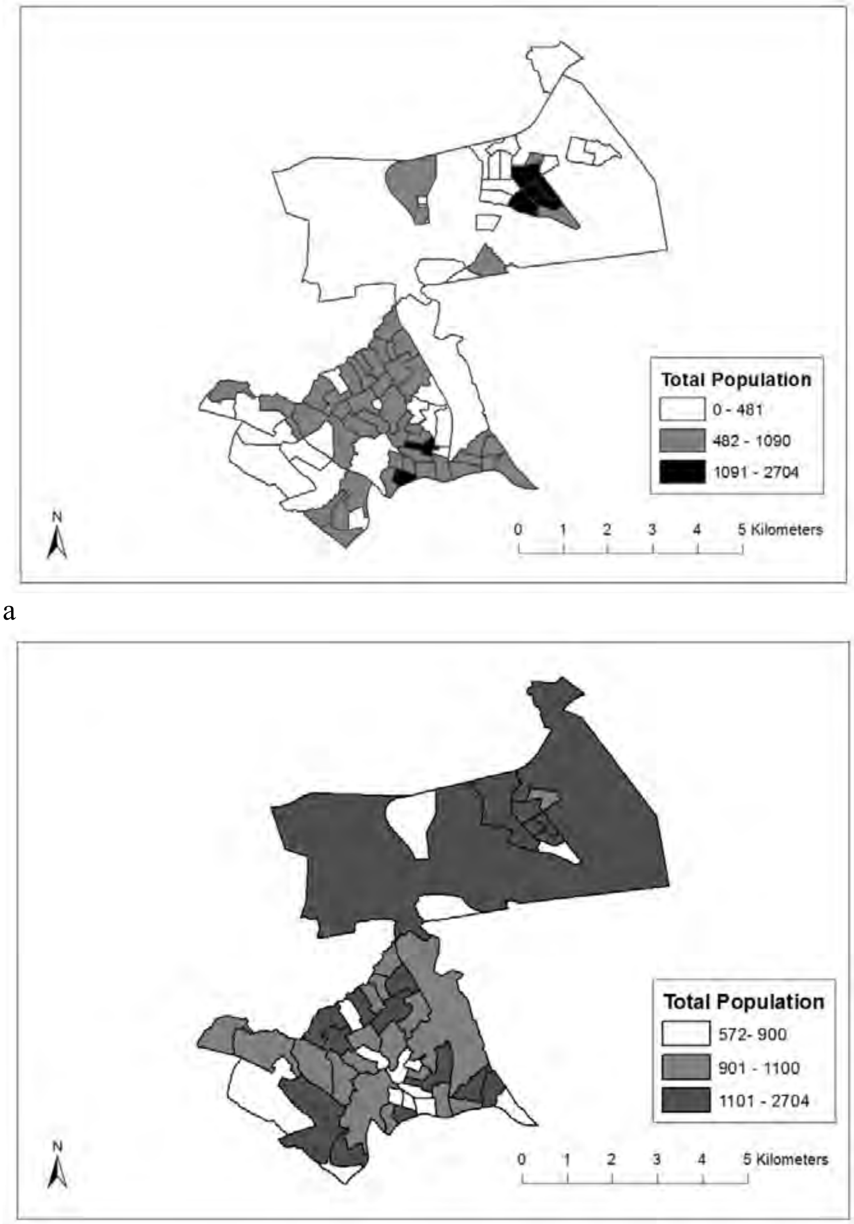
The ArcGIS software was also used to display the results from the AZTool program. The AZTool results were written in a tract composition (.csv) file and contained the Building Block ID and the Tract ID (ID for newly created Output Area) to which it had been assigned. In order to visualize the newly created tracts or output areas, the tract composition (.csv) file was incorporated into the Building Block shapefile using Building Block ID as the field in the table to base the join on. The building block boundaries were then dissolved based on the Tract ID in order to create the Tracts (Output Areas). Further statistical analyses were performed in SPSS (see Section 1.5).

3.3 Results

3.3.1 Statistical characteristics of EAs and output areas in rural areas

Figure 3.2a shows the boundaries of original EAs of Phuthaditjhaba mainplace and indicates an EA which is widely spread on the northern part of the study area. This is typical for rural areas in South Africa. In most cases EAs that are large in terms of size in the rural areas are sparsely populated. Figure 3.2b shows the newly created output areas for the same area of Phuthaditjhaba. Donut EAs or building blocks (areas that are completely surrounding other areas) such as the ones on the north eastern part of the study area are no longer showing in these output areas. These have been combined with other building blocks to form the largest output area in terms of size. However, the most spread EA, largest in terms of coverage or area, is not the highest with regard to total population. This indicates that some of the original EAs which formed this new output area were not as populated as some of their counter parts in the same northern part or in the southern part of the study area.

Table 3.2 highlights the statistical characteristics of the original EAs and the newly created output areas for the rural areas in all four geographical regions. It is important to note that the original EAs were slightly more homogeneous and compact than the newly created output areas at all geographical levels. The latter was further proven by inferential statistics which showed that the difference between P2A means of the AZTool output areas and the original EAs was statistically significant ($p < 0.05$). The confidentiality threshold of 500 was not breached at any of the four geographical levels (mainplace, municipality, district, and provincial). The results show that there is a steady increase in terms of the IAC from the lower geographical level (0.22) to the higher geographical level (0.59), meaning that the degree of homogeneity within-area increased as the geographical level increased. The mean population sizes were also close to the targeted mean with reasonable standard deviations, but the mainplace level had a higher mean value and standard deviation compared to the municipality level.



a

b

Figure 3.2: Population for Phuthaditjhaba mainplace, a) the original building blocks EAs and b) the newly created output areas

The mean shape tended to increase from lower geographical levels to higher geographical levels, indicating that the output areas at higher geographical levels were much less compact in shape than lower geographical levels which had lower mean accompanied by lower standard deviation. However, the difference between P2A means of different geographical levels was not statistically significant ($p > 0.05$).

Table 3.2: Statistical characteristics of EAs and output areas at different geographical levels in rural settings

	Number		Population			Shape		Homogeneity
	of Zones	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
EAs								
Phuthaditjhaba Mainplace	86	0	2704	621	451	25	9	0.25
Maluti-a-Phofung Municipality	747	0	2704	480	313	26	9	0.58
Thabo Mofutsanyane District	1412	0	6196	518	410	26	9	0.66
Free State Province	5182	0	9269	519	454	26	10	0.65
Output Areas								
Phuthaditjhaba Mainplace	49	572	2704	1090	341	27	10	0.22
Maluti-a-Phofung Municipality	349	610	2704	1027	232	32	13	0.5
Thabo Mofutsanyane District	667	581	5292	1087	403	33	13	0.56
Free State Province	2440	547	9269	1101	489	31	12	0.59

3.3.2 Statistical characteristics of EAs and output areas in urban areas

For urban areas, a similar trend was also noticed whereby the EAs were slightly more homogeneous and compact than the newly created output areas at all geographical levels. The difference between P2A means of the AZTool output areas and the original EAs was statistically significant ($p < 0.05$) as it was the case in rural areas. Table 3.3 shows that the IAC increased dramatically from 0.09 at mainplace level to 0.46 for the district/metro level. The provincial level experienced a slight decrease to 0.45. The mean population limit was also adhered to at all geographical levels as it was for the rural areas. For both rural and urban areas, the IAC values at the lower geographical levels were lower than any higher geographical levels. This means that higher geographical levels produced more homogeneous output areas than lower levels. This might be due to the fact that at higher geographical level, there are many building blocks which output areas could be constructed from, whereas at lower geographical levels there are fewer building blocks and hence the AZTool has limited number of options with regard to improving the IAC as well as other constraints. With regard to the compactness of the shape of the output areas, a contradiction to what happened in rural areas was noticed. The lower geographical levels output areas were less compact compared to those of higher geographical areas even though the difference was not statistically significant ($p > 0.05$).

Table 3.3: Statistical characteristics of EAs and output areas at different geographical levels in urban settings

	Number		Population			Shape		Homogeneity
	of Zones	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
EAs								
Pretoria Mainplace	865	0	4625	610	358	24	9	0.11
City of Tshwane District	2115	0	8802	726	538	24	9	0.50
Gauteng Province	13200	0	9627	667	563	24	8	0.50
Output Areas								
Pretoria Mainplace	500	621	5026	1056	320	28	11	0.09
City of Tshwane District	1276	502	8802	1203	514	27	10	0.46
Gauteng Province	7253	501	9627	1214	520	27	9	0.45

3.3.3 Statistical characteristics of Free State and Gauteng province output areas

Table 3.4 shows the comparison of both rural and urban provinces as well as their combined results. The mean population threshold was not breached at either provinces or when the two were combined. The urban province, Gauteng, was outperformed by the rural province with regard to the degree of homogeneity while it outperformed the rural province with regard to compactness of the output shapes. Similar trends were also noticed at other geographical levels. The IAC for all provinces combined was higher than that of both provinces ran separately, while the shape of output areas for combined was more compact than that of the Free State province. The higher degree of homogeneity for all combined (urban and rural provinces), the IAC of 0.62, suggests that the selected variables could be used as good indicators of social homogeneity in creating homogeneous output areas across the entire country.

Table 3.4: Statistical characteristics of Free State and Gauteng province output areas and the two provinces combined

	Output		Population			Shape		Homogeneity
Region	**Areas**	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
Gauteng Province	7253	501	9627	1214	520	27	9	0.45
Free State	2440	547	9269	1101	489	31	12	0.59
All Combined	9773	502	9627	1176	515	28	10	0.62

3.3.4 Optimal number of AZTool runs

Table 3.5 displays the use of different number of AZTool runs at the rural mainplace of Phuthaditjhaba in the Free State. Results show that there was only a slight improvement in the results when the runs were increased up to 1000. In essence, the increasing number of runs did not consistently increase the IAC values, as the IAC ranged from 0.22 to 0.24, and did not consistently decrease the P2A mean values from 10, 20, 50, 100, 500 and 1000 runs.

Table 3.5: Statistical characteristics of Phuthaditjhaba mainplace with different runs

Number	Output	Population				Shape		Homogeneity
of Runs	Areas	Min	Max	Mean	SD	Mean	SD	IAC
10	49	572	2704	1090	341	27	10	0.22
20	47	718	2704	1137	326	28	10	0.21
50	50	572	2704	1068	356	26	8	0.24
100	50	704	2704	1068	335	26	12	0.22
500	49	572	2704	1090	334	27	9	0.23
1000	50	572	2704	1068	334	27	9	0.24

For the urban areas, the same number of runs, as was done in Phuthaditjhaba mainplace, was performed for Pretoria mainplace. Table 3.6 shows that the IAC values remained constant throughout the runs (10 – 1000) at 0.09 while the shape compactness mean only slightly declined from 28 to 27 after 500 runs. It is also worth noting that the higher number of runs came at a price of increased processing time. Therefore, if no tangible improvement with regard to output areas is achieved with higher number of runs it may be wise to stick to a low number of runs, hence, 10 runs were kept in this case. On average, it took approximately three to four hours processing time when Gauteng and Free State provinces were combined with 10 runs. It is important to note that the Free State province on its own as well as lower geographical regions in both rural and urban areas were taking much shorter time to complete. It is anticipated that if this creation of census output areas using the AZTool program is considered for the entire country it might take between 10 and 18 hours. But with increased number of runs (such as 1000 runs) this would take even longer.

Table 3.6: Statistical outputs of Pretoria mainplace with different runs

Number	Output		Population			Shape		Homogeneity
of Runs	**Areas**	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
10	500	621	5026	1056	320	28	11	0.09
20	505	532	4625	1045	312	28	11	0.09
50	504	606	4625	1048	312	28	11	0.09
100	496	606	4625	1064	313	28	11	0.09
500	499	619	4779	1058	315	27	11	0.09
1000	505	619	4625	1045	305	27	11	0.09

3.3.5 Respecting higher geographical levels or regions

In an effort to make sure that output areas were nested with higher geographical regions such as mainplace, municipality and district, the “Region to use” rule was set in the .xml parameter file for the AZTool program. The AZTool could not successfully produce any solutions when any of the higher geographical regions were respected. To overcome this, higher geographical regions were analysed separately and merged at the end to produce an overall output in the Free State province (Table 3.7). The results show that an average IAC score (0.46) for the five districts was below the exact IAC score (0.59) of the Free State province. The importance of census output areas nesting within higher geographical levels is to enable exact statistics to be compiled for geographical areas used for applications such as elections or public resource allocation. However, these higher geographical levels change regularly as population grows, which makes it difficult to keep census output areas nested within them. Hence, some countries such as Australia, England and Wales have removed the requirement for census output areas to be nested within certain higher geographic levels.

Table 3.7: Statistical outputs of merged districts against Free State province single run

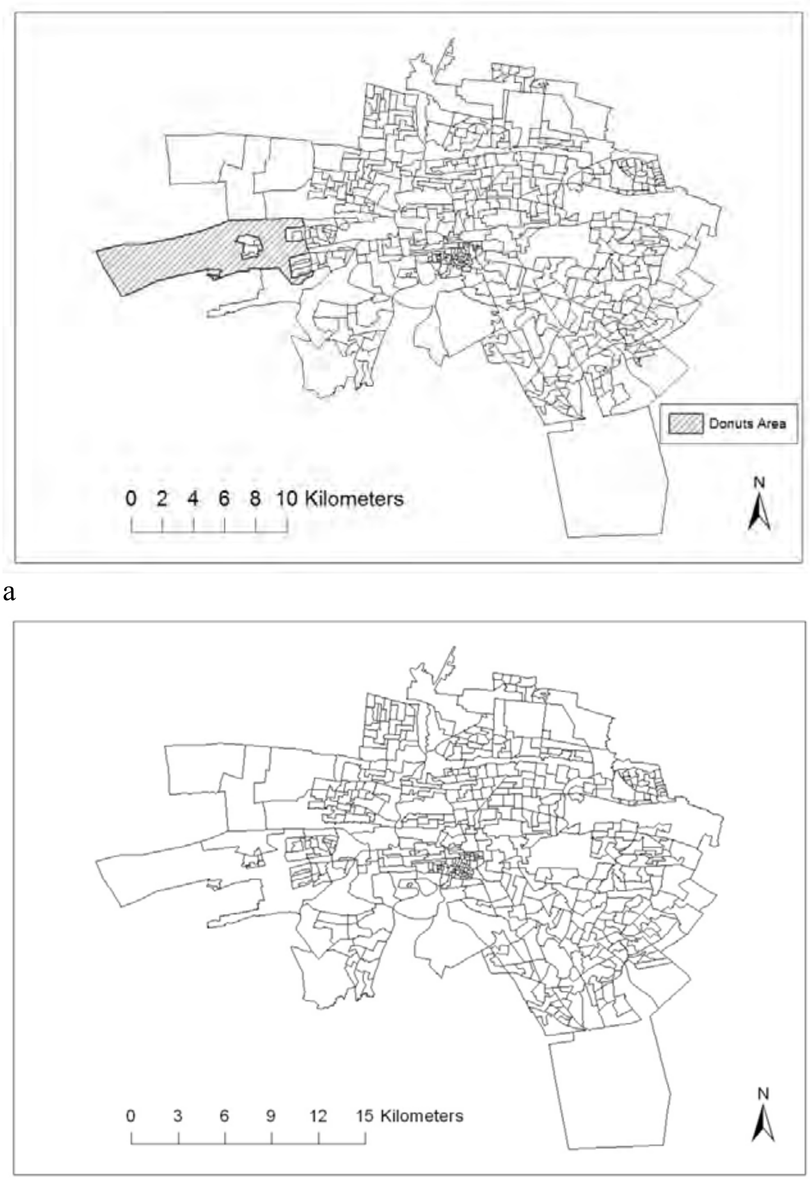
	Output		Population			Shape		Homogeneity
Region	**Areas**	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
Lejweleputswa	558	541	9269	1143	580	30	12	0.40
Motheo	669	517	6252	1093	425	31	12	0.60
Northern Free State	409	573	7116	1116	551	30	11	0.44
Thabo Mofutsanyane	667	581	5292	1087	403	33	13	0.56
Xhariep	123	578	5183	1105	525	33	11	0.34
Merged Districts	**2426**	**517**	**9269**	**1108**	**487**	**N/A**	**N/A**	**0.47**
Free State	**2440**	**547**	**9269**	**1101**	**489**	**31**	**12**	**0.59**

3.3.6 Boundary length constraint

Boundary length is the length of the perimeter of boundaries that is shared between adjacent building blocks. When minimum boundary length was set to 5%, it was found that the shape of the output areas for Pretoria mainplace slightly improved compared to the shape of the original EA building block while the population mean also increased away from the population target mean. The IAC measure did change compared to when the minimum boundary length was ignored. Care should be taken when taking this option as many building blocks might become isolated due to boundary length restrictions. This was the case for Phuthaditjhaba mainplace. The isolated building blocks may be analysed separately and then appended or incorporated to the main output areas at a later stage. Alternately, one could alter the .aat file by manually setting the boundary lengths of these isolated building blocks to negative one (-1) with the implication that their boundary lengths would be ignored in further analysis.

3.3.7 Donuts constraint

Overall, donut areas, which are areas that are completely surrounding other areas, were allowed in all analysis. Figure 3.3a shows some donuts output area (shaded area in the map) in the western part of Pretoria mainplace. Further experiments were undertaken with donuts not allowed in the final output areas for mainplace, municipality and district levels for the two provinces. When comparing donuts allowed against donuts not allowed, results indicate that not allowing donuts did not have nor had little impact on the shape compactness of the output as well as on the IAC score, degree of homogeneity. For example, the western part of Pretoria mainplace did not contain donuts any more, as illustrated in Figure 3.3b. In general, the inclusion of donuts constraint made no real difference in this experiment. This criterion might be of importance in a broader application for avoiding disjointed census output areas, especially when output areas are created for mapping and analysis.



a

b

Figure 3.3: **Newly created output areas for Pretoria mainplace, a) donuts allowed and b) donuts not allowed**

3.4 Discussion

This study illustrates the potential of automated zone design techniques and the potential challenges that may occur when applying such techniques in the creation of optimised output areas in South Africa. Generally, it is important to note that the original building blocks were slightly more homogeneous and compact than the newly created output areas at all geographical levels as well as in both rural and urban settings. The IAC values at the lower geographical levels were lower than those of any higher geographical level in both rural and urban areas. This indicates that higher geographical levels produced more homogeneous output areas than lower geographical levels. One of the reasons for this might be that at a higher geographical level there are many building blocks from which output areas could be constructed, whereas at lower geographical levels there are fewer building blocks. Hence, the AZTool has limited number of options with regard to improving IAC and other constraints. Similarly, in New Zealand, Ralphs and Ang (2009) found that larger areas seemed to be more homogenous with each other compared to smaller areas i.e. evidence of scale effect of MAUP. The lower IAC scores for lower geographical levels (mainplace levels) were also experienced in previous studies at detailed levels (Cockings *et al.*, 2011; 2013; Martin *et al.*, 2013).

When comparing the performance of the AZTool at the two different spatial settings, urban and rural areas, the newly created output areas from the rural areas had higher degrees of homogeneity than their counterparts. However, the urban areas were more compact than the rural areas. Overall, the higher degree of homogeneity for all provinces combined (urban and rural provinces), the IAC of 0.62, suggests that the selected variables could be used as good indicators of social homogeneity in creating homogeneous output areas across South Africa. Generally, the IAC of 0.5 is regarded as a very reasonable degree of homogeneity (Sabel *et al.*, 2013). It is also important to mention that in all experiments that were performed in urban and rural areas at all geographical levels, the confidentiality limit was adhered to.

Ideally, increasing the number of runs tends to improve the AZTool’s solution, as it enables finding better optimal output areas. On the contrary, results from this study did not show reasonable improvement of optimal output areas when different numbers of runs were explored. These results concur with previous studies that were conducted with the AZTool such as Ralphs and Ang (2009) and Sabel *et al.* (2013). When increasing the number of runs they found that increasing number of iteration runs did little significant improvement in the quality of outputs while significantly increasing computing time. Therefore, there is confidence that setting the number of runs to 10 could still produce quality output areas even when expanding the analysis to the entire country.

The donuts constraint did not have an impact on the quality of output areas with regard to shape and degree of homogeneity. Therefore, there was no restriction made to exclude donuts in the final output areas. In order to make sure that output areas nested within higher geographical level or region, the AZTool was set to respect higher geographical regions. Unfortunately the program did not produce any solutions when any of the higher geographical levels were respected. Cockings *et al.* (2011) argued that having to respect a higher geographical region constraint is particularly restrictive and often prevents solutions being found at all. Further investigations need to be performed to see the cause of this in the context of South African geographical areas. To overcome this, higher geographical regions could be analysed separately and merged at the end to produce an overall output even though this might be time consuming for larger samples.

The uniqueness of the approach in this study is that the performance of the AZTool program for both urban and rural areas at different geographical levels or regions was considered. This provided a clear indication as to how the program is likely to perform when the whole of South Africa is analysed. In addition, the current version of the AZTool has potential for application in developing countries, including South Africa, as it does not require Arc Info licence for preparing the contiguity files, .aat and .pat files. However, further consultations with other relevant stakeholders should be taken before output areas from this study could be considered for possible use for any census dissemination as each set of output areas is the product of a set of criteria set by the author. From a policy and practice point of view, it is important to note that this research was a stand-alone project with the aim of influencing policies and practice of government stakeholders such as Stats SA. It is believed that the positive findings from these initial experiments regarding the AZTool applications in the creation of census output areas in South Africa would encourage future possible collaboration between the candidate and the government stakeholders such as Stats SA as well as other South African census data users.

Regarding the limitations, the accessibility of data, at lower geographical levels such as household and EA levels as well as recent the 2011 census data, was not successful. Hence, only the 2001 census EA estimates data was used as building blocks. Globally, there seems to be a challenge with regard to accessing census data at lower geographical levels for research purposes and other purposes such as business and marketing due to confidentiality. Alternatively, a dwelling frame data could have been used, but this was challenging because the data had a lot of missing information or dwellings that were not captured in some areas across the country. The use of household level data would have minimised the flaws carried by administrative data (EAs), which were created for a different purpose, as building blocks into the created output areas. Therefore, caution should be taken when using pre-existing input areas to aggregate them into larger areas, as the flaws that are inherent in the building blocks would be carried over into the output areas as well as possible bias and potential errors associated to the MAUP (Ralphs and Ang, 2009; Drackley *et al.*, 2011; Cockings *et al.*, 2013). Among the barriers of using AZTool for creating census output areas is that respecting a higher geographical region constraint is restrictive and often prevents solutions being found at all, which was also the case in this research. In the real world, there are legal or administrative requirements for census output areas to nest within higher geographical levels to enable exact statistics to be compiled for geographical areas used for either elections or public resource allocation. However, these administrative areas change overtime as population grow, making it difficult to keep census output areas nested within them, hence some countries such as Australia, England and Wales have removed the requirement for census output areas to be nested within certain higher geographic levels.

3.5 Conclusions

The success of this study was measured by the fact that the primary criterion of minimum population threshold of 500 people was kept and not breached throughout all newly created output areas at different geographical levels as well as in both rural and urban areas. In addition, the second most prioritised criterion of homogeneity of output areas showed the IACs of 0.45 for Gauteng province, 0.52 for the Free State, and 0.62 for both provinces combined. These IAC values are encouraging as international studies show that the IAC of 0.5 is regarded a very reasonable degree of homogeneity within output areas. It is important to indicate that the AZTool output areas were significantly ($p < 0.05$) less compact in shape than the original EAs in both rural and urban settings. Based on these findings from different spatial settings as well as different geographical levels, it was concluded that the AZTool software could be used to effectively and objectively create optimized output areas in South Africa. The availability or accessibility of data at lower geographical level, such as household level (or updated dwelling frame data in South Africa), is highly recommended as this would improve developments of robust and optimized output areas using automated zone design techniques.

CHAPTER 4

THE STATISTICAL QUALITIES OF THE AZTOOL CENSUS OUTPUT AREAS

This chapter is based on

Mokhele TA., Ahmed A. and Mutanga O. The statistical qualities of the AZTool census output areas. (*In Preparation*).

Abstract

The statistical qualities of the census output areas are of great importance especially when the purpose of output areas is to understand the statistical properties of the population rather than mapping only. If the purpose of creating census output areas is solely for displaying results in map format, shape compactness of output areas is prioritised. In that case, other statistical characteristics such as population, population mean and social homogeneity are often ignored. This chapter explored the statistical qualities of the newly AZTool generated census output areas using the 2001 census EAs as building blocks in South Africa. The statistical qualities were mainly based on population target mean as a way of controlling population distribution, minimum population threshold, social homogeneity as well as shape compactness. The homogeneity variables that were selected from the 2001 census data were dwelling type and geotype (geography type). The results showed that the AZTool generated output areas substantially out-performed the original EAs and SALs in terms of the minimum population threshold and population distribution statistical qualities. It is worth noting though that the AZTool output areas were less compact and homogeneous than the original EAs in both urban and rural settings. The fact that a minimum population threshold of 500 was respected by AZTool output areas in both rural and urban settings was a huge success from confidentiality point of view. It is therefore concluded that the AZTool could be utilized to produce robust and high-quality optimised output areas for population census dissemination in South Africa.

Keywords: AZTool; Census; Enumeration areas; Output areas; South Africa.

4.1 Introduction

The statistical qualities of the census output areas are of great importance especially when the purpose of output areas is to understand the statistical properties of the population rather than mapping only. In this study, statistical qualities are based on the characteristics of output areas with regard to their shape, social homogeneity and population targets. For instance, if the purpose of creating census output areas is solely for displaying results in map format, shape compactness of output areas is prioritised. In that case, other statistical characteristics such as population, population mean and social homogeneity are often ignored. Therefore this chapter aimed to determine the statistical quality of the newly AZTool generated census output areas using South African EAs as building blocks.

Applications of the AZTool software are well described in the following references (Flowerdew *et al.,* 2008; Ralphs and Ang, 2009; Cockings *et al.*, 2011; 2013; Martin *et al.*, 2013; Sabel *et al.*, 2013). For instance, Cockings *et al.* (2011) employed AZTool to modify the 2001 Census output geographies within six local authority districts in England and Wales in order to make them suitable for the release of contemporary population-related data. This was done such that zones which still meet the design criteria were retained while those that were no longer fit for purpose were split or merged. The use of AZTool for maintenance of an existing system was found to be a more iterative and constrained problem than designing a completely new system; design constraints frequently had to be relaxed and manual intervention was occasionally required. In addition, their findings suggested that it would be easier to resolve under-threshold zones than over-threshold zones.

Martin *et al.* (2013) further explored the application of AZTool for creating workplace zones (WZ) with England and Wales 2001 census microdata. They found that the prototype areas displayed much improved statistical properties, with more uniform sizes of workforce, less extreme values and compliance by design with the specified threshold values. Their results further showed that there were a small number of WZs which could not be automatically resolved by using the parameters evaluated in their study, either because no suitable neighbouring zones were available for merging or their constituent postcodes are inappropriately configured. Their approach was further adopted or incorporated in England and Wales 2011 census output plans.

None of these studies strictly focused on the statistical quality of the created optimised output areas or zones except Ralphs and Ang (2009). They attempted to determine statistical quality of automatically developed geographies by comparing them with existing official geographies in New Zealand. They found that the automatically generated geographies substantially out-performed the existing geographies across almost all of their optimisation criteria. For instance, the automatically created geographies effectively satisfied minimum and target population thresholds, while the population distributions were much narrower in range than the existing reporting geographies. Therefore, aim of this chapter was to determine statistical qualities of the newly AZTool generated output areas in comparison to the original building blocks, EAs in South Africa.

4.2 Methods

Two out of the nine provinces in South Africa were selected for this study (see Section 3.2). These were Free State and Gauteng provinces which were representative of rural and urban areas, respectively. In an effort to get a better picture of the statistical qualities of AZTool output areas at different geographic levels, the district, municipality and mainplace levels were also analysed.

The 2001 census estimates data developed by HSRC (2005) were used to get data at the EA level as the original data was not accessible at EA level from Stats SA. The data for the two provinces that were extracted from these census data include total population, homogeneity variables as well as different spatial level boundaries. The homogeneity variables that were selected from the 2001 census data are dwelling type and geotype (see Table 4.1). The dwelling type also known as housing type is the commonly used variable as proxy for social built environment homogeneity measure (Martin *et al.*, 2001; Ralphs and Ang, 2009) while the geotype (geographic type) has been used as a homogeneity rule for development of SAL which was used to disseminate the 2001 census data in South Africa (Verhoef and Grobbelaar, 2005).

Table 4.1: Homogeneity variables from the 2001 census data

Dwelling Type	Geotype (Geography type)
1=House or brick structure on a separate stand or yard	1=Formal Urban
2=Traditional dwelling/hut/structure made of traditional materials	2=Informal Urban
3=Flat in block of flats;	3=Informal Rural (Tribal areas)
4=Town/cluster/semi-detached house (simplex, duplex, triplex)	4=Formal Rural (Farms)
5 =House/flat/room in back yard	
6=Informal dwelling/shack in back yard	
7=Informal dwelling/shack NOT in back yard, e.g. in an informal/squatter settlement	
8=Room/flatlet not in back yard but on a shared property	
9=Other dwelling	

The EAs from the EA level data of the 2001 census were then used as building blocks for the development of optimised census output areas and were generated using AZTool version 1.0.3 (Cockings *et al.*, 2011). The minimum population threshold, population target, shape and homogeneity criteria were pre-defined in the creation of these optimised output areas. A minimum population of 500 and a population target of 1000 were set. For homogeneity, this study employed the IAC while P2A was used as a measure of shape compactness. For more details on the methodology, see Sections 1.5 and 3.2 of this thesis. Further statistical analyses such as ANOVA and Shapiro-wilk test were performed in SPSS (see Section 1.5 for more details).

4.3 Results

Figure 4.1 highlights the comparison of the original EAs used as building blocks with the AZTool census output areas in Phuthaditjhaba mainplace. Figure 4.1a shows that there was a significant number of areas that had lower than 500 people. The original EAs population distribution also had large population range which means it could not be easy to compare individual areas based on population size. The higher variance further indicates that the original EAs had broader population distribution compared to the optimised AZTool output areas. In addition, the population means of the AZTool output areas were more close to the target mean of 1000 with lower standard deviations compared to the original EAs (Figure 4.1b). This indicates that the output areas had much narrower and tighter population distributions than their counterparts. As it was in Chapter 3, the confidentiality limit of 500 people was also not breached for output areas, which is a success from confidentiality point of view. This was further proven statistically by running Shapiro-wilk test which showed that the population distribution for the AZTool output areas was normal ($p > 0.05$) while for the counter-part it was not normal ($p < 0.05$).

In order to depict the general picture at the urban settings, a similar population distribution figure was displayed for Pretoria mainplace (Figure 4.2). This figure shows that similar trends to those of the rural areas were experienced. The AZTool output areas respected the confidentiality limit and had much tighter population distributions (Figure 4.2b). It is important to highlight that none of these population distributions was normal as the Shapiro wilk test revealed significant ($p < 0.05$) results in both cases.

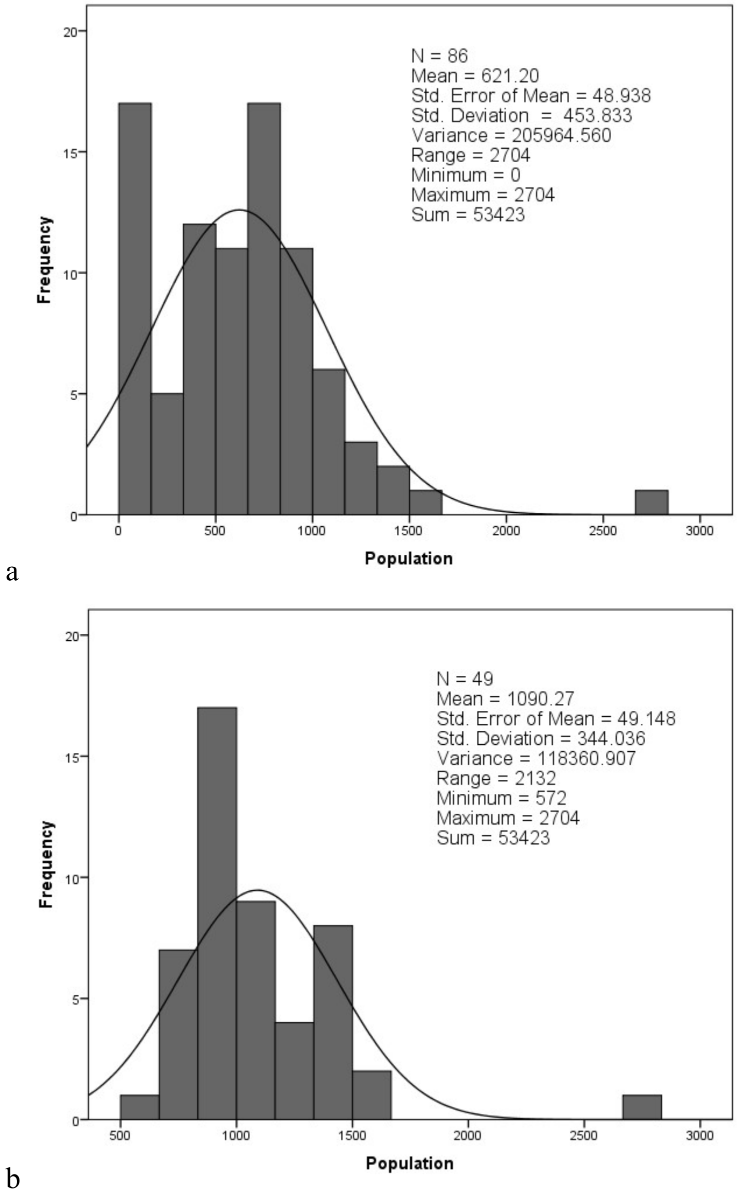


Figure 4.1: Population distribution for a) the original EAs and b) the AZTool census output areas for Phuthaditjhaba mainplace

Furthermore, full results showing statistical characteristics of the AZTool generated output areas and the original EAs in rural and urban areas, were shown in Tables 3.2 and 3.3 in the previous chapter. The results showed that confidentiality was adhered to at all geographical levels in AZTool output areas in both rural and urban areas compared to the original EAs where it was breached at all spatial levels. However these newly created AZTool output areas had higher shape mean at all geographical levels indicating that they were slightly less compact compared to the original EAs in both rural and urban settings.

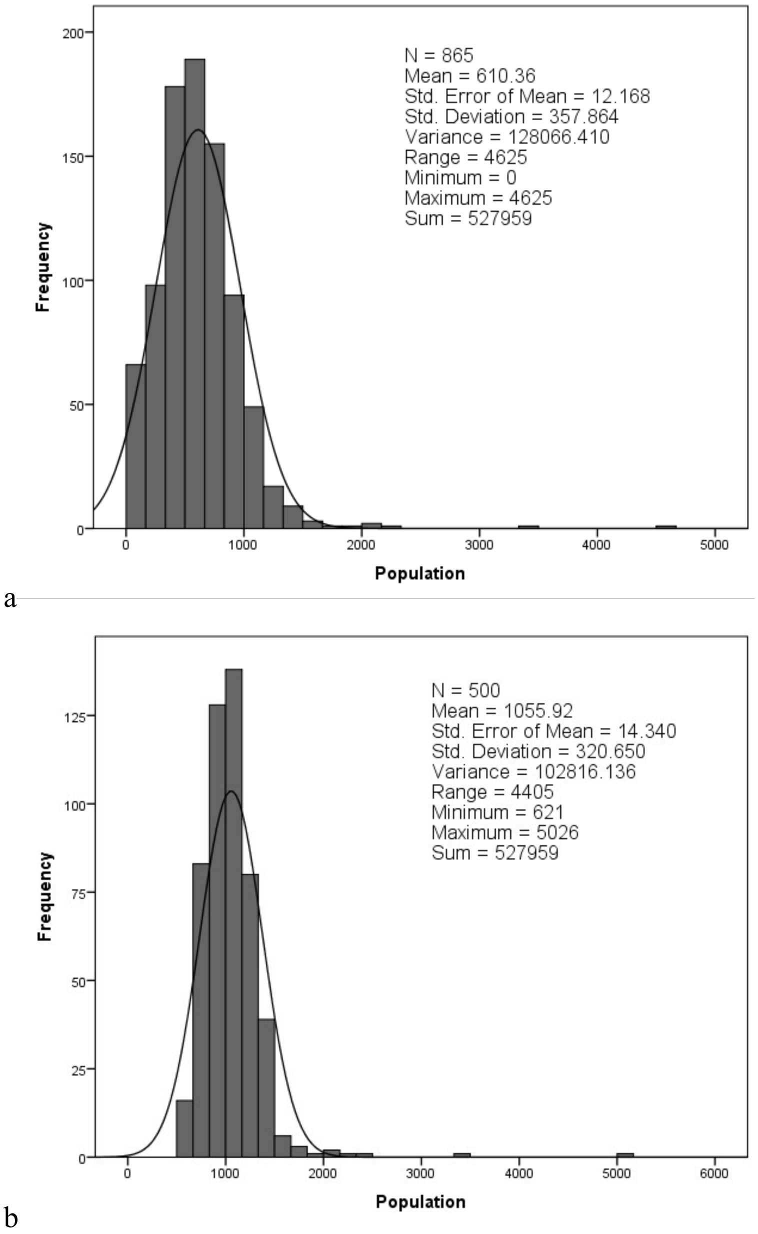


Figure 4.2: Population distribution for a) the original EAs and b) the AZTool census output areas for Pretoria mainplace

As a follow-up testing of number of runs at lower geographical regions in Chapter 3, a further test was performed to see if increased number of AZTool runs would improve statistical characteristics of output areas at the district level in both rural and urban areas. The results showed that increasing number of runs did not improve statistical qualities of optimised output areas in all areas (see Tables 4.2 and 4.3).

Table 4.2: Statistical outputs of Thabo Mofutsanyane district with different runs

Number	Output			Population			Shape			Homogeneity
of Runs	**Areas**	**Min**	**Max**	**Mean**	**SD**	**Score**	**Mean**	**SD**	**Score**	**IAC**
10	667	581	5292	1087	403	113616375	33	13	21695	0.56
20	667	516	5292	1087	404	113921055	32	13	21430	0.56
30	678	587	5364	1070	404	113876601	32	13	21670	0.56
40	676	527	5292	1073	403	113479469	32	12	21389	0.56
50	672	610	5364	1079	401	112337947	32	12	21633	0.56
100	669	581	5292	1084	401	112260839	32	12	21263	0.56
500	663	597	5292	1094	403	113704181	32	13	21364	0.56
1000	676	578	5364	1073	399	111041831	32	12	21593	0.56

Table 4.3: Statistical outputs of Tshwane district with different runs

Number	Output			Population			Shape			Homogeneity
of Runs	**Areas**	**Min**	**Max**	**Mean**	**SD**	**Score**	**Mean**	**SD**	**Score**	**IAC**
10	1276	502	8802	1203	514	389103794	27	10	33940	0.46
20	1273	517	8802	1205	514	390442028	26	10	33623	0.46
30	1262	517	8802	1216	507	383577766	27	10	33682	0.46
40	1267	507	8802	1211	509	385210614	27	10	33750	0.46
50	1265	517	8802	1213	512	388941006	27	11	33581	0.46
100	1271	502	8802	1207	512	387293712	27	10	33732	0.46
500	1273	517	8802	1205	506	379605664	27	10	33799	0.46
1000	1281	502	8802	1198	506	377658462	27	11	33970	0.46

Different weights for homogeneity, population target and shape were also explored to see their statistical effects on the output areas. For instance, when homogeneity weight was set to the weight of 200, 300, 400, 500, and 1000 respectively, the other two (population and shape weights) were left at default weight of 100 and vice versa. Figure 4.3 shows that different shape weights make a substantial improvement on the shape measure of the output areas. There is clear evidence that when the shape (P2A) weight increases, the shape measure decreases, resulting in more compact output areas. For instance, when the shape weight increased from 100 – 1000, the P2A measure decreased from 1340 – 664.

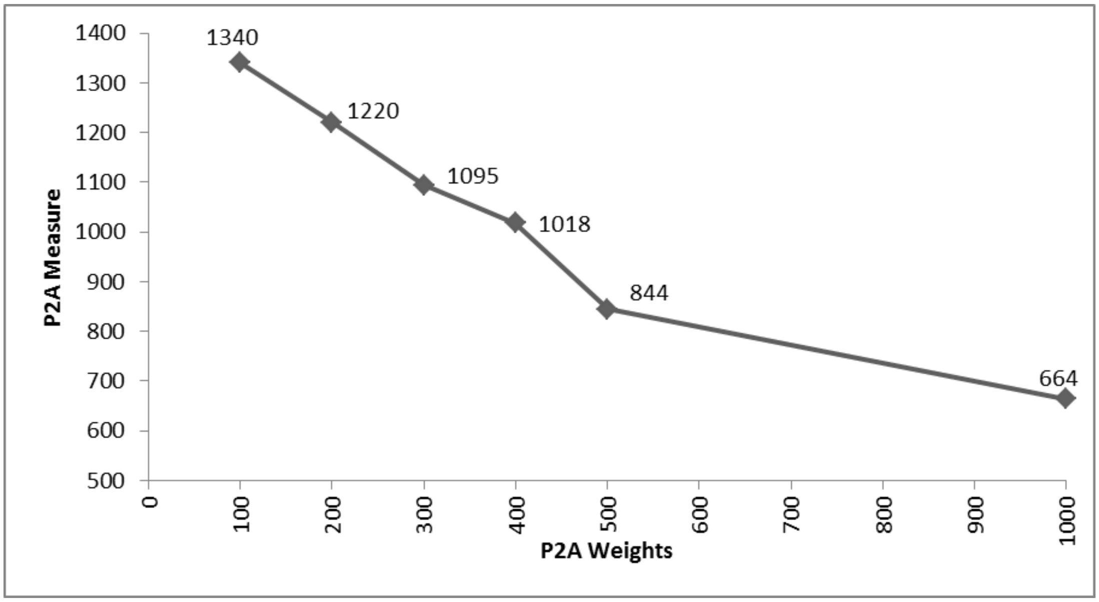


Figure 4.3: Effects of different shape weights on the P2A measure of output areas for Phuthaditjhaba mainplace

Effects of different population weights on the population characteristics of AZTool output areas were also explored for Phuthaditjhaba mainplace. Figure 4.4 highlights that both minimum and maximum population did not change when different population weights were applied. The population target means changed a bit but were also constant after population weights of 500 and1000 were considered.

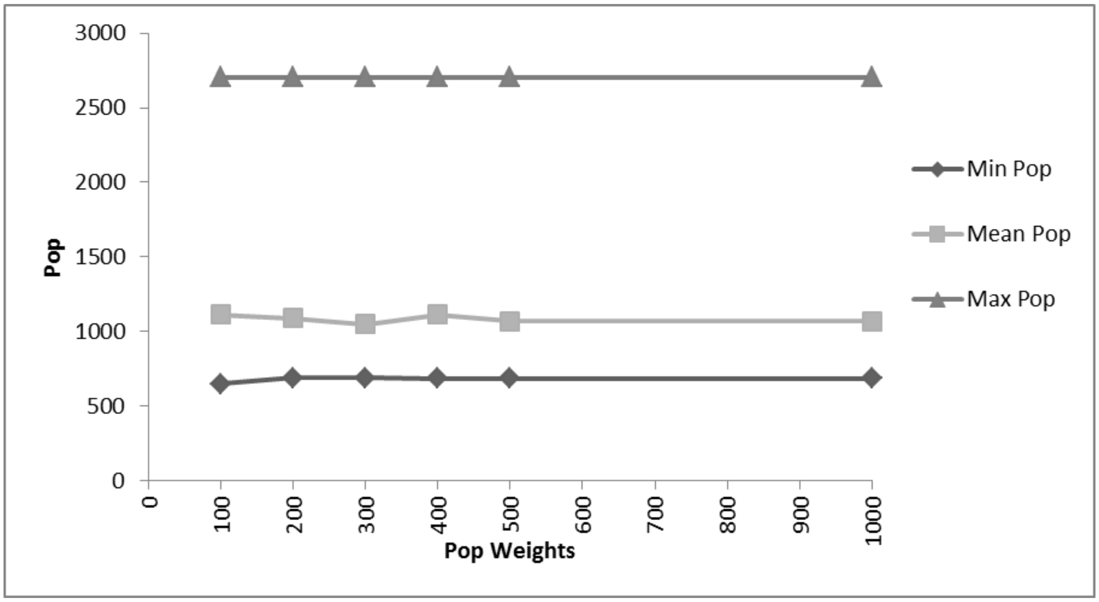


Figure 4.4: Effects of different population weights on the population characteristics of AZTool output areas for Phuthaditjhaba mainplace

Figure 4.5 shows the impact of different shape weights on the AZTool optimised output areas for Phuthaditjhaba mainplace. Clearly, the visual displays highlight that there is improvement from Figure 4.5a (original EAs) to Figure 4.5b (output areas with shape weight of 100) in terms of shape compactness. The shape weights of 500 and 1000 (Figures 5c and d) show even more compact shapes. This indicates that, if the priority to have more compact output areas, especially for mapping, different weights could be applied for Phuthaditjhaba, especially higher weights. It is noteworthy that this application of higher shape weights would come at a compromise of other design criteria such as population target and social homogeneity.

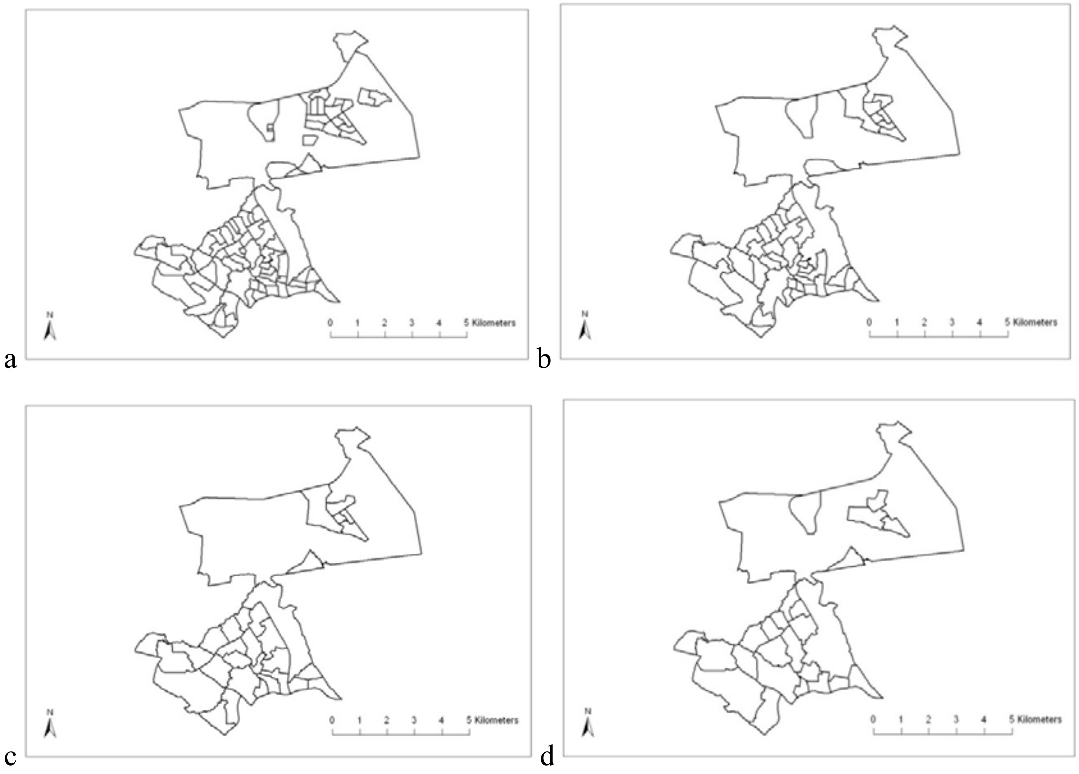


Figure 4.5: Phuthaditjhaba mainplace a) original EAs, b) P2A100weight, c) P2A500weight, and d) P2A1000weight output areas

The 2011 census data was released at SAL level, however there was a significant number of areas that were below the official minimum threshold of 500 people, especially in Free State province whereby almost half (42.2%) of the areas had below 500 people compared to around 27% in Gauteng province. Therefore, the SALs from the 2011 census data were also used as building blocks in an effort to further determine statistical qualities of the AZTool generated output areas. The same criteria set for the generation of output areas using the EAs were employed. The results highlight that the AZTool output areas substantially out-performed the original SALs with regard to confidentiality as none of the output areas were below the 500 minimum population thresholds (Table 4.4). In addition, the population means of the output areas were closer to the set population target of 1000 than the ones of the original SALs at all spatial levels. Hence the output areas had tighter population distribution than the original SALs. The output areas were less compact compared to the SALs at all spatial levels as they had significantly ($p < 0.05$) higher P2A means than their counter-parts. With regard to homogeneity, the SALs produced results at higher level (provincial level) only. Hence only this level could be compared with IAC score for the optimised output areas. Results also highlight that the optimised output areas were less homogeneous than the original SALs.

Table 4.4: Statistical characteristics of the original SALs and the AZTool generated output areas at all levels in the Free State province

	Number	Population				Shape		Homogeneity
	of Zones	Min	Max	Mean	SD	Mean	SD	IAC
SALs								
Phuthaditjhaba	105	42	1065	521	128	25	8	N/A
Maluti-a-Phofung	729	15	1080	460	122	27	9	N/A
Thabo Mofutsanyane	1513	9	1326	486	167	26	8	N/A
Free State	5114	9	5586	536	228	25	9	0.62
Output Areas								
Phuthaditjhaba	51	639	1677	1072	210	33	14	0.18
Maluti-a-Phofung	334	642	1563	1005	166	35	13	0.21
Thabo Mofutsanyane	721	612	1674	1021	188	33	12	0.45
Free State	2596	594	5586	1056	264	31	11	0.55

4.4 Discussion

The results showed that confidentiality was largely adhered to at all geographical levels in AZTool output areas in both rural and urban areas compared to the original EAs where the minimum population was zero at all geographic levels. As indicated earlier in Chapter 1, census data or national statistics have to be released at level where disclosure of personal information of individuals, households, or organisations is avoided by all means, even if other systems such as registers or any administrative datasets are used to collect these data (Valente, 2010; Cockings *et al.*, 2011; Flowerdew, 2011). Furthermore, the AZTool optimised output areas had much narrower and tighter population distributions than the original EAs. This was further proven statistically by Shapiro-wilk test results which showed that the population distribution for the AZTool output areas was normal ($p > 0.05$) whereas for the one of the EAs was not normal ($p < 0.05$). However, these newly created AZTool output areas had higher shape mean at all geographical levels indicating that they were statistically ($p < 0.05$) slightly less compact compared to the original EAs in both rural and urban settings. This shows that a compromise had to be considered at some point (Ralphs and Ang, 2009; Cockings and Martin, 2005; Drackley *et al.*, 2011).

Findings from this study also showed that different shape weights had a great improvement on the visual display of the output areas. This was proven by the fact that when the criterion for the shape was set to carry ten times more weight than population and homogeneity, the shapes of output areas were more circular and less elongated. It is noteworthy that this application of higher shape weights would of course come at a compromise of other design criteria such as population target and social homogeneity. No previous studies which reported on direct impact of different AZTool weights on the statistical qualities of the optimised output areas were found for comparative purposes.

In addition, when the 2011 census data was explored, the results highlighted that the AZTool output areas substantially out-performed the original SALs with regard to confidentiality as none of the output areas were below the 500 minimum population thresholds. The population means of the output areas were more close to the set population target of 1000 than the ones of the original SALs at all spatial levels. Hence the AZTool optimised output areas had tighter population distribution than the original SALs (Ralphs and Ang, 2009; Martin et al., 2013). The output areas were less compact compared to the SALs at all spatial levels. With regard to homogeneity, the SALs produced results at higher level (provincial level) only. Hence only this level could be compared with IAC score for the optimised output areas. Results also showed that the output areas were less homogeneous than the SALs.

4.5 Conclusions

Based on the two different datasets explored in this chapter, it was further proven that the AZTool generated output areas substantially out-performed the original EAs and the SALs in terms of minimum population threshold and population distribution statistical qualities. To substantiate this, Shapiro-wilk test results showed that the population distribution for the AZTool output areas was normal ($p > 0.05$) whereas for the one of the EAs was not normal ($p < 0.05$). However, the AZTool output areas were less compact and homogeneous than the original EAs in both urban and rural settings. The fact that confidentiality limit of 500 persons was respected by AZTool output areas in both rural and urban settings was a huge success from a confidentiality point of view (similar results were found in Chapter 3). Results further showed that different shape weights had a great improvement on the visual display of the AZTool output areas. For instance, when the criterion for the shape was set to carry ten times more weight than population and homogeneity, the shapes of output areas were more circular and less elongated. It was therefore concluded that the AZTool could be utilized to produce robust and high-quality optimised output areas for population census disseminations in South Africa. However, a compromise had to be taken when setting the criterion based on the purpose the output areas would be utilised for.

CHAPTER 5

EFFECTS OF DIFFERENT BUILDING BLOCKS DESIGNS ON THE STATISTICAL CHARACTERISTICS OF AZTOOL OUTPUT AREAS

This chapter is based on

Mokhele TA., Mutanga O. and Ahmed A. Effects of different building blocks designs on the statistical characteristics of AZTool output areas. *International Journal of Geographical Information Science. (Submitted).*

Abstract

It is common practice that prior to any census, the country usually gets demarcated into small geographic units called census enumeration areas, districts or blocks. In most countries, these small geographic units are also used for census dissemination. In cases where they are not used for census release, they are normally used as building blocks for developing output areas or they are aggregated to higher spatial levels in an effort to preserve privacy or confidentiality. Buildings blocks are therefore, of significant importance towards results that could be drawn from either aggregated higher level or from output areas developed using these small geographic areas. This study aimed at evaluating the effects of different building blocks on the statistical characteristics of output areas generated using the AZTool computer program. Different spatial layers (EAs, SALs and SubPlaces) from the 2001 census data were used as building blocks for the generation of census output areas with AZTool program in both rural (Free State province) and urban (Gauteng province) areas of South Africa. One way-ANOVA was also performed to determine statistical significance of the AZTool results. Results from this study show that the AZTool output areas generated from smaller areas (EAs and SALs) tend to be more homogeneous than the ones generated from larger areas (SubPlaces) when using dwelling type and geotype as homogeneity variables. The output areas from smaller areas also had narrower population distribution and more compact shapes than their counter-parts. In addition, the AZTool optimised output areas from the smaller areas allowed a clear distinction of the scale effects than output areas from larger areas. It was concluded that indeed different building blocks did have an impact on the statistical qualities of the AZTool optimised output areas in both rural and urban settings in South Africa.

Keywords: AZTool; Building blocks; Enumeration areas; Output areas; Scale effects; Small Area Layers; SubPlaces.

5.1 Introduction

It is common practice around the world that before any census is conducted the country usually gets demarcated into census EAs. These small areas are normally designed in such a way that they are of a size enough to be covered by one census enumerator within the census period. In South Africa, these small areas (EAs) normally contain from 100 to 250 households (Stats SA, 2003; Stats SA and HSRC, 2007; Verhoef and Grobbelaar, 2005). As indicated in Chapter 2, the criteria for the design of EAs are that: firstly, they should not overlap; secondly, they should be compact without pockets or disjointed sections and should cover the entire country; thirdly, they should have boundaries that could be identified on the ground; and last but not the least, they should be of approximately equal population size to enable an enumerator to cover each one in the allocated census period (Stats SA and HSRC, 2007). Before 1996, the boundaries for EAs were hand-drawn, which is traditional demarcation (Laldaparsad, 2007). The 1996 census represented a transition from traditional demarcation and mapping methods towards an electronic geographic database. For the 2001 census, GIS technology was used to draw EAs (80787 EAs) and for map production. For instance, about 80% of the 2001 EA demarcation was done in the office on a GIS using photography and digital topographical maps (Stats SA, 2003; Laldaparsad, 2007). For the other 20%, field inspection and other alternatives were considered (Stats SA, 2003; Laldaparsad, 2007). For the 2011 census, there were 103576 newly demarcated EAs across the country.

The 2001 census was not released at EA level in an effort to preserve confidentiality of individuals, but instead data was released at SubPlace level. These SubPlaces were too large for most census data users and did not have tighter and narrower population distribution hence comparability of areas with respect to population size was a challenge. A new spatial layer, the SAL, was therefore created using the non-zone design approach in 2005 for release of the 2001 census data at lower level. A similar non-zone design approach was also employed in the creation of SAL for the 2011 census data.

Some countries such as Australia have moved towards having nationally consistent small areas, mesh blocks, which would be used as a stable basis for their output zones and systems for many years to come (Cockings *et al.*, 2013). This move perhaps is worth an investigation in South Africa as this would allow trend analysis and comparisons between different censuses at smaller areas. Generally, geographic shape compactness is of concern with regard to urban morphology, political districting, and accuracy of enumeration unit values (MacEachren, 1985). Zones or areas with compact shapes whose boundaries follow recognisable features on the ground are often desirable for mapping purposes whereas homogeneity of population size is often preferable for statistical analysis (Cockings *et al.*, 2013). Social homogeneity of zones or areas on the other hand could also be of high importance as this could be used as an indication of where resource allocation or service deliveries should be prioritized by governments and Non-Government Organisations (NGOs). However, practical considerations often out-compete more conceptual aspects when designing these small areas or building blocks design (Cockings *et al.*, 2013).

In most countries, these small geographic units are also used for census dissemination. In cases where they are not used for census release, they are normally used as building blocks for developing output areas or zones or they are aggregated to higher spatial levels (Cockings *et al.*, 2013). This aggregation is often done on the basis of geographical location and usually data are made available at two or more spatial levels (Flowerdew, 2011). It is noteworthy that small areas or building blocks would always be of high importance for the dissemination of national population statistics due to confidentiality issues even if census is replaced by other systems such as registers or any administrative datasets, like in Denmark and Finland (Valente, 2010). Buildings blocks are therefore of significant importance towards results that could be drawn from either aggregated level or from output areas developed using these small geographic areas.

The fact that census data is collected for individual households but is usually released at higher levels to preserve confidentiality raises some concerns. This results in a problem called the MAUP originally discovered by Gehlke and Biehl (1934). As indicated earlier in Chapter 2, the MAUP has two components, the scale effect and the zonation (Openshaw, 1977; 1984; Reynolds, 1998; Ratcli¨e and McCullagh, 1999; Kitchin and Tate, 2000; Heywood *et al.*, 2002; Duque *et al.*, 2007; Dumedah *et al.*, 2008; Ralphs and Ang, 2009; Flowerdew, 2011). These two effects occur due to the fact that spatial processes generating the observed data may exist at scales and for particular areal units that may be reflected more or less accurately by the boundaries that are used (Manley *et al.*, 2006).

Cockings *et al.* (2013) evaluated the influence of two sets of building blocks (street blocks and postcodes) on output zone characteristics using six local authorities in England and Wales. Their findings indicated that postcodes were more effective building blocks than street blocks as they provided more uniform population and household sizes. On the other hand street blocks were found to produce more compact output zones with greater internal homogeneity of tenure and accommodation type. They also found that the scale effect of the modifiable areal unit problem and the specific geographical patterning of variables were important factors when designing building blocks. Therefore, this chapter was aimed at evaluating the effects of different building blocks on the statistical characteristics of the AZTool optimised census output areas in South Africa.

5.2 Methods

Two provinces were selected for this study; Free State and Gauteng, which were representative of rural and urban areas, respectively (see Section 3.2). In each province, different spatial or geography levels such as district, municipality and mainplace were selected for subsequent analysis. There were no provincial boundary changes for Free State province in 2011 and its total population did not change substantially between 2001 and 2011 hence comparisons of the two censuses data could be undertaken where necessary for this study area. As both rural and urban settings were represented, findings from these study areas are likely to apply in many other parts of South Africa.

The original 2001 census data from Stats SA at SAL and SubPlace levels for the two provinces were extracted. It is noteworthy to mention that the 2001 census estimates (HSRC, 2005) were used for EA-level data as this data was not accessible from Stats SA. The 2011 census data at SAL level was also extracted for Free State province to allow comparison with the 2001 census data as this province did not exhibit a significant population change between 2001 and 2011 as well as its boundaries which did not change. The extracts from the data included total population, homogeneity variables (dwelling type and geotype) as well as spatial levels related information. Therefore, different spatial layers (2001 EAs, 2001 Subplaces, 2001 and 2011 SALs) were used as building blocks for the generation of census output areas in order to determine the impact of building blocks of output areas.

These output areas were generated using the AZTool version 1.0.3 (Cockings *et al.*, 2011) with pre-defined design criteria such as minimum population threshold, population target, shape and homogeneity. All the AZTool output areas were generated using different building blocks with a population threshold of 500 (as practised by Stats SA) and a population target of 1000. The IAC was used to measure the degree of homogeneity within the AZTool output areas (Tranmer and Steel, 1998; 2001; Martin *et al.*, 2001; Flowerdew, 2011). For instance, higher IAC values indicate a higher degree of homogeneity within-area and a higher degree of heterogeneity between areas (Tranmer and Steel, 1998; Martin *et al.*, 2001; Cockings *et al.*, 2013). The P2A (Cockings and Martin, 2005; Haynes *et al.*, 2007) was employed as a measure of shape compactness. Briefly, low P2A mean values indicate more compact shapes (see Sections 1.5 and 3.2 for more details). SPSS was also employed for further statistical analysis (see Section 1.5 in Chapter 1).

5.3 Results

5.3.1 Effect of building blocks on statistical qualities of output areas in rural settings

Table 5.1 summarises characteristics of output areas developed using three different building blocks (EAs, SALs and SubPlaces) at the rural settings. The confidentiality limit of 500 persons was adhered to for all output areas from the three different building blocks (also see Chapters 3 and 4). The AZTool output areas from the EAs had slightly higher population means and lower standard deviations than the ones developed with SALs as building blocks. This means that the SALs built output areas were slightly tighter than the ones created from the EAs with regard to population distribution. The output areas from the SubPlaces on the other hand had higher population means and higher standard deviations. With regard to shape compactness, the lower P2A mean values indicated that output shapes were more compact whereas higher P2A mean values indicated that output areas were less compact. The P2A mean values for output areas from the EAs and the SALs were almost similar but the latter had slightly higher standard deviations at all levels. The output areas from the SubPlaces had higher P2A means and higher standard deviations than the ones generated from the EAs and the SALs at all spatial levels. Clearly, this shows that the output areas created using the EAs and the SALs were significantly ($p < 0.05$) more compact than those developed using the SubPlaces as building blocks. The post-hoc test results showed that P2A means for output areas from both the EAs and the SALs were not significantly different ($p > 0.05$). The results further indicated that the difference between P2A means of those generated from the SubPlaces and the EAs and the difference between P2A means of those created from the SubPlaces and the SALs was statistically significant ($p < 0.05$).

For homogeneity, only the AZTool optimised output areas from the EAs and SubPlaces yielded reasonable results. The SALs ones did not have enough homogeneity variables hence the IAC score produced not a number i.e. the SALs data did not have the dwelling type variable. At lower levels, the IAC scores for the output areas from the EAs were lower than those from the SubPlaces while at higher spatial levels the opposite was the case. The IAC score for the output areas developed using the EAs as building blocks was 0.59 while that of using the SubPlaces was 0.51 at provincial level. These statistics indicate that the output areas from the EAs were less homogeneous than those from the SubPlaces at lower levels (mainplace and municipality) while EAs output areas were more homogeneous than the SubPlaces ones at the provincial level. The two sets of output areas were homogeneously the same at the district level as they both had IAC score of 0.56.

Table 5.1: Statistical characteristics of output areas from the EAs, the SALs and the SubPlaces for Free State

	Number		**Population**			**Shape**		**Homogeneity**
	of Zones	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
EA Output Areas								
Phuthaditjhaba	48	649	2704	1113	346	28	9	0.21
Maluti-a-Phofung	349	610	2704	1027	232	32	13	0.50
Thabo Mofutsanyane	667	581	5292	1087	403	33	13	0.56
Free State	2440	547	9269	1101	489	31	12	0.59
SAL Output Areas								
Phuthaditjhaba	53	701	1602	1003	229	26	14	N/A
Maluti-a-Phofung	361	646	2071	999	230	32	14	N/A
Thabo Mofutsanyane	726	615	6701	1000	353	32	15	N/A
Free State	2707	619	7747	1000	342	31	13	N/A
SubPlace Output Areas								
Phuthaditjhaba	10	722	10507	5318	3902	47	32	0.32
Maluti-a-Phofung	53	516	22496	6807	4657	41	19	0.55
Thabo Mofutsanyane	112	631	26411	6482	4475	39	18	0.56
Free State	352	530	93782	7690	7890	37	21	0.51

5.2.3 Effect of building blocks on statistical qualities of output areas in urban settings

Table 5.2 presents the statistical qualities from similar analysis but this time for urban areas. For the mean population target, similar trends were noticed as the output areas from the SALs were having lower means (almost similar to the target mean) and standard deviations than those developed using the EAs. In addition, the AZTool output areas from the SubPlaces also had higher population means and higher standard deviations than those developed from the EAs and the SALs as in rural areas. Similar trends as to those in rural areas were also seen for the optimised output areas from all the three different building blocks from the shape compactness of the shape point of view. However, the P2A mean values of the output areas from the SubPlaces were not as higher as they were in rural areas but they were statistically different from their counter-parts with one-way ANOVA revealing p-value less than 0.05 (p = 0.006).

In contrast to rural areas, the IAC scores for the output areas from the EAs were higher than those of the output areas from the SubPlaces at lower levels. At higher level, provincial level, the IAC score for the output areas developed using the EAs as building blocks was still higher than those from the SubPlaces. This highlights that for the urban areas, the automated zone design output areas generated using the EAs as building blocks were more homogeneous than their counter-parts at all spatial levels.

Table 5.2: Statistical characteristics of output areas from the EAs, SALs and SubPlaces for Gauteng

	Number		**Population**			**Shape**		**Homogeneity**
	of Zones	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
EA Output Areas								
Pretoria	500	621	5026	1056	320	28	11	0.09
City of Tshwane	1276	502	8802	1203	514	27	10	0.46
Gauteng	7253	501	9627	1214	520	27	9	0.45
SAL Output Areas								
Pretoria	525	640	4227	1001	281	27	12	N/A
City of Tshwane	1527	587	8092	1000	429	26	11	N/A
Gauteng	8837	586	8400	1000	365	26	11	N/A
SubPlace Output Areas								
Pretoria	74	623	26773	7100	5425	29	12	0.06
City of Tshwane	135	577	82002	11311	11930	29	14	0.42
Gauteng	967	523	131662	9139	10052	31	17	0.40

5.3.3 Effect of building blocks from different censuses in rural settings

As the 2001 SALs did not have enough homogeneity variables, the 2011 SALs (from the 2011 census data) were used as building blocks to determine effects of the SALs on the statistical qualities of AZTool output areas in terms of degree of homogeneity. Only SAL data for Free State province was extracted from the 2011 census data as this province did not change boundaries from 2001 and as its total population had only slight increase while Gauteng province had changes on its provincial boundaries and its total population increased substantially. It is noteworthy to mention that the boundaries of all other lower spatial levels changed from the 2001 census, hence only the provincial level of the 2011 census results could be compared with the 2001 census ones. Although both the 2001 EAs and 2011 SALs data had the dwelling type and the geotype homogeneity variables, there were slight differences in terms of their categories, that is, the 2001 EAs had 9 fields for dwelling type and 4 for geotype while the 2011 SALs data had 12 fields (cluster house, townhouse and caravan as extra fields) for dwelling type and 3 for geotype (Only Urban, no more Formal and Informal Urban).

The 2001 EAs optimised output areas had slightly higher population mean of 1101 and higher standard deviation of 489 compared with 1056 and 264 of the output areas generated using the 2011 SALs as building blocks (Table 5.3). This highlights that the output areas from the 2001 EAs were slightly less tight than their counter-parts with regard to population distribution. However, it should be mentioned that the population means for the optimised output areas from both sets of building blocks were close to the target mean of 1000 people which was set on the design criteria. The output areas from the two sets of building blocks were similar with regard to shape compactness. The AZTool output areas developed from the 2001 EAs were more homogeneous than the ones created using the 2011 SALs as building blocks with IAC score of 0.59 and 0.55 respectively. Table 3 further indicates that there is continuous decreasing trend with regard to IAC scores for output areas from smaller areas to larger areas as output areas from the SubPlaces recorded the lowest IAC value of 0.51.

Table 5.3: Statistical characteristics of output areas using 2001 EAs, 2011 SALs and 2001 SubPlaces as building blocks for the Free State province

Free State	Number	Population				Shape		Homogeneity
Province	of Zones	Min	Max	Mean	SD	Mean	SD	IAC
2001 EA Output Areas	2440	547	9269	1101	489	31	12	0.59
2011 SAL Output Areas	2596	594	5586	1056	264	31	11	0.55
2001 SubPlaces Output Areas	352	530	93782	7690	7890	37	21	0.51

The fact that almost half (42.7%) of the SALs for the 2011 census breached the confidentiality limit of 500 people in Free State province, prompts further generation of census output areas in South Africa, if confidentiality is taken seriously. The argument is that there is a pressing need for the creation of the 2011 census output areas which truly respect confidentiality limit as much as possible. The AZTool program was then used to explore effects of different homogeneity variable pairs on statistical qualities of census output areas using the 2011 SALs as building blocks in Free State province at all spatial levels. The homogeneity variable pairs were: dwelling type and geotype; tenure type and geotype; dwelling type and tenure type; and all three homogeneity variables together.

Results highlighted that statistical qualities of AZTool output areas developed using different combinations of homogeneity variable pairs were slightly similar in terms of population means and shape compactness. The statistical characteristics differed when it comes to degree of homogeneity. Figure 5.1 shows that tenure type and geotype homogeneity variable pair had higher IAC scores than all the other variable pairs at all spatial levels. The dwelling type and geotype homogeneity variable pair became second, and then all three homogeneity variable pair and lastly dwelling type and tenure type. The dwelling type and tenure type homogeneity variable pair had very low IAC scores; hence if the social homogeneity is one of the design criteria, this pair could not be used. For example, at provincial level, the pair resulted in output areas that were almost three times less homogeneous and two times less homogeneous than the ones from tenure type and geotype and all three homogeneity variable pairs, respectively.

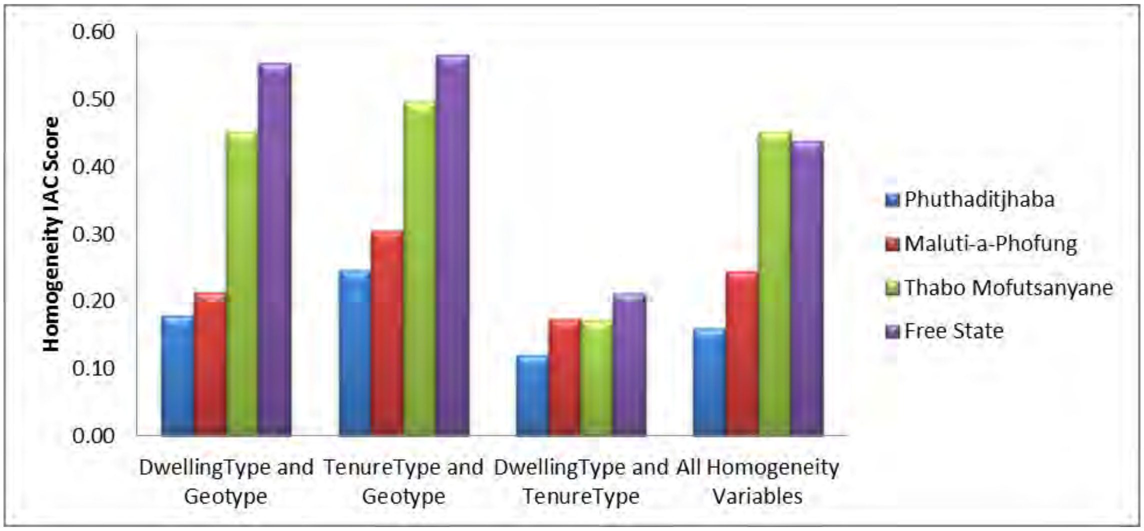


Figure 5.1: Different homogeneity variable pairs' IAC scores for AZTool output areas in Free State

5.4 Discussion

Findings of this study highlight that different building blocks do have an impact on the statistical qualities of the AZTool optimised output areas. Generally, all output areas from the three different building blocks adhered to the confidential limit of 500 persons; this is a huge success from personal privacy perspective. When the EAs and the SALs were used as building blocks in all study areas, statistics showed that output areas from the EAs had slightly higher population means and lower standard deviations than the ones from the SALs. However, the means from the two sets of optimised output areas were close to user-defined population target mean. Clearly, this highlights that the EAs output areas had slightly broader population distributions than their counter-parts. This might be due to the fact that EAs had maximum population of 9269 and low population average of 519 compared to 6701 and 782, respectively, of the SALs. This shows that the AZTool program had to do more effort to bring the mean value of 519 to the target mean of 1000 than it was for 782 to 1000. Unsurprisingly, the output areas from the SubPlaces had higher population means and standard deviations than the ones from the two sets of buildings blocks at all levels in both rural and urban settings. This was expected as the SubPlaces are much bigger in size than the two sets of building blocks and the two sets nest within the these SubPlaces in the South African geography hierarchy.

With regard to shape compactness, the optimised output areas generated from the EAs and the SALs were almost similar at all study areas. The output areas from the SubPlaces were the less compact compared to the ones from EAs and the SALs at all spatial levels in both rural and urban settings. However, the P2A mean values and standard deviations of the output areas from the SubPlaces in urban areas were not as high as they were in rural areas; in fact they were close to the mean values and standard deviations of output areas from the EAs and the SALs. Therefore the effects of different building blocks on the shape characteristics of the AZTool output areas tend to be noticed more in the rural areas, especially between lower level building blocks (EAs and SALs) and higher level building blocks (SubPlaces). Findings from all the three different building blocks further showed that the AZTool optimised output areas from urban areas were more compact than their counter-parts at all levels of geography. In support of these findings, Cockings *et al.* (2013) discovered that output areas in rural areas (Isle of Anglesey) were less compact than those in the urban areas (Camden, Manchester).

For degree of homogeneity, only the AZTool optimised output areas from the EAs and SubPlaces yielded reasonable results. The SALs ones did not have enough homogeneity variables hence did not produce reasonable output results. Therefore 2011 SALs were explored, but only for the Free State province as it did not change its provincial boundaries from 2001. In general, the output areas created using the EAs as building blocks were more homogeneous than those created from the SubPlaces in both rural and urban settings. Few exceptions were found in rural areas where output areas from the EAs at lower geographic levels (Mainplace and municipality) were less homogenous than the ones from the SubPlaces.

In terms of homogeneity at the SAL level, findings from the Free State province (only at provincial level) showed that the AZTool output areas created using 2011 SALs as building blocks were less homogeneous than the ones from EAs but more homogenous than those from the SubPlaces. This is indicative that the AZTool output areas generated from smaller areas tend be more homogeneous than the ones generated from larger areas when using dwelling type and geotype as homogeneity variables. Similarly, it was found that there was a tendency for smaller areas to capture more between-neighbourhood variations than larger areas, hence clustering appeared to be most marked at the very local scale (Haynes *et al.*, 2008).

In addition to measuring the degree of homogeneity, IAC scores could also be used as an assessment of magnitude of the scale effect because the IAC scores are adjusted for population size (Manley *et al.*, 2006; Flowerdew, 2011). Generally, the higher IAC scores indicate the higher scale effects. The higher IAC scores produced by output areas from smaller areas indicate that scale effects are clearly identified when smaller areas are used as building blocks than when larger areas are considered. This also supports arguments by previous studies such as (Openshaw, 1984; Cockings *et al.*, 2013) that the scale effect of the MAUP is generally greater than the zonation effect.

In general, the AZTool output areas from the SubPlaces had higher population means and higher standard deviations than those developed from the EAs and the SALs at all levels in both rural and urban areas. This shows that the SubPlaces are not ideal building blocks from user's perspective as comparisons of individual areas in terms of population size is not possible. In addition, the output areas from the SubPlaces were less compact in shape and less homogenous than the output areas from their counter-parts.

When looking at different combinations of homogeneity variable pairs, it was found that tenure type and geotype homogeneity variable pair and the dwelling type and geotype homogeneity variable pair made it more possible to identify scale effects than all three homogeneity variable pair and the dwelling type and tenure type. The dwelling type and tenure type homogeneity variable pair had very low IAC scores which indicated that output areas were less heterogeneous between each other hence low scale effect.

Among limitations to this study was the accessibility of data at lower levels from Stats SA. The accessibility of census data at household level could have allowed the exploration of other building blocks design such as grid squares which were found to minimize the effect of MAUP in France by Sabel *et al.* (2013). In addition, the 2011 census data at the SAL level excluded zero-populated areas; therefore this resulted in 15 isolated building blocks being picked by the AZTool program in Free State province. They were excluded for further analysis as the AZTool works with contiguous building blocks. Even though these isolated building blocks constituted only 0.15% of the total population of Free State province, they might have some slight contribution on the statistical characteristics of the AZTool output areas generated using the 2011 SALs as building blocks.

5.5 Conclusions

It was concluded that based on results from this study, different building blocks did have an impact on the statistical qualities of the AZTool optimised output areas in both rural and urban settings in South Africa. Although the output areas from the smaller areas (EAs and SALs) were almost similar, they differed slightly. The output areas generated from the EAs were slightly more homogeneous while the ones from the SALs had slightly narrower population distributions. Therefore, between these two building blocks, a choice would depend on the user needs. For instance, for better allocation of resources and prioritising targeting of the population that is the most in need, the EAs would be better suited as building blocks than the SALS. However, the latter would be ideal for statistical analysis as individual areas are comparable in terms of population size. The SubPlaces on the other hand are not ideal to be used as building blocks as they produced output areas with higher population means and higher standard deviations than those developed from the EAs and the SALs at all levels in both rural and urban areas. In addition, the output areas generated from the

SubPlaces were less compact in shape and less homogenous than the output areas from their counter-parts. From these findings, it was also concluded that that the AZTool output areas generated from smaller areas tend be more homogeneous than the ones generated from larger areas when using dwelling type and geotype as homogeneity variables. In addition, the AZTool optimised output areas from the smaller areas allowed a clear distinction of the scale effects than output areas from larger areas. The accessibility of census data at lower levels such as household would have allowed exploration of other building blocks such as grid squares. Therefore, it is recommended that such data should be made accessible even if it is under secure condition for future research.

CHAPTER 6

COMPARISON OF THE AZTOOL OUTPUT AREAS WITH EXISTING OFFICIAL CENSUS DISSEMINATION AREAS IN SOUTH AFRICA

This chapter is based on

Mokhele TA., Mutanga O. and Ahmed A. Comparison of AZTool census output areas with existing output areas in South Africa. *International Journal of Geographical Information Science. (Submitted).*

Abstract

South Africa is one of the few countries that have stopped using the same EAs for census enumeration and dissemination. The advantage of this change is that confidentiality issue could be addressed for census dissemination as the design of geographic unit collection is mainly covered by one enumerator. The objective of this chapter was to evaluate the performance of automated zone design output areas against non-zone design developed geographies using the 2001 census data, and 2011 census to some extent, as the main input. The comparison of the AZTool census output areas with the SALs and SubPlaces based on confidentiality limit, population distribution, and degree of homogeneity as well as shape compactness was undertaken. Further, SPSS was employed for validation of the AZTool output results. The results showed that AZTool developed output areas out-perform the existing official SAL and SubPlaces with regard to minimum population threshold, population distribution and to some extent to homogeneity. However, the AZTool created output areas were less compact in shape than the SALs and SubPlaces in all geographical regions. In general, there was statistically significant ($p< 0.05$) difference in P2A means between the output areas, the SALs and. Therefore, it was concluded that AZTool program provides a new alternative to the creation of optimised census output areas for disseminations of population census data in South Africa.

Keywords: AZTool; Output Areas; Small Area Layers; SubPlaces; South Africa.

6.1 Introduction

Census data is a powerful tool for development and poverty reduction. It is a foundation for a wide range of research and analyses required to improve the standard of living of people in any country. Population projections are one of the most important analytical outputs based on census information (Schwabe, 2003; Stats SA, 2012). The characteristics of all individuals within a given area are recorded simultaneously in the census data collection. This data is utilised to inform government policy making, planning and administration. They are also used for demographics, social research and research to inform business, industry, labour and the public (Margeot and Ramjith, 2001; UN, 2002; 2004; 2009; Stats SA and HSRC, 2007; Owiti, 2008; Stats SA, 2012). In addition, census data provides a sampling framework for surveys that provide further insights into demographic and socio-economic trends that could be used to assess, monitor and evaluate the implementation of government policies and programs (Margeot and Ramjith, 2001; Stats SA and HSRC, 2007; Stats SA, 2012).

Many countries conduct censuses at regular intervals of five or ten years. In South Africa, the Statistics Act No. 6 of 1999 mandates Stats SA to carry out a census in a five-year cycle, but a decision was taken by Cabinet in 2004 that censuses would be undertaken in every ten years (Stats SA and HSRC, 2007). South Africa is one of the countries that have moved from using the same geographic unit for census enumeration and dissemination. For the 1991 and 1996 censuses, the same EAs were used for both census enumeration and dissemination. For the 2001 census, it was decided that census data must be released on an area larger than an EA due to confidentiality (Stats SA, 2003; Verhoef and Grobbelaar, 2005; Stats SA and HSRC, 2007). Stats SA then attached two names each EA and a spatial layer was created from the name attributes (SubPlaces and MainPlaces). Most users of the census data believed that these areas were too large. This resulted in the creation of the SAL using a non-zone design approach with the aim of meeting South African census user needs. A similar non-zone design approach was also employed in the creation of SAL for the 2011 census data. As indicated earlier, the main objective of the SAL was to have a spatial area layer that corresponded as much as possible to the EA layer, but remained within the confidentiality limit of 500 people (Verhoef and Grobbelaar, 2005). For instance, for the creation of SAL in 2005, the following criteria were set and adhered to as far as possible: firstly, EAs could only be merged if they are within the same SubPlace; secondly, EAs could only be merged if they have the same EA geography type; thirdly, an EA could only be merged if its population is less than 500; and lastly, the resulting small area polygons must have a population total of 500 and more (Verhoef and Grobbelaar, 2005).

In South Africa, it has not been established whether automated zone design generated census output areas could perform better than the existing official census dissemination areas with respect to certain design criteria or not. Automated zone design procedures tend to offer more efficient, systematic, and objective methodologies for designing optimised zoning systems than non-zone design methods. However, their success is dependent on the extent to which it is possible to model real-world phenomena and whether it is feasible to parameterise the required design criteria (Cockings *et al.*, 2011). Applications of the automated zone design, especially the AZTool program, are well described in previous studies (Flowerdew *et al.*, 2008; Ralphs and Ang, 2009; Cockings *et al.*, 2011; 2013; Martin *et al.*, 2013; Sabel *et al.*, 2013). Automated zone design methods offer more efficient, systematic, and objective methodologies for designing optimised zoning systems than manual methods (Cockings *et al.*, 2011). In the United Kingdom, Haynes *et al.* (2007) compared automated zone design program ''A2Z'' zones, developed by Daras (2006), with areal units identified subjectively by local government officers as communities in the city of Bristol. Their findings showed that the first automated zone design was much more successful in identifying homogenous deprivation areas than the subjective community (cf. the ICC values of 0.82 and 0.61), and was equally successful in identifying homogeneous areas of a particular housing type (0.51 and 0.51) even though zone design was much less compact in shape than the subjective areas. Their results further highlighted that automated zone design was close to replicating the subjective communities when the balance of objectives and boundary constraints was adjusted. In New Zealand, Ralphs and Ang (2009) compared the AZTool new geographies with existing official geographies. They found that the new geographies substantially out-performed the existing geographies across almost all of their optimisation criteria. In France, Sabel *et al.* (2013) compared the AZTool new zones with existing IRIS census areas to explore relationships between asthma and deprivation in Strasbourg. Their results indicated that the newly produced synthetic neighbourhood solution performed better than the then existing IRIS census areas, measured by improved statistical relationships between asthma and deprivation. Therefore the objective of this chapter was to compare the newly AZTool developed census output areas with existing official census output geographies such as SALs and SubPlaces in South Africa with the aim of evaluating the AZTool application in South Africa.

6.2 Methods

The study areas were Free State and Gauteng provinces of South Africa, which were representative of rural and urban settings, respectively (see Section 3.2).

The EAs from the 2001 census estimates data (HSRC, 2005) were used as building blocks for the development of new census output areas using automated zone design procedure. The 2001 SubPlaces data and the 2001 SALs data, from Stats SA were used for comparisons with the newly created census output areas. The 2001 SAL data did not have dwelling type variable (see Table 1.1). For geotype, it had 3 categories, namely, urban, rural, and mixed instead of 4 categories that were in the 2001 census EAs data, namely, formal urban, informal urban, informal rural (tribal Areas); and formal rural (farms). Hence the comparison on social homogeneity could not yield fruitful results. The 2001 SubPlaces data had similar variables with the EAs data which was used as building blocks for the newly developed output areas, therefore comparisons were made with all design criteria. The 2011 SALs data from Stats SA was also explored.

In order to create optimised census output areas in South Africa, the AZTool version 1.0.3 (Cockings *et al.*, 2011) was used. ESRI's ArcGIS 10.2 and Microsoft excel were employed for data preparation to be used by the AZTool software and for displaying AZTool output results.

As indicated in Table 6.1, the design criteria were that all output areas must not breach a minimum population threshold of 500, must be as homogeneous with regard to dwelling type and geotype and be as compact with regard to shape as possible. The population mean target was also set in order to control the population distribution. In this study, like in other studies (such as Cockings and Martin, 2005; Flowerdew *et al.*, 2008; Haynes *et al.*, 2008; Ralphs and Ang, 2009), output areas have been developed by taking existing areas (the 2001 census EAs) and using them as building blocks to create larger areas that are optimised based on the required design criteria.

Table 6.1: Design criteria for newly developed census output areas

Criteria	Description	Weighting
Minimum threshold population size[1]	500	N/A
Mean target population	1000	100
Homogeneity[2]	IAC score for dwelling type and geotype	100
Shape compactness[3]	Perimeter squared per area (P2A)	100

[1] Minimum population threshold used by Statistics South Africa in creation of SAL (Verhoef and Grobbelaar, 2005)

[2] Intra-Area Correlation (IAC) (Tranmer and Steel, 1998; 2001; Martin *et al.*, 2001; Flowerdew, 2011; Cockings *et al.*, 2013)

[3] Shape compactness (Cockings and Martin, 2005; Haynes *et al.*, 2007)

In order to statistically validate the results from the AZTool program, further quantitative statistical analyses were undertaken using SPSS. These included one-way ANOVA

Kolmogorov sminov test and a paired t-test. See Methods section in Chapter 1 for detailed methods.

6.3 Results

Table 6.2 shows the statistical characteristics of the newly developed output areas, the SALs and SubPlaces data for the 2001 census at all spatial levels in the Free State province. This table indicates that the confidentiality limit of 500 was respected at all spatial levels for the newly created output areas whereas for both the SALs and SubPlaces, this threshold was breached at all levels. Setting the mean target in output areas also made output areas to have much narrower and tighter population distribution than that of SALs. For instance, population distribution of the newly created output areas was compared with that of the SALs for Maluti-a-Phofung Municipality in Free State province (Figure 6.1). It is important to mention that maximum populations for output areas are a bit larger than those of the SALs at all levels. As indicated earlier, in many instances, the SubPlaces were too large for most census data users. This is illustrated in Table 6.2 as maximum population for a SubPlace could go as high as 93290 persons in the Free State province. With regard to social homogeneity, only the newly created output areas and SubPlaces could be compared as SAL social homogeneity could not yield fruitful IAC results due to lack of homogeneity variables as indicated in methods section. The IAC scores for the newly AZTool created output areas were slightly lower than those of SubPlaces at most levels except for provincial level where they both recorded the same IAC score of 0.59. When comparing compactness of the shapes, the output areas had slightly higher P2A mean values with lower standard deviations than the SALs at all spatial levels. This means that the newly created output areas were less compact in shape than the SALs in all regions. In general, there was statistically significant ($p< 0.05$) difference in P2A means between the output areas, the SALs and SubPlaces based on one-way ANOVA results. The LSD post-hoc test revealed that difference between the P2A means for the output areas and the SALs was not statistically significant ($p > 0.05$). The SubPlaces had higher P2A mean values with higher standard deviations than the output areas and the SALs. In addition, the P2A means difference between the SupPlaces and the output areas was statistically significant ($p < 0.05$) as well as between the SubPlaces and the SALs ($p < 0.05$). This shows that the SubPlaces were less compact in shape compared to both the output areas and the SALs.

Table 6.2: Statistical characteristics of newly developed output areas, the 2001 SALs and SubPlaces for Free State

	Number		Population			Shape		Homogeneity
	of Zones	Min	Max	Mean	SD	Mean	SD	IAC
Output Areas								
Phuthaditjhaba	48	649	2704	1113	346	28	9	0.21
Maluti-a-Phofung	349	610	2704	1027	232	32	13	0.50
Thabo Mofutsanyane	667	581	5292	1087	403	33	13	0.56
Free State	2440	547	9269	1101	489	31	12	0.59
Small Area Layers								
Phuthaditjhaba	68	408	1144	782	169	26	13	N/A
Maluti-a-Phofung	474	0	2071	761	248	30	14	N/A
Thabo Mofutsanyane	901	0	6701	806	359	30	13	N/A
Free State	3463	0	6701	782	318	29	13	N/A
SubPlaces								
Phuthaditjhaba	13	410	10507	4091	3565	43	29	0.29
Maluti-a-Phofung	110	0	22496	3280	4250	38	22	0.54
Thabo Mofutsanyane	223	0	25500	3255	3977	36	19	0.57
Free State	791	0	93290	3422	5974	34	22	0.59

Similar analyses were undertaken in Gauteng province at all spatial levels in order to get an understanding of comparisons at urban settings (Table 6.3). It is satisfying to note that AZTool created output areas adhered to the minimum population at all levels as it was in rural areas, which is very reassuring from a confidentiality perspective. As it was for rural areas, the SALs and SubPlaces breached the confidentiality threshold with minimum population of zero being record at all levels. Table 6.3 further shows that output areas recorded higher IAC scores than SubPlaces at all levels. In contrary to rural settings, this shows that the newly created output areas were more homogeneous than the SubPlaces based on dwelling type and geotype homogeneity variables. In terms of shapes, the AZTool optimised output areas had higher P2A mean values and standard deviations than the SALs showing that output areas were less compact than their counter-parts at all spatial levels. The difference between three P2A means was statistically significant ($p < 0.05$). On the contrary to the rural areas, the P2A mean difference between the output areas and the SALs was statistically significant ($p < 0.05$). Similar to rural areas, the SubPlaces had statistically significantly ($p < 0.05$) higher P2A mean values and standard deviations than the output areas and the SALs. It is interesting to see that all levels recorded the same P2A mean value of 29 even though their standard deviations tend to increase with spatial level, thus Pretoria Mainplace had standard deviation of 12, then 15 and 17 for City of Tshwane and Gauteng province, respectively.

Table 6.3: Statistical characteristics of the newly developed output areas, the 2001 SALS and SubPlaces for Gauteng

	Number		Population			Shape		Homogeneity
	of Zones	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
Output Areas								
Pretoria	500	621	5026	1056	320	28	11	0.09
City of Tshwane	1276	502	8802	1203	514	27	10	0.46
Gauteng Province	7253	501	9627	1214	520	27	9	0.45
Small Area Layers								
Pretoria	662	0	4227	794	301	26	10	N/A
City of Tshwane	1723	0	8092	886	442	25	11	N/A
Gauteng	10177	0	8092	868	389	25	10	N/A
SubPlaces								
Pretoria	157	0	26773	3346	4599	29	12	0.07
City of Tshwane	315	0	82002	4848	8764	29	15	0.45
Gauteng	2222	0	131662	3977	7403	29	17	0.44

Figure 6.1 compares the population distribution for AZTool newly created output areas and the SALs in Maluti-a-Phofung municipality. Figure 6.1a shows that AZTool successfully respected the confidentiality rule by having more than 500 people in all the areas. Kolmogorov sminov test results showed that both the output areas and the SAL population distributions were not normal ($p < 0.05$ in both cases). Furthermore, Figure 6.1 shows that AZTool newly created output areas population distribution follows a normal curve more than the SALs. For instance, frequencies are higher on the left side of the normal curve for the SALs instead of in the middle. This shows that the newly created output areas, with population target set to 1000, had a much narrower and tighter population distribution than the SALs. This makes the newly created output areas more ideal from user's perspective as the individual areas could be comparable to each other with regard to population size.

In general, the percentages of areas breaching the population thresholds for the SALs were 6.3% and 4.7%, for Free State and Gauteng provinces, respectively. In the SubPlaces data, 24.7% of areas fell below the 500 population confidentiality limit in the Free State province and 21.2% in Gauteng province. None of the areas breached the population confidentiality for the newly created output areas in both rural and urban areas. The rural areas seem to be more homogeneous than the urban areas at all regions for both newly created output areas and SubPlaces.

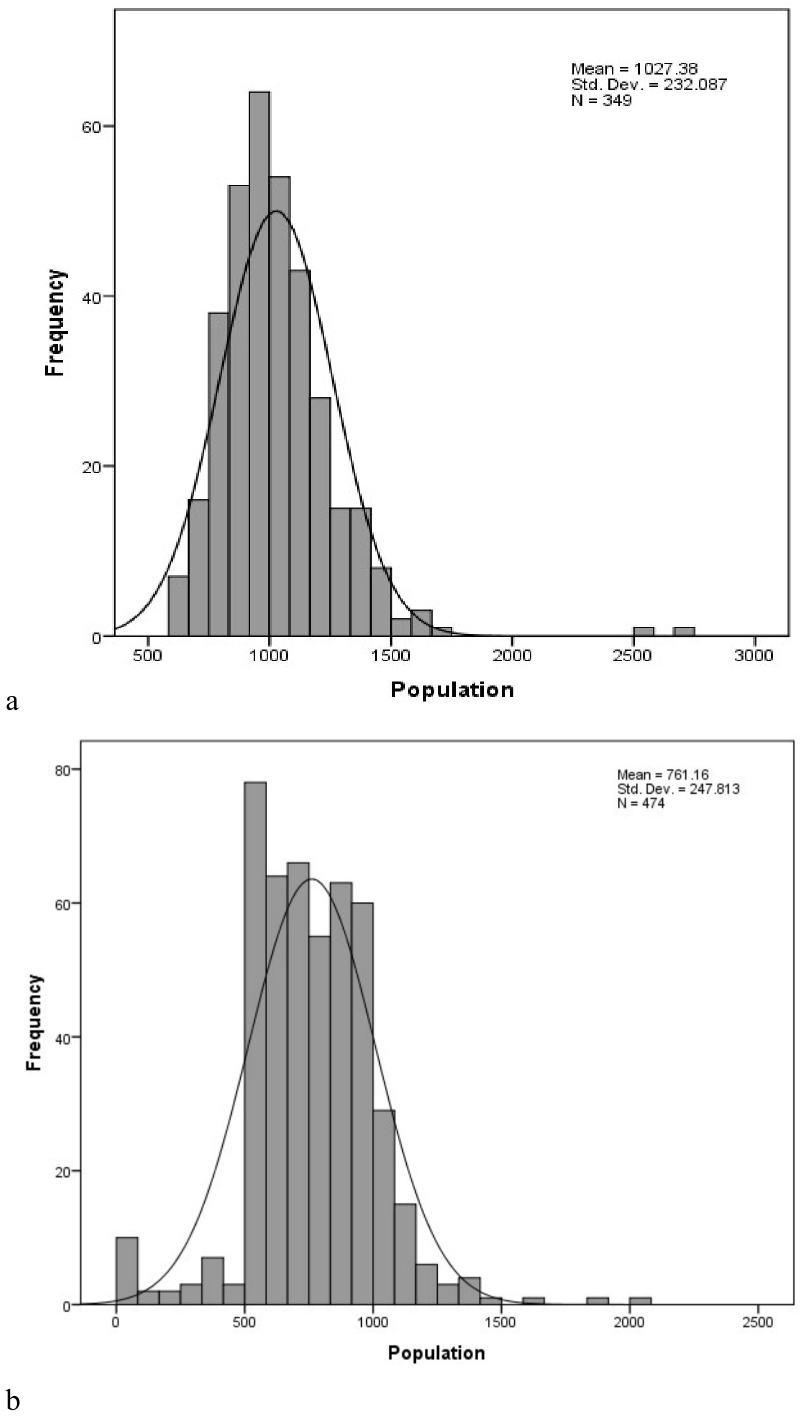


Figure 6.1: Population distribution for, a) AZTool output areas and b) SALs in Maluti-a-Phofung municipality

In terms of shape compactness, a further examination was done at provincial levels for both rural and urban settings for the newly created output areas, SALs and SupPlaces. Figure 6.2 indicates that output areas had higher P2A mean values and lower standard deviations than the SALs in both rural and urban provinces. The SubPlaces recorded higher shape mean values for both provinces but their standard deviations were also too high. The fact that standard deviations overlapped indicate that there were not significant differences between the P2A means. This was proven by performing ANOVA, which further revealed that the P2A mean difference between these three groups was not statistically significant ($p > 0.05$). When comparing the two provinces for all three areas, the rural province had slightly less compact shapes than the urban province as P2A means were higher than those of urban province. However, the differences in P2A means were also not significant as the standard deviations were overlapping.

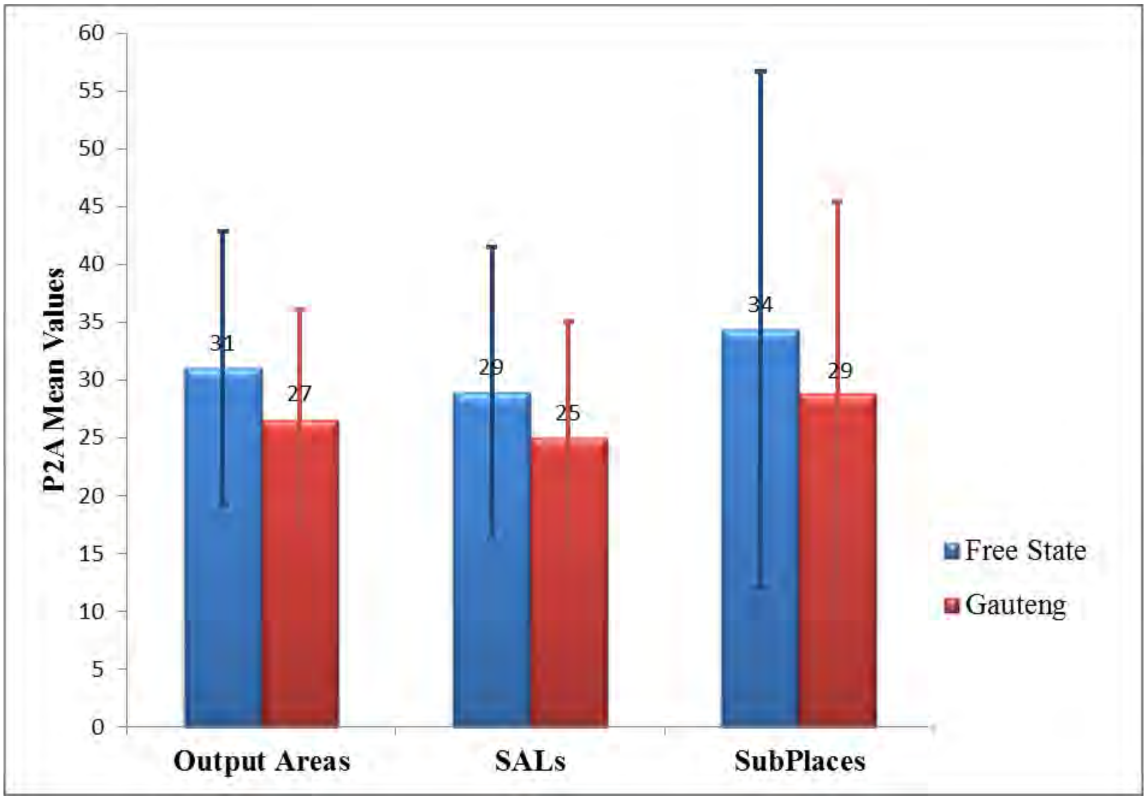
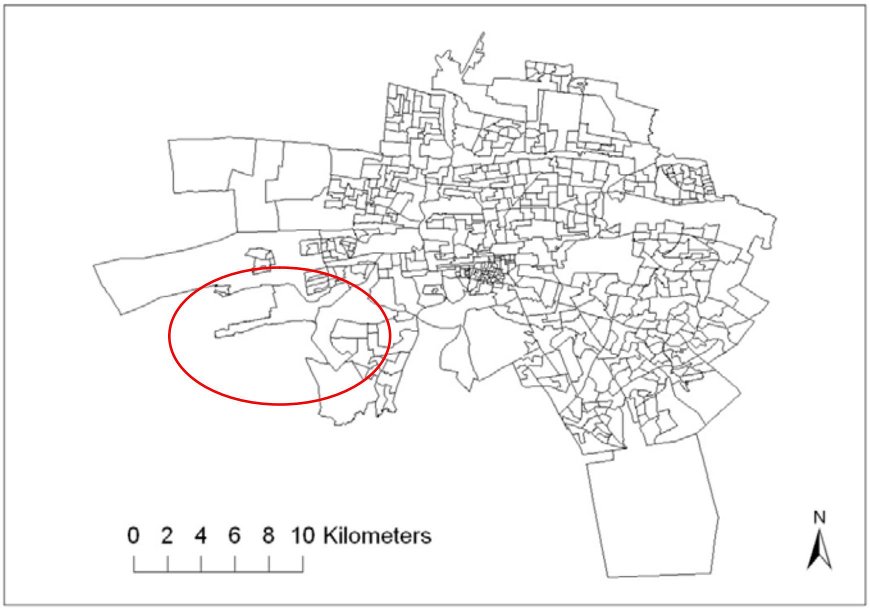
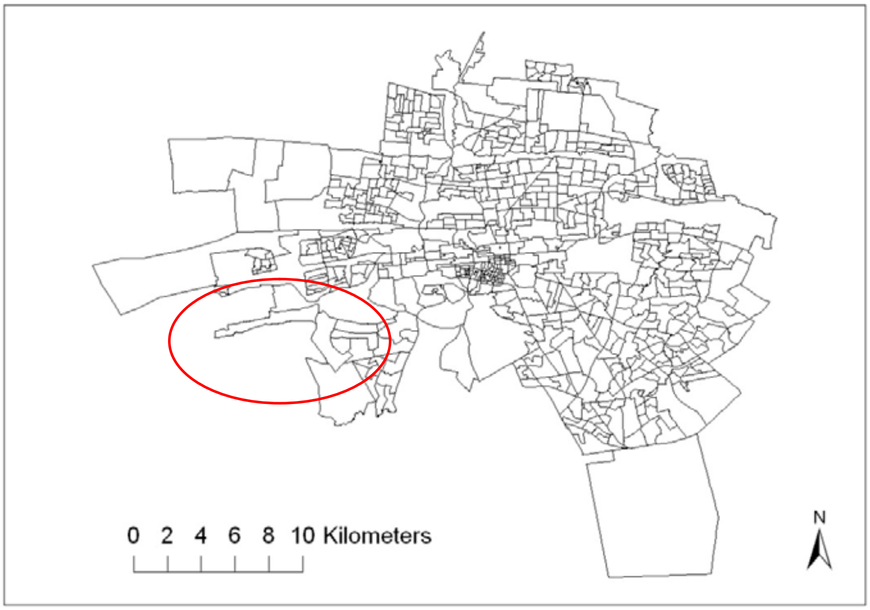


Figure 6.2: Shape means and standard deviations of the output areas, the SALs and the SubPlaces for the Free State and Gauteng provinces

Figure 6.3 shows visual comparison of shape compactness of the AZTool output areas and the SALs are for Pretoria mainplace. Figure 6.3a confirms that the output areas are more compact in shapes than the SALs. This can clearly be seen on the South Western part of the Pretoria mainplace where the SALs have some elongated areas (shown by red circle in Figures 6.3a and b).



a



b

Figure 6.3: Visual comparison of shape compactness for, a) output areas and b) SALs for Pretoria mainplace

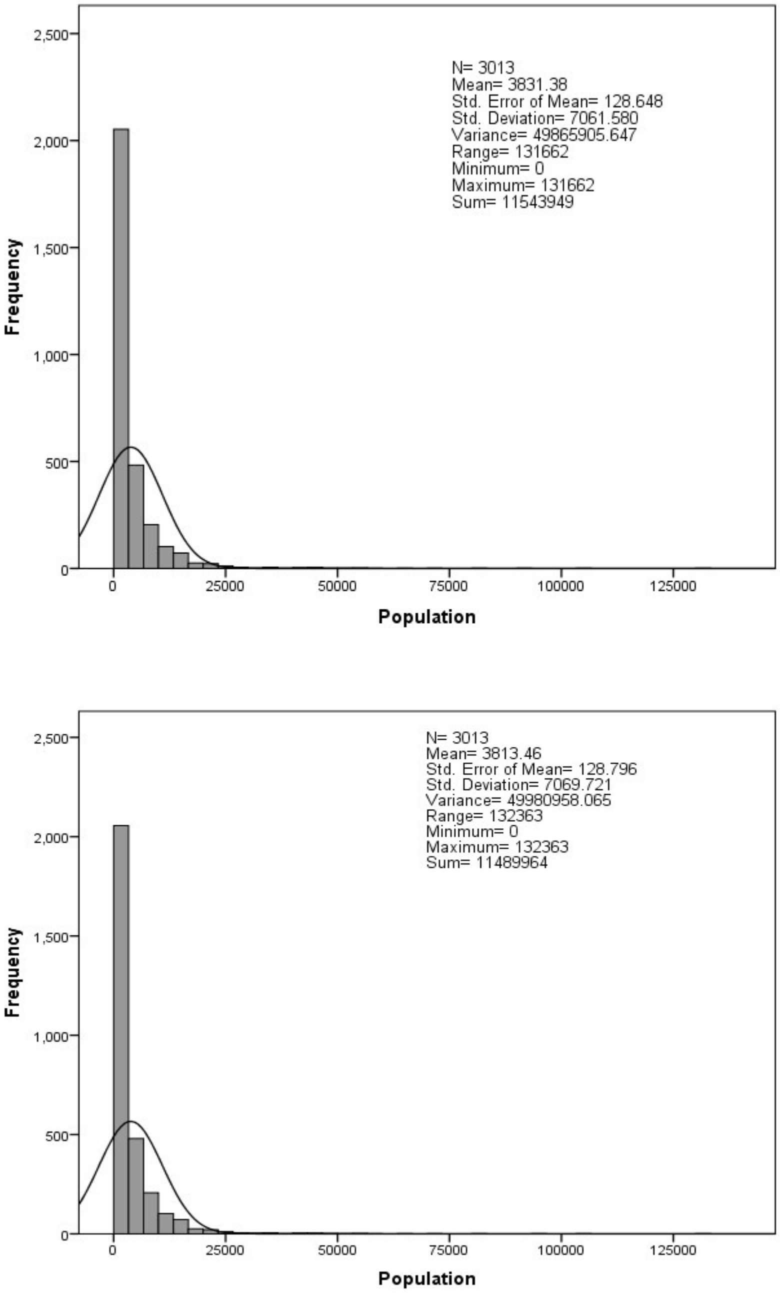
In an effort to compare newly developed output areas with SALs with all design criteria, the 2011 SAL data for the Free State was explored. Only Free State was used as indicated earlier that this province experienced low population growth and its provincial boundaries did not change from 2001. It is worth to note that for the 2011 SAL, only populated areas were captured hence all zero-populated areas were not included in the data. Table 6.4 indicates that minimum population for the 2011 SALs, the 2001 SALs and newly created output areas were 9, 0 and 547, respectively. Additionally, the newly created output areas were more compact in shape than the 2011 SALs and 2001 SALs. The 2001 SALs were less compact than the 2011 SALs. In terms of social homogeneity, the output areas were slightly less homogeneous with IAC score of 0.59 compared to 0.62 of the 2011 SALs. However, the geotype homogeneity variables for the 2011 SALs had only three categories (Urban, Rural and Farms) while the AZTool output areas had four categories which were Formal Urban, Informal Urban, Formal Rural (Tribal Areas) and Informal Rural (Farms).

Table 6.4: Statistical characteristics of the output areas, 2001 SALs, and 2011 SALs for Free State province

	Number		**Population**			**Shape**		**Homogeneity**
Free State	**of Zones**	**Min**	**Max**	**Mean**	**SD**	**Mean**	**SD**	**IAC**
Output Areas	2440	547	9269	1101	489	31	12	0.59
Small Area Layers 2001	3463	0	6701	782	318	29	13	N/A
Small Area Layers 2011	5129	9	5586	535	228	25	9	0.62

Further attempts were made to test how accurate or reliable the EAs 2001 estimates data was. The 2001 EAs estimates were aggregated to SubPlaces level in order to compare them with the original SubPlaces data from Stats SA. Figure 6.4 shows population distributions for the Free State and Gauteng province combined for both the aggregated SubPlaces and the original SubPlaces. As illustrated in Figures 6.4a and b, all statistics were almost similar for both datasets. In particular, when the total populations were compared, the difference was only 0.47%, indicating that the estimates were highly accurate.

Populations for both the aggregated SubPlaces data and the original SubPlaces data from Stats SA for Phuthaditjhaba mainplace were displayed in Figure 6.5. This shows that populations for the aggregated SubPlaces, derived from the 2001 EAs estimates data, were slightly higher than those of the original SubPlaces data in each individual areas.



a

b

Figure 6.4: Population distributions for a) original SubPlaces and b) aggregated SubPlaces

A paired t-test was perfomed to see if the means from these two datasets were the same. The results (t = 3.944, p = 0.002) showed that difference in mean populations from the aggregated data and the original data was statistically significant. The mean difference between the two datasets was 18. 77 with the 95% confidence interval ranging from 8.401 to 29.137. This indicates that, although the difference in means was statistically significant, it was actually relatively small. In order for these results to be valid, the differences between the paired values should be approximately normally distributed. Therefore, a simple Kolmogorov sminov test revealed that indeed the distribution of differences was normal (p > 0.05).

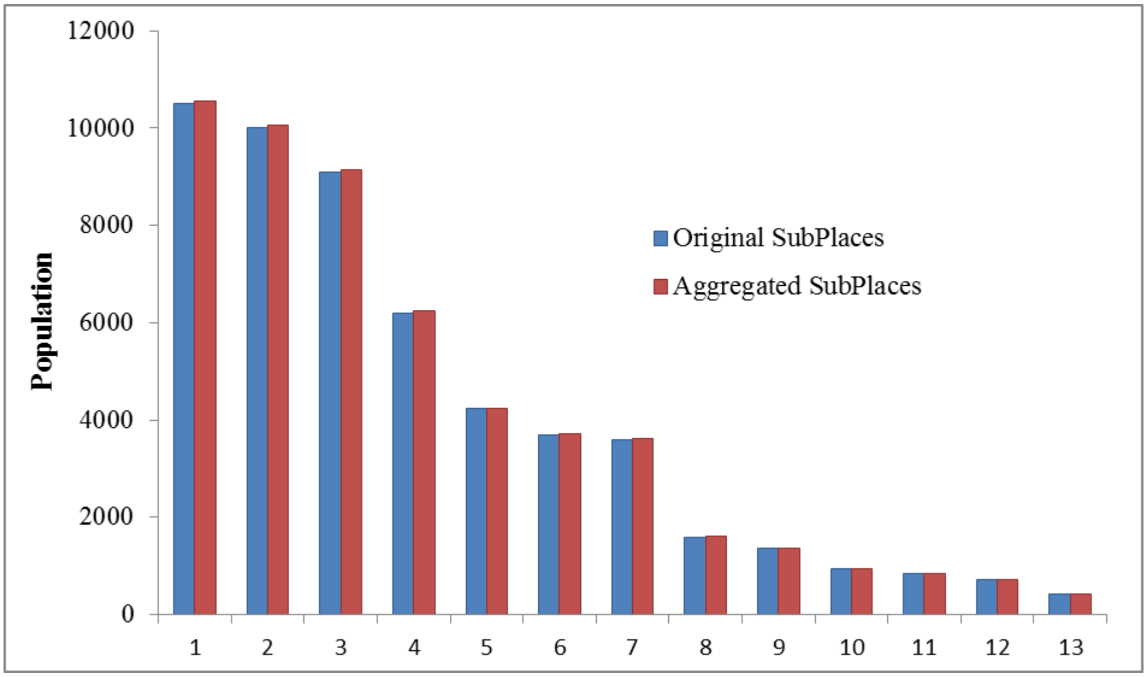


Figure 6.5: Comparison of the original SubPlaces with the aggregated SubPlaces population data for Phuthaditjhaba mainplace

A further comparison of the aggregated SubPlaces data with the original SubPlaces data was performed using the AZTool results outputs for all levels in both rural and urban settings. Table 6.5 shows that statistical qualities of these areas were mostly the same. It is important to note that when comparing IAC scores at each spatial level for both aggregated SubPlaces and original SupPlaces, IAC scores were exactly the same for rural areas. The urban areas showed a slight difference in these IAC scores as the ones for aggregated SubPlaces were slightly higher than those of the original SubPlaces. These comparisons provide some confidence with regard to the use of the 2001 EAs estimates as building blocks in the development of the AZTool output areas.

Table 6.5: Statistical characteristics of the original SubPlaces and the aggregated SubPlaces

	Number	Population				Shape		Homogeneity
	of Zones	Min	Max	Mean	SD	Mean	SD	IAC
Original SubPlaces								
Phuthaditjhaba	13	410	10507	4091	3565	43	29	0.29
Maluti-a-Phofung	110	0	22496	3280	4250	38	22	0.54
Thabo Mofutsanyane	223	0	25500	3255	3977	36	19	0.57
Free State	791	0	93290	3422	5974	34	22	0.59
Aggregated SubPlaces								
Phuthaditjhaba	13	412	10554	4109	3581	43	29	0.29
Maluti-a-Phofung	110	0	22594	3260	4281	38	22	0.54
Thabo Mofutsanyane	223	0	25612	3253	4001	36	19	0.57
Free State	791	0	93701	3397	5969	34	22	0.59
Original SubPlaces								
Pretoria	157	0	26773	3346	4599	29	12	0.07
City of Tshwane	315	0	82002	4848	8764	29	15	0.45
Gauteng	2222	0	131662	3977	7403	29	17	0.44
Aggregated SubPlaces								
Pretoria	157	0	26915	3363	4625	29	12	0.08
City of Tshwane	315	0	82440	4872	8815	29	15	0.46
Gauteng	2222	0	132363	3962	7415	29	17	0.45

6.4 Discussion

Results from this study show that the newly developed output areas using the AZTool are very much an improvement over the SALs and the SubPlaces. This was proven by the fact that newly developed output areas effectively satisfied minimum and target population thresholds, while the population distributions were much narrower in range than those of the existing SALs and SubPlaces. The confidentiality limit of 500 people was respected at all spatial levels in both rural and urban settings for the newly created output areas whereas for both SALs and SubPlaces the confidentiality limit of 500 persons was breached at all levels. The fact that the AZTool generated output areas did not breach minimum population throughout all study areas (chapters 3, 4, 5 and 6) is very reassuring from a confidentiality perspective. Similarly, Ralphs and Ang (2009) found that AZTool successfully constrained all tracts to be of at least the required minimum size.

The population target criterion also yielded positive results as the AZTool census output areas had a much narrower and tighter population distribution than that of the SALs. The summary of rules set for the creation of the SALs did not have population target, which would have made them to have a better distribution than the current one. The importance of tighter and narrower population distribution is that it makes the newly created output areas more ideal from a census data user's point of view as the individual areas could be easily compared in terms of their population size distribution. This supports previous arguments by Verhoef and Grobbelaar (2005) that in many instances the SubPlaces were too large for most census data users, hence the initiative taken to develop the SALs in 2005. It is worth mentioning though that some of the AZTool output areas had very large population sizes. This is due to the fact that the 2001 EAs were used as building blocks for the creation of these output areas. The availability of data at the lower level than EAs, household level, would allow the optimisation algorithm to have more options in generating output areas that meet target population sizes as much as possible. Other studies such as Cockings and Martin (2005); Flowerdew *et al.* (2008); Haynes *et al*. (2007); Haynes *et al*. (2008) and; Ralphs and Ang (2009) also identified similar challenges as they used existing areas as building blocks hence the flaws of such areas were inherited into the generated output areas (Drackley *et al.*, 2011).

With regard to homogeneity, only the SubPlaces were comparable to the newly generated census output areas as the SALs did not produce IAC score due to insufficient homogeneity variables. The newly AZTool created output areas were slightly less homogeneous than the SubPlaces at most levels in rural areas. The provincial level was an exception as the output areas and SubPlaces shared the same degree of homogeneity in terms of dwelling type and geotype variables, with both having IAC score of 0.59. In contrary, the urban settings showed that the newly created output areas were more homogeneous than the SubPlaces based on dwelling type and geotype homogeneity variables at all levels (Haynes *et al.*, 2007).

A further attempt was undertaken to use the 2011 SALs which had both dwelling type and geotype variables in order to be able compare the AZTool output areas with the SALs. The census output areas generated from the AZTool program were slightly less homogeneous with IAC score of 0.59 compared to 0.62 of the 2011 SALs for Free State province. This might be due to the fact that the geotype homogeneity variable for the 2011 SALs had only three categories while the AZTool output areas had four categories. Although these results are from two different censuses, it is believed that they are good indication of how homogeneity variable would perform in the comparisons as this province did change at all in terms provincial boundaries and did not change much in terms of population growth. It is important though to note that due to infrastructure development the dwelling type variable might have been affected from 2001 to 2011.

Findings from both rural and urban areas showed that the AZTool newly created output areas were less compact in shape compared with the SALs at all regions. This is in line with previous findings by Haynes *et al.* (2007) and Ralphs and Ang (2009), where automated zone design output areas were slightly less compact than original existing geographies. The SubPlaces had less compact shapes than both the output areas and the SALs.

Comparing the 2001 EAs estimates data, which were used as building blocks for the output areas, with the original 2001 SubPlaces data brought some confidence in the AZTool newly created output areas as these have to be close to reality as much as possible. This does not rule out the fact that the original EAs data from Stats SA would have been preferable had it been available. There are some positives to be drawn from this study as the comparison of automated zone design census output areas with existing official census output areas had not been reported before in South Africa. Therefore, findings from this study provide a new alternative to the creation of optimised census out areas for population census disseminations.

6.5 Conclusions

In general, the AZTool generated census output areas out-performed the existing official SALs and SubPlaces, non-zone design developed geographies. This was proven by the fact that AZTool output areas effectively satisfied minimum and target population thresholds, while the population distributions are much narrower in range than those of the existing SALs and SubPlaces. However, the AZTool census output areas were less compact in shape compared with the SALs at all spatial levels in both rural and urban environments. A comparison of automated zone design census output areas with existing official census output areas has not been reported before in South Africa. Therefore, it was concluded that findings from this study provide a new alternative to the creation of optimised census output areas for upcoming population census disseminations in South Africa.

CHAPTER 7

SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

Many countries use the same demarcation areas for both data collection and the dissemination of their census data. Prior to the 2001 census, this was also the case in South Africa. For the 2001 census, it was decided that census data must be disseminated on an area larger than an EA due to confidentiality issues. Then two names were attached to each EA and a spatial layer was created from the place name attributes (SubPlaces and MainPlaces) by Stats SA. Most users of the census information had concerns as these areas were too big. Therefore, the SAL was created in an effort to meet needs of the census data users in 2005. The idea behind the SAL was to have a spatial area layer that corresponds as much as possible to the EA layer, but lies within confidentiality limits (Verhoef and Grobbelaar, 2005; Grobbelaar, n.d.). Only the first two concerns directly considered, the census confidentiality limits and population size. The social homogeneity and output shape were not directly addressed. Even the confidentiality issue was not fully addressed as there was still lot of SALs that breached confidentiality limit. Therefore the advanced techniques of automated zone design methods, such as AZTool program, are required in the development of optimized census output areas in South Africa that would address these four issues as much as possible.

The overall aim of this study was to develop an optimized census output areas using AZTool program in South Africa. The specific objectives were as follows: 1) to create optimised output areas using AZTool with 2001 census EAs as building blocks, 2) to determine the statistical qualities of the AZTool output areas, 3) to determine the effect of building blocks designs on the statistical characteristics of AZTool output areas, 4) to compare the AZTool developed output areas with existing official census output areas in South Africa, and 5) to evaluate the AZTool application in South Africa.

It is important to note that all these objectives were achieved by this study. Objectives 1and 2 were achieved through applications of various criteria such as number of iterations, respecting higher geographic region, donuts constraint, applying different weights using 2001 EAs as the building blocks. Results from both Chapters showed that the AZTool output areas were better off than the original EAs with regard to the primary criterion of minimum population threshold. Objective 3 was achieved through comparisons of AZTool optimised outputs areas developed from the EAs, the SALs and Subplaces. Objective 4 was attained through a comparative evaluation of the AZTool output areas with existing official census dissemination areas such as the SALs and the SubPlaces. Objective 5 was a cross-cutter as all the 4 objectives had an impact on it. In general, all these objectives contributed, with objective 1 being the core, in the robust development of census output areas in South Africa as per the title of this thesis.

7.2 Summary of findings

This study illustrates the potential applications of automated zone design techniques and the potential challenges that may occur when applying such techniques in the creation of optimised output areas in South Africa. This was highlighted in Chapter 3. The 2001 census EAs, from the estimates data, were used as building blocks for the creation of AZTool output areas. The IAC values at the lower geographical levels were lower than those of any higher geographical level in both rural and urban areas. This indicated that higher geographical levels produced more homogeneous output areas than lower geographical levels. The newly AZTool created output areas from the rural areas had higher degrees of homogeneity than those from urban areas. However, the urban areas were more compact than the rural areas. Overall, the higher degree of homogeneity for all provinces combined (urban and rural provinces), the IAC of 0.62, suggests that the selected variables could be used as good indicators of social homogeneity in creating homogeneous output areas across South Africa. Generally, the IAC of 0.5 is regarded as a very reasonable degree of homogeneity (Sabel *et al.*, 2013). In addition, in all experiments that were performed in both urban and rural areas at all geographical levels or regions, the confidentiality limit of 500 people was adhered to. This was a huge success as this was a challenge in the SALs created for the 2001 census dissemination.

Results further showed the donuts constraint did not have impact on the quality of output areas with regard to shape and degree of homogeneity. Therefore, there was no restriction made to exclude donuts in the final output areas. In order to make sure that output areas nested within higher geographical level or region, the AZTool was set to respect higher geographical regions. Unfortunately, the program did not produce any solutions when any of the higher geographical levels were respected. To solve this, higher geographical regions could be analysed separately and merged at the end to produce an overall output even though this might be time consuming for larger samples.

The positive findings in Chapter 3 prompted the interest in further determining the statistical qualities of the newly AZTool generated census output areas in Chapter 4. The results showed that confidentiality was adhered to at all geographical levels in AZTool output areas in both rural and urban areas compared to the original EAs where the minimum population was zero at all geographic levels. In terms of population distribution, the AZTool optimised output areas had much narrower and tighter population distributions than their counterparts. Furthermore, Shapiro-wilk test results indicated that the population distribution for the AZTool output areas was normal ($p > 0.05$) while for the EAs it was not normal ($p < 0.05$). However, the newly created AZTool output areas were slightly less compact compared to the original EAs in both rural and urban settings. Findings from Chapter 4 also showed that different shape weights had a great improvement on the visual display of the output areas. This was proven by the fact that when the criterion for the shape was set to carry ten times more weight than population and homogeneity, the shapes of output areas were more circular and less elongated. Furthermore, when the 2011 census data was explored, the findings showed that the AZTool output areas substantially out-performed the original SALs with regard to confidentiality as none of the output areas were below the 500 minimum population thresholds. The population means of the output areas were more close to the set population target of 1000 than the ones of the original SALs at all spatial levels. Hence the output areas had tighter population distribution than the original SALs. However, the output areas were less compact compared to the SALs at all spatial levels or regions.

In Chapter 5, the effects of different building blocks designs on the statistical characteristics of the AZTool census output areas were explored. Different spatial layers (2001 EAs, 2001 Subplaces, 2001 and 2011 SALs) were used as building blocks for the generation of census output areas in order to test these. Findings of this study highlighted that different building blocks did have an impact on the statistical qualities of the AZTool optimised output areas. In general, all output areas from all the different building blocks respected the minimum population thresholds, which was a huge success from confidentiality point of view. When the EAs and the SALs were used as building blocks in all study areas, statistics showed that output areas from the EAs had slightly higher population means and lower standard deviations than the ones from the SALs. However, the means from the two sets of optimised output areas were close to user-defined population target mean. Unsurprisingly, the output areas from the SubPlaces had higher population means and standard deviations than the ones from the two sets of buildings blocks at all levels in both rural and urban settings. This was expected as the SubPlaces are much bigger in size than the two sets of building blocks and the two sets nest within the these SubPlaces in the South African geography hierarchy.

In terms of shape compactness, the optimised output areas generated from the EAs and the SALs were almost similar in all study areas. The output areas from the SubPlaces had higher P2A means and higher standard deviations than the ones generated from the EAs and the SALs at all spatial levels. Clearly, this shows that the output areas created using the EAs and the SALs were more compact than those developed using the SubPlaces as building blocks. This was statistically significant at 95% confidence level ($p < 0.05$). The LSD post-hoc test results showed that P2A means for output areas from both the EAs and the SALs were not significantly different ($p > 0.05$). The results further indicated that the difference between P2A means of those generated from the SubPlaces and the EAs as well as the difference between P2A means of those created from the SubPlaces and the SALs was statistically significant ($p < 0.05$ in both cases). However, the P2A mean values and standard deviations of the output areas from the SubPlaces in urban areas were not as higher as they were in rural areas, but they were statistically ($p < 0.05$) different from their counter-parts based on one-way ANOVA results. The P2A mean values and standard deviations for urban areas were close to the mean values and standard deviations of output areas from the EAs and the SALs. Therefore the effects of different building blocks on the shape characteristics of the AZTool output areas tend to be noticed more in the rural areas, especially between lower level building blocks (EAs and SALs) and higher level building blocks (SubPlaces). Findings from all the three different building blocks further showed that the AZTool optimised output areas from urban areas were more compact than their counter-parts at all levels of geography.

With regard to degree of homogeneity, only the AZTool optimised output areas from the EAs and SubPlaces yielded reasonable results. Results showed that the output areas created using the EAs as building blocks were more homogeneous than those created from the SubPlaces in both rural and urban settings. As the 2001 SALs did not have enough homogeneity variables, the 2011 SALs were explored, but only for the Free State province as it did not change its provincial boundaries from 2001. The AZTool output areas created using 2011 SALs as building blocks were less homogeneous than the ones from 2001 EAs. This is an indication that the AZTool output areas generated from smaller areas tend be more homogeneous than the ones generated from larger areas when using dwelling type and geotype as homogeneity variables. Furthermore, the IAC scores could also be used as an assessment of magnitude of the scale effect because the IAC scores are adjusted for population size. In general, the higher IAC scores indicate the higher scale effects. Therefore, the higher IAC scores produced by output areas from smaller areas in this chapter indicate that scale effects are clearly identified when smaller areas are used as building blocks than when larger areas are considered.

Chapter 6 explored the comparisons of the newly AZTool developed census output areas with existing official census output areas (the SALs and the SupPlaces) in South Africa. Findings from this Chapter showed that the newly developed output areas using the AZTool were a significant improvement over the SALs and the SubPlaces. This was proven by the fact that newly developed output areas effectively satisfied minimum and target population thresholds, while the population distributions were much narrower in range than those of the existing SALs and SubPlaces. The confidentiality limit of 500 people was respected at all spatial levels in both rural and urban settings for the newly created output areas whereas for both SALs and SubPlaces, the minimum population threshold was breached at all levels. The fact that the AZTool generated output areas did not breach minimum population threshold throughout all study areas is very reassuring from a confidentiality perspective.

Findings from Chapter 6 further showed that the newly AZTool created output areas were slightly less homogeneous than the SubPlaces at most levels in rural areas. The provincial level was an exception as the output areas and SubPlaces shared the same degree of homogeneity in terms of dwelling type and geotype variables, with both having IAC score of 0.59. In contrary, the urban settings showed that the newly created output areas were more homogeneous than the SubPlaces based on dwelling type and geotype homogeneity variables at all levels. The 2011 SALs which had both dwelling type and geotype variables were evaluated. Results highlighted that the AZTool census output areas, generated from 2001 EAs, were slightly less homogeneous with IAC score of 0.59 compared to 0.62 of the 2011 SALs for the Free State province.

When comparing compactness of the shapes, the output areas had slightly higher P2A mean values with lower standard deviations than the SALs at all spatial levels in rural areas. This means that the newly created output areas were less compact in shape than the SALs in all regions. In general, there was statistically significant ($p < 0.05$) difference in P2A means between the output areas, the SALs and SubPlaces based on one-way ANOVA results. The LSD post-hoc test revealed that difference between the P2A means for the output areas and the SALs was not statistically significant ($p > 0.05$). The SubPlaces had higher P2A mean values with higher standard deviations than the output areas and the SALs. In addition, the P2A means difference between the SupPlaces and the output areas as well as between the SubPlaces and the SALs was statistically significant ($p < 0.05$). This shows that the SubPlaces were less compact in shape compared to both the output areas and the SALs. When similar analysis were undertaken in urban areas, results indicated that the AZTool optimised output areas had higher P2A mean values and standard deviations than the SALs showing that output areas were less compact than their counter-parts at all spatial levels. The ANOVA analysis revealed that there was a statistically significant ($p < 0.05$) difference between three P2A means. On the contrary to the rural areas, the P2A mean difference between the output areas and the SALs was statistical significant ($p < 0.05$). The SubPlaces had significantly ($p < 05$) higher P2A mean values than both the output areas and the SALs, as it was for the rural settings. Generally, the findings from both rural and urban areas showed that the AZTool newly created output areas were less compact in shape compared with the SALs at all regions. The SubPlaces had less compact shapes than both the output areas and the SALs.

The comparison of the 2001 EAs estimates data, aggregated to SubPlace level, with the original 2001 SubPlaces data from Stats SA brought some confidence in the newly established AZTool created output areas as these have to be close to the reality as much as possible. For instance, findings in Chapter 6 showed that populations from the aggregated SubPlaces were slightly higher than those from the original SubPlaces data in each individual areas. A paired t-test results ($t = 3.944$, $p = 0.002$) showed that difference in mean populations from the aggregated data and the original data was statistically significant. The mean difference between the two datasets was 18. 77 with the 95% confidence interval ranging from 8.401 to 29.137. This indicates that, although the difference in means was statistically significant, it was actually relatively small. In order for these results to be valid, the differences between the paired values should be approximately normally distributed. Therefore, a simple Kolmogorov sminov test results indicated that indeed the distribution of differences between the aggregated SubPlaces and the original SubPlaces was normal ($p > 0.05$). These findings do not rule out the fact that the original EAs data from Stats SA would have been preferable had it been available. However, there are some positives to be drawn from this study as the comparison of automated zone design census output areas with existing official census output areas had not been reported before in South Africa. Therefore, these positive results from this study provide a new alternative to the creation of optimised census out areas for population census disseminations.

7.3 Contribution to knowledge

Findings from this study contribute to the research in general and to the potential applications of automated zone design methods in developing countries. This was shown by the capabilities and advantages of using AZTool program to create optimised census output areas in developing countries, with South Africa being used as example. This was also proven by the fact that the application of AZTool program improves the overall statistical quality of a census output areas especially adhering to the confidentiality limit and narrower population distribution. The data dynamics such as exploring different building blocks in the creation of AZTool optimised census output areas add value to the existing knowledge. The evaluations of the AZTool program at different geographical levels or regions in both urban and rural areas provide new contribute to the existing knowledge. Finally, the comparison of automated zone design census output areas with existing official census output areas had not been reported before in South Africa.

7.4 Study limitations

Among the limitations of this study is that the accessibility of data at lower geographical levels such as household and EA levels as well as recent the 2011 census data was not successful. Hence, only the 2001 census EA estimates data was used as building blocks. There seems to be a challenge with regard to accessing census data at lower geographical levels for research purposes and other purposes such as business and marketing due to confidentiality. The use of household level data would have minimised the flaws carried by administrative data (EAs), which were created for a different purpose, as building blocks into the created output areas. Therefore, caution should be taken when using pre-existing input areas to aggregate them into larger areas, as the flaws that are inherent in the building blocks would be carried over into the output areas as well as possible bias and potential errors associated to the MAUP. Secondly, among the barriers of using the AZTool program for creating census output areas is that respecting a higher geographical region constraint is restrictive and often prevents solutions being found at all, which was also the case in this study. Higher geographical regions or administrative areas change overtime as population grow, making it difficult to keep census output areas nested within them, hence some countries such as Australia, England and Wales have removed the requirement for census output areas to be nested within certain higher geographic levels. Lastly, the 2011 census data at the SAL level excluded zero-populated areas therefore, this resulted in 15 isolated building blocks which were excluded for further analysis as the AZTool works with contiguous building blocks. Even though these isolated building blocks constituted only 0.15% of the total population of Free State province, they might have some slight contribution on the statistical characteristics of the AZTool output areas created using the 2011 SALs as building blocks.

7.5 Conclusions

The main aim of this study was to develop an optimized census output areas using AZTool program in South Africa. Based on the results reported in Chapters 3 to 6, there is potential in application of automated zone design methods, particularly AZTool program, in the creation of optimized census output areas in South Africa. Furthermore, comparisons of the newly AZTool created output areas with existing South African official census output areas in Chapter 6 support this conclusion. In addition, findings from this study contribute to the research in general and to the potential applications of automated zone design methods. These concluding remarks emanated from the following findings in this thesis:

- The primary criterion of minimum population threshold of 500 people was kept and not breached throughout all the AZTool newly created output areas at different geographical levels as well as in both rural and urban areas (Chapters 3 to 6).

- The second most prioritised criterion of homogeneity of output areas showed the IACs of 0.45 for Gauteng province, 0.52 for the Free State, and 0.62 for both provinces combined. These IAC values are encouraging as international studies show that the IAC of 0.5 is regarded a very reasonable degree of homogeneity within output areas (Chapter 3).
- The AZTool generated output areas substantially out-performed the original EAs and the SALs in terms to minimum population threshold and population distribution statistical qualities (Chapter 4).
- Chapter 4 further highlighted that the AZTool optimised output areas had much narrower and tighter population distributions than the original EAs. This was further proven statistically by running Shapiro-wilk test which showed that the population distribution for the AZTool output areas was normal ($p > 0.05$) whereas for the EAs population distribution was not normal ($p < 0.05$).
- Different building blocks designs (EAs, SALs and SubPlaces) did have an impact on the statistical qualities of the AZTool optimised output areas in both rural and urban settings in South Africa (Chapter 5).
- The AZTool output areas generated from smaller areas (EAs) were more homogeneous than the ones generated from larger areas (SubPlaces) when using dwelling type and geotype as homogeneity variables (Chapter 5).
- The AZTool optimised output areas from the smaller areas allowed a clear distinction of the scale effects than output areas from larger areas (Chapter 5).
- The newly AZTool developed census output areas out-performed the existing official SALs and SubPlaces, non-zone design developed geographies. This was proven by the fact that AZTool output areas effectively satisfied minimum and target population thresholds, while the population distributions were much narrower in range than those of the existing SALs and SubPlaces (Chapter 6).
- The newly developed AZTool census output areas were less compact in shape compared with the SALs at all spatial levels in both rural and urban settings (Chapter 6).
- A comparison of automated zone design census output areas with existing official census output areas had not been reported before in South Africa. Hence, the findings from this study provide a new alternative to the creation of optimised census output areas for upcoming population census disseminations in South Africa (Chapter 6).

7.6 Recommendations and suggestions for future research

Future research and general work should evaluate the applications of automated zone design methods, such as AZTool computer program, in the creation of census output areas across the entire country. This is from the basis that this study was conducted in only two provinces out of the nine provinces of South Africa. In taking this research or work forward, it is important that the following recommendations are considered:

- The availability or accessibility of data at lower geographical level, such as EA or household levels, is highly recommended as this would improve developments of robust and optimized output areas using automated zone design techniques. The accessibility of census data at lower levels such as household would also allow exploration of other building blocks such as grid squares. In addition to these, the flaws of the original building blocks (EAs) are often inherited into the AZTool output areas. Therefore, household level data should be made accessible even if it is under secure condition for future research.
- The creation of optimised output areas using other census homogeneity variables should also be explored.
- The total household size should also be explored as a confidentiality threshold.
- Unfortunately, the AZTool program did not produce any solutions when any of the higher geographical levels were respected. Hence, it is recommended that further research should be explored to see the cause of this in the context of South African geographical areas.
- Other AZTool program design criteria which were not explored in this study should also be evaluated.
- The effects of zero-populated building blocks on the AZTool output areas should also be investigated in South African context.
- Future research should look into perceptions of South African census data users.
- From a policy and practice perspective, as indicated in Chapter 3, it is important to note that this research was a stand-alone project with the aim of influencing policies and practice of government stakeholders such as Stats SA. It is believed that these initial experiments regarding the AZTool applications in the creation of census output areas in South Africa would encourage future possible collaboration between the candidate and the government stakeholders such as Stats SA as well as other South African census data users.

- The fact that official 2011 SAL census data had a significant number of SALs that fell below the official minimum threshold of 500 persons is worry some, For instance 42.2% of the SALs had below 500 persons in Free State while Gauteng had 27% of the SALs which breached the confidentiality limit. Therefore, one of the implications from this study is that there should be a change with regard to the current policy and practice in census dissemination. This change should be guided by evidence based or practical research such as the one from this study.

REFERENCES

Ajayi, M.T.A., Nuhu, M.B., Bello, M.Z., Shuaib, S.I., Owoyele, G., Onuigbo, I., Babalol, A. and Alias, A. (2015). A GIS based assessment of the relationship between housing conditions and rental value in government built housing estates in Minna. *Journal of Building Performance*, 6 (1), 50 – 62.

Avenell, D., Noble, M. and Wright, G. (2009). South African datazones: A technical report about the development of a new statistical geography for the analysis of deprivation in South Africa at a small area level, *CASASP Working Paper No. 8, Oxford:* Centre for the Analysis of South African Social Policy, University of Oxford.

Bajat, B., Krunić, N., Petrović, M.S. and Kilibarda, M. (2013). Dasymetric modelling of population dynamics in urban areas. *Geoditski Vestnik,* 54 (4), 777 – 792.

Basson, C. (2007). The use of Geographic Information Systems, Global Positioning Systems and automated demarcation technologies in surveys and census mapping at Statistics South Africa. *The Regional Workshop on Census Cartography and Management*, 8 – 12 October 2008, Zambia.

Christopher, A. J. (2001). Urban segregation in Post-apartheid South Africa. *Urban Studies,* 38 (3), 449 – 466.

Christopher, A. J. (2009). Delineating the nation: South African censuses 1865–2007. *Political Geography*, 28 (2), 101 – 109.

Christopher, A. J. (2010). A South African domesday book: The first union census of 1911, *South African Geographical Journal*, 92, 1, 22 – 34.

Cockings, S. and Martin, D. (2005). Zone design for environment and health studies using pre-aggregated data. *Social Science and Medicine*, 60, 2729 – 2742.

Cockings, S., Harfoot, A., Martin, D. and Hornby, D. (2011). Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 census output geographies for England and Wales. *Environment and Planning A*, 43, 2399 – 2418.

Cockings, S., Harfoot, A., Martin, D. and Hornby, D. (2013). Getting the foundations right: spatial building blocks for official population statistics. *Environment and Planning A*, 45, 1403 – 1420.

Daras, K. (2006). *An information statistics approach to zone design in the geography of health outcomes and provision*. Unpublished PhD Thesis. University of Newcastle, England.

Dhonju, H.K., M.S.R. and Duwal, S. (2015). Dasymetric mapping of census data for Nepal towards improved disaster risk assessment studies. *FIG – ISPRS workshop, 2015: International Workshop on Role of Land Professionals and SDI in Disaster Risk Reduction: In the Context of Post 2015 Nepal Earthquake*, 25 – 27 November 2015, Kathmandu, Nepal.

Drackley, A., Newbold, K.B., and Taylor, C. (2011). Defining socially-based spatial boundaries in the Region of Peel, Ontario, Canada. *International Journal of Health Geographics*, 10 (38), 1 – 12.

Dube, C. (2005). Census geography of South Africa. *AfricaGIS Conference,* 31 October – 4 November 2005, South Africa.

Duke-Williams, O. and Rees, P. (1998). Can offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure. *International Journal of Geographical Information Science,* 12 (6), 579 – 605.

Dumedah, G., Schuurman, N. and Yang, W. (2008). Minimizing effects of scale distortion for spatially grouped census data using rough sets. *Journal of Geographical Systems*, 10, 47 – 69.

Duque, J.C., Ramos, R. and Surinach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review,* 30 (3), 195 – 220.

Eagleson, S., Escobar, F. and Williamson, I. (1999). Spatial hierarchical reasoning applied to administrative boundary design using GIS. http:///www.sli.unimelb.edu.au/research/publications/IPW/ipw_paper36.pdf [Accessed 19 November 2009].

Exeter, D.J., Feng, P.B.Z., Flowerdew, R., and Schierloh, N. (2005). The creation of 'Consistent Areas Through Time' (CATTs) in Scotland, 1981 – 2001. *Population Trends,* 119, 28 – 37.

Eze, C.G. (2009). The role of Satellite Remote Sensing data and GIS in population census and management in Nigeria: A case study of an Enumeration Area in Enugu, Nigeria. *Scientific Research and Essay,* 4 (8), 763 – 772.

Flowerdew, R., Manley, D.J. and Sabel, C.E. (2008). Neighbourhood effects on health: Does it matter where you draw the boundaries? *Social Science and Medicine*, 66, 1241 – 1255.

Flowerdew, R. (2011). How serious is the Modifiable Areal Unit Problem for analysis of English census data? *Population Trends*, 145, 106 – 118.

Gehlke, C.E. and Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29, 169 – 170.

GeoSpace, (n.d.). Population and housing census mapping solutions. GeoSpace International. Available from:
http://www.geospace.co.za [Accessed 10 November 2010].

Gregory, I.N. (2002). The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, 26 (4), 293 – 214.

Gregory, I.N. and Ell, P.S. (2005). Analysing spatiotemporal change by use of National Historical Geographical Information Systems; population change during and after the Great Irish Famine. *Historical Methods,* 38 (4), 149 – 167.

Grobbelaar, N. (n.d.). The development of a Small Area Spatial Layer to serve as the most detailed geographical entity for the dissemination of census 2001 data. Statistics South Africa. Available from:
http://mapserver2.statssa.gov.za/geographywebsite/Docs/AfricaGIS/685_Development%20of%20the%20Small%20Are%20Layer.pdf [Accessed 19 November 2009].

Hakizimana, J. (2009). Importance of new technologies for census; a South African Experience. *5th African Symposium for Statistical Development (ASSD) Workshop,* 19 - 21 October 2009, South Africa.

Haynes, R., Daras, K., Reading, R., and Jones, A. (2007). Modifiable neighbourhood units, zone design and residents' perceptions. *Health and Place*, 13, 812 – 825.

Haynes, R., Jones, A., Reading, R., Daras, K. and Emond, A. (2008). Neighbourhood variations in child accidents and related child and maternal characteristics: does area definition make a difference? *Health and Place,* 14, 693 – 701.

Heywood, I., Cornelius, S. and Carver, S. (2002). *An introduction to Geographical Information Systems.* 2nd ed. Pearson Prentice Hall, London.

Hofstee, P. and Islam, M. (2004). Disaggregation of census districts: Better population information for urban risk management. *25th Asian Conference on Remote Sensing*, 22 – 26 November 2004, Thailand.

HSRC, (2005). 2001 census EA estimates. Human Sciences Research Council in collaboration with Prof DJ Stoker. Pretoria, South Africa.

Khalfani, A.K. and Zuberi, T. (2001). Racial classification and the modern census in South Africa, 1911 – 1966. *Race and Society,* 4, 161 – 176.

Khatun, H., Falgunee, N. and Kutub, M.J.R. (2015). Analysing urban population density gradient of Dhaka Metropolitan Area using Geographic Information Systems (GIS) and Census Data. *Malaysian Journal of Society and Space*, 11(13), 1 – 13.

Kitchin, R. and Tate, N.J. (2000). *Conducting research in human geography: Theory, methodology and practice*. Pearson Prentice Hall, London.

Klosterman, R.E. (1995). The appropriateness of Geographic Information Systems for regional planning in the developing world. *Computers, Environment and Urban Systems*, 19 (1), 1 – 13.

Laldaparsad, S. (2007). Census mapping and the use of geo-spatial technologies (A case of South Africa). *United Nations Expert Group Meeting on Contemporary Practices in Census Mapping and Use of Geographical Information Systems,* 29 May – 1 June 2007, New York.

Lehohla, P. (2005). Statistics needs geography; Geography needs statistics. *AfricaGIS Conference,* 31 October – 04 November 2005, South Africa.

Linard, C., Gilbert, M., Snow, R.W., Noor, A.M. and Tatem, A.J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS ONE*, 7 (2), e31743.

Lombaard, M. (n.d.). South African postcode geography. Statistics South Africa, South Africa.

Loots, H. (2005). Addressing the challenges associated with census mapping in Africa. *15th Conference of Commonwealth Statistics,* 5 – 9 September 2005, South Africa.

Maantay, J.A., Maroko, A.R. and Herrmann, C. (2007). Mapping population distribution in the urban environment: The Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science,* 34 (2), 77 – 102.

MacEachren, A. (1985). Compactness of geographic shape: comparison and evaluation of measures. *Geografiska Annaler B,* 67, 53 – 67.

Manley, D., Flowerdew, R. and Steel, D. (2006). Scales, levels and processes: studying spatial patterns of British census variables. *Computers, Environment and Urban Systems,* 30, 143 – 160.

Mansour, S., Martin, D. and Wright, J. (2012). Problems of spatial linkage of a geo-referenced Demographic and Health Survey (DHS) dataset to a population census: A case study of Egypt. *Computers, Environment and Urban Systems*, 36 (4), 350–358.

Margeot, H. and Ramjith, S., (2001). The South African census 2001 spatial information system data capture problems. *International Conference on Spatial Information for Sustainable Development,* 2 – 5 October 2001, Kenya.

Martin, D. (1997). From enumeration districts to output areas: Experiments in the automated creation of a census output geography. *Statistical Commission and Economic Commission for Europe Conference of European Statistics, Work Session on GIS,* 22 – 25 September 1997, United Kingdom.

Martin, D. (1998a). 2001 Census output areas: From concept to prototype, *Population Trends,* 94, 19 – 24.

Martin, D. (1998b). Optimizing census geography: The separation of collection and output geographies. *International Journal of Geographical Information Science*, 12, 673 – 685.

Martin, D. (2000). Towards an integrated national socioeconomic GIS: The geography of the 2001 census in England and Wales. *3rd AGILE Conference on Geographic Information Science,* 25 – 27 May 2000, Finland.

Martin, D., Nolan, A. and Tranmer, M. (2001). The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A,* 33, 1949 – 1962.

Martin, D. (2003). Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, 17 (2), 181 – 196.

Martin, D. (2002). Geography for the 2001 census in England and Wales, *Population Trends,* 108, 7 – 15.

Martin, D. (2004). Neighbourhoods and area statistics in the post 2001 census era. *Area,* 36 (2), 136 – 145.

Martin, D., Cockings, S. and Harfoot, A. (2013). Development of a geographical framework for census workplace data. *Journal of Royal Statistical Society*, 176 (2), 1 – 18.

Mbogoni, M. (2012). Report of the United States of America on the 2010 World Programme on population and housing censuses. *United Nations International Seminar on population and housing censuses: Beyond the 2010 Round,* 27 - 29 November 2012, Republic of Korea.

Mobbs, J.D. (1998). Australia comes to its census: The Public Sector Mapping Agencies and the 1996 Australian census of population and housing. Public Sector Mapping Agencies, Australia. Available from:

http://www.geom.unimelb.edu.au/fig7/Brighton98/Comm7Papers/SS34Mobbs.html [Accessed 14 April 2012].

Murayama, Y. (2001). The contribution of GIS to geographical research. *GeoJournal,* 52, 163 – 164.

Murray, A. T. (2010a). Quantitative geography. *Journal of Regional Science*, 50 (1), 143 – 163.

Murray, A. T. (2010b). Advances in location modelling: GIS linkages and contributions. *Journal of Geographical Systems*, 12, 335 – 354.

Ndubi, J.N. (2007). Use of GIS in census management and mapping: Kenyan experience. *The Regional Workshop on Census Cartography and Management*, 8 – 12 October 2007, Zambia.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers, NS,* 2 (4), 459 – 472.

Openshaw, S. (1984). Problem concepts and techniques in modern geography: The Modifiable Areal Unit. *CATMOG 38,* 1 – 39.

Openshaw, S. and Rao, L. (1995). Algorithms for reengineering 1991 Census geography. *Environment and Planning A,* 27 (3), 425 – 446.

Owiti, P.O. (2008). Embedded GIS for census mapping. Nairobi, Kenya. Available from: http://www.map-gis-rs.blogspot.com/2008/03/gis-for-census.html [Accessed 8 January 2012].

Prasad, M. (2006). Maps and census. *GIS Development: Asia Pacific. The Monthly Magazine on GIS, December 2006,* 10 (12), 7 – 48.

Rain, D. (2008). Handbook on geographic databases and census mapping (Draft Version). *United Nations Expert Group Meeting on Measuring the Economically Active Population in Censuses,* 7 - 10 April 2008, New York.

Ralphs, M. and Ang, L. (2009). Optimized geographies for data reporting: Zone design tools for census output geographies. *Statistics New Zealand Working Paper No 09–01.* Statistics New Zealand, Wellington.

Ratcli¨e, J.H. and McCullagh, M.J. (1999). Hotbeds of crime and the search for spatial accuracy. *Journal of Geographical Systems*, 1, 385 – 398.

Reynolds, H.D. (1998). *The Modifiable Unit Problem: Empirical analysis by statistical simulation,* Unpublished PhD Thesis. University of Toronto, Canada.

Sabel, C.E., Kihal, W., Bard, D. and Weber, C. (2013). Creation of synthetic homogeneous neighbourhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France. *Social Science and Medicine*, 91, 110 – 121.

Schlossberg, M. (2003). GIS, the US census and neighbourhood scale analysis. *Planning, Practice and Research,* 18 (2-3), 213 – 217.

Schwabe, C. (2003). *The South African census user's handbook: Analysing data from 1996 census.* Human Sciences Research Council Publishers, Cape Town, South Africa.

Seid, Y. and Gutu, S. (2009). Census undertaking in pastoral areas and application of new technologies in 2007 population and housing census. *The 57th Session of The International Statistical Institute,* 16 – 22 August 2009, South Africa.

Stats SA, (2003). Census 2001. How the count was done. Statistics South Africa, South Africa.

Stats SA and HSRC, (2007). *Using the 2001 census: Approaches to analysing data.* A collaboration between Statistics South Africa (Stats SA) and the Human Sciences Research Council (HSRC), 244 p, South Africa.

Stats SA, (2012). *Census 2011 methodology and highlights of key results*. Statistics South Africa. Pretoria: Report no. 03-01-42, 29 pp., ISBN 978-0-621-41389-2.

Stevens, F.R., Gaughan, A.E., Linard, C., and Tatem, A.J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE,* 10(2), e0107042. doi:10.1371/journal.pone.0107042.

Tabatabai, P., Henke, S., Susˇac, K., Kisanga, O.M.E., Baumgarten, I., Kynast-Wolf, G., Ramroth, H. and Marx, M. (2014). Public and private maternal health service capacity and patient flows in southern Tanzania: using a geographic information system to link hospital and national census data. *Global Health Action,* 7, 1 – 11.

Tye, A. (2009). The emergence of GIS and GPS integrating with census data and its impact for the developing world. *The 57th Session of The International Statistical Institute,* 16 – 22 August 2009, South Africa.

UN, (2000). *Handbook on Geographic Information Systems and digital mapping. Studies in Methods, Series F No. 79.* New York: United Nations Publication.

UN, (2004). Integration of GPS, digital imagery and GIS with census mapping. *United Nations Expert Group Meeting to Review Critical Issue Relevant to the Planning of the 2010 Round of Population and housing Censuses,* 15 – 17 September 2004, New York.

UN, (2007). *Report of Sub-regional workshop on census cartography and management.* 8 – 12 October 2007, Zambia.

UN, (2009). *Handbook on geospatial infrastructure in support of census activities. Studies in Methods, Series F No. 103.* New York: United Nations Publication.

Valente, P. (2010). Census taking in Europe: How are populations counted in 2010? *Population and Societies,* 467. Available from: http://www.ined.fr/fichier/s_rubrique/19135/pesa467.en.pdf [Accessed 20 February 2015].

Verhoef, H. and Grobbelaar, N. (2005). The development of a Small Area Layer for South Africa for census data dissemination. Statistics South Africa. Available from: http://www.cartesia.org/geodoc/icc2005/pdf/poster/TEMA26/HELENE%20VERHOEF.pdf [Accessed 19 November 2009].

Wall, H. (n.d.). GIS based dissemination of census data in Trinidad and Tobago: A Caribbean experience. Central Statistical Office, Ministry of Planning and Development, Republic of Trinidad and Tobago. Available from: http://www.unstats.un.org/unsd/demographic/meetings/wshops/Trinidad_22Oct07/docs/countries_presentations/Trinidad_Tobago_Paper.pdf [Accessed 27 November 2010].

Wang, X., Liu, Z. and Chen, S. (2001). Some issues on urban census GIS designing, *Geo-spatial Information Science*, 4 (4), 25 – 31.

Weir-Smith, G. (2014). An overview of the geographic data of unemployment in South Africa. *South African Geographical Journal*, 96 (2), 134 – 152.

Weir-Smith, G. and Ahmed, F. (2013). Unemployment in South Africa: Building a spatio-temporal understanding. *South African Journal of Geomatics*, 2(3), 218 – 230.

Wu, C. and Murray, A. T. (2007). Population estimation using Landsat Enhanced Thematic Mapper Imagery. *Geographical Analysis*, 39, 26 – 43.

Wu, S., Wang, L. and Qiu, X. (2008). Incorporating GIS building data and census housing statistics for sub-block-level population estimation. *The Professional Geographer,* 60 (1), 121 – 135.

Yuan, Y., Smith, R.M. and Limp, W.F. (1997). Remodelling census population with spatial information from Landsat TM imagery. *Computers, Environment and Urban Systems*, 21 (3 – 4), 245 – 258.

Zimstat, (2013). Census 2012. National Report, Zimbabwe National Statistics Agency, Zimbabwe.