Integrative Use of Census and GIS

Verno Andries Ferreira B.Sc. IT (GIS), B.Sc.Hons (Geography) (2010073746)

Research report presented in partial fulfilment of the requirements for a Master's degree (in Town and Regional Planning) at the University of the Free State

Supervisor: M. Campbell

Nov 2015

DEPARTMENT OF TOWN AND REGIONAL PLANNING

UNIVERSITY OF THE
FREE STATE
UNIVERSITEIT VAN DIE
VRYSTAAT
YUNIVESITHI YA
FREISTATA

UFS
UV

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

November 2015

ABSTRACT

Census is an ancient phenomenon, and Geographic Information Systems (GIS) is a modern day marvel. What both have in common is their direct relationship to geography. Despite the wealth of information available in the census, unearthing this information with GIS is largely underutilised. This research essay opens with a review on census and GIS as two components for integration. To assess integrative use between census and GIS for decision making, a custom framework was developed called CENGIS (derived from census and GIS) to assess integrative use through key aspects such as tabulation, representation, aggregation and disaggregation. Each integrative aspect is then evaluated according to frequency of use and overall usability from which the degree of integrative use is determined. In conclusion the study ends with a synthesis on its key findings as well as proposals for future research.

Keywords: census, geographic information systems, integrative use, tabulation, census cartography, decision making

ACKNOWLEDGEMENTS

Supervisor

To my supervisor for her innovative contribution toward the completion of this research project.

Family

To my family for their patience and support in this venture.

Employer

To my current boss (Stephanus Minnie) for allowing me sufficient time toward completing this project.

All

Towards all who participated indirectly in completion of this project.

TABLE OF CONTENTS

TABLES

FIGURES

ACRONYMS AND ABBREVIATIONS

CENGIS	– Census and GIS
CGIS	– Canada Geographic Information System
DUF	– Dwelling Unit Frame
EAs	– Enumerated Areas
GIS	– Geographic Information System
GIT	– Geographic Information Technology
IDP	– Integrated Development Plan
MAUP	– Modifiable Area Unit Problem
NDP	– National Development Plan
SAL	– Small Area Layer
SDF	– Spatial Development Framework
SPOT5	– Satellite Pour I‘Observation de le Terre No. 5
StatsSA	– Statistics South Africa

CHAPTER 1: INTRODUCTION AND RESEARCH QUESTION

"The first census in 1790 asked just six questions: the name of the head of the household, the number of free white males older than 16, the number of free white males younger than 16, the number of free white females, the number of other free persons, and the number of slaves."

Tom G. Palmer.

Census has come a long way with the first official count done under the supervision of King Servius Tullius of Rome, which resulted in a mere 87 000 people in total (Tenney, 1930). The word itself is derived from *Latin*, which means to keep track of adult males fit for military service. The famous historical account of a census in Biblical chronology under the order by Caesar Augustus is documented in Luke chapter two. Census formed part of the foundation stone of the ancient Roman civilisation. It dynamically transformed the military and political outlook of the empire, who esteemed themselves more than just a barbarian horde, also a populous capable of collective action.

Over the year's census have progressed from solely population counts to a highly sophisticated enumeration of household profiles, service provision, income, expenditure, education etc. Nowadays, census information is even questioned for putting people's personal safety at risk when disclosing private information for 'governmental planning'. In spite of the questionable breach in personal safety, census nevertheless remains an intrinsic part of administrative action. The fact that census has a strong relationship with geography greatly enhances its usefulness in solving spatial disparities. According to the current statistician-general of South Africa, Pali Lehohla: "there is nothing as powerful as small area information, a statistical representation of the area in a map, for that talk in much better understanding…"

1.1. OVERVIEW AND BACKGROUND

Integrative use between census and Geographic Information Systems (GIS) evolved since 1996 until present. This evolution can be accredited to the technological revolution that started especially since the 1990s. Census on the one hand, is an ancient phenomenon, with the first official count recorded early 500 B.C. at the dawn of the Roman Empire. GIS, on the other hand, considered a modern-day marvel, which was created by Roger Tomlinson in the 1960s. The need to map, manage and analyse large areas of terrain gave rise to the first functional GIS (Burrough, 2001, p. 361). The first official GIS was called the Canada Geographic Information System (CGIS), capable of digitally representing old cartographic maps and allowing users to seamlessly connect multiple maps into one great mosaic from which users could query information. Over the next 50 years this idea evolved into a fully-fledged GIS used in almost every conceivable discipline that uses spatial data in its analysis (McGrath & Sebert, 1999).

In addition to GIS, census data has become the norm in strategic planning. Graphs, charts and tables derived from censuses feature frequently in municipal, provincial and national planning regulations, such as the Integrated Development Plan (IDP), Spatial Development Frameworks (SDFs), National Development Plan (NDP), sector plans, precinct plans etc. Apart from the government, the private sector also utilises census data in planning related activities. Academic institutions in particular are a notable user of census data, especially for research papers, where statistical profiles are derived from census data. Census data is very useful for any large-scale-planning initiative, regardless of sector, discipline or institution.

Despite the statistical use of census information, the geographic component is seldom explored or utilised (Kakembo & van Niekerk, 2014, p. 451). Despite the apparent value residing within census and GIS integration, active utilization thereof has very limited in reports produced by public or private planning institutions. GIS also known as digital cartography, offers census data users great value for planning facilitation. Census seamlessly integrates into the core functionalities of GIS for storage, retrieval, analysis and display of spatial data (Burrough, 2001, p. 363). GIS converts census data into digital cartography capable of displaying vast amounts of data in a condensed way.

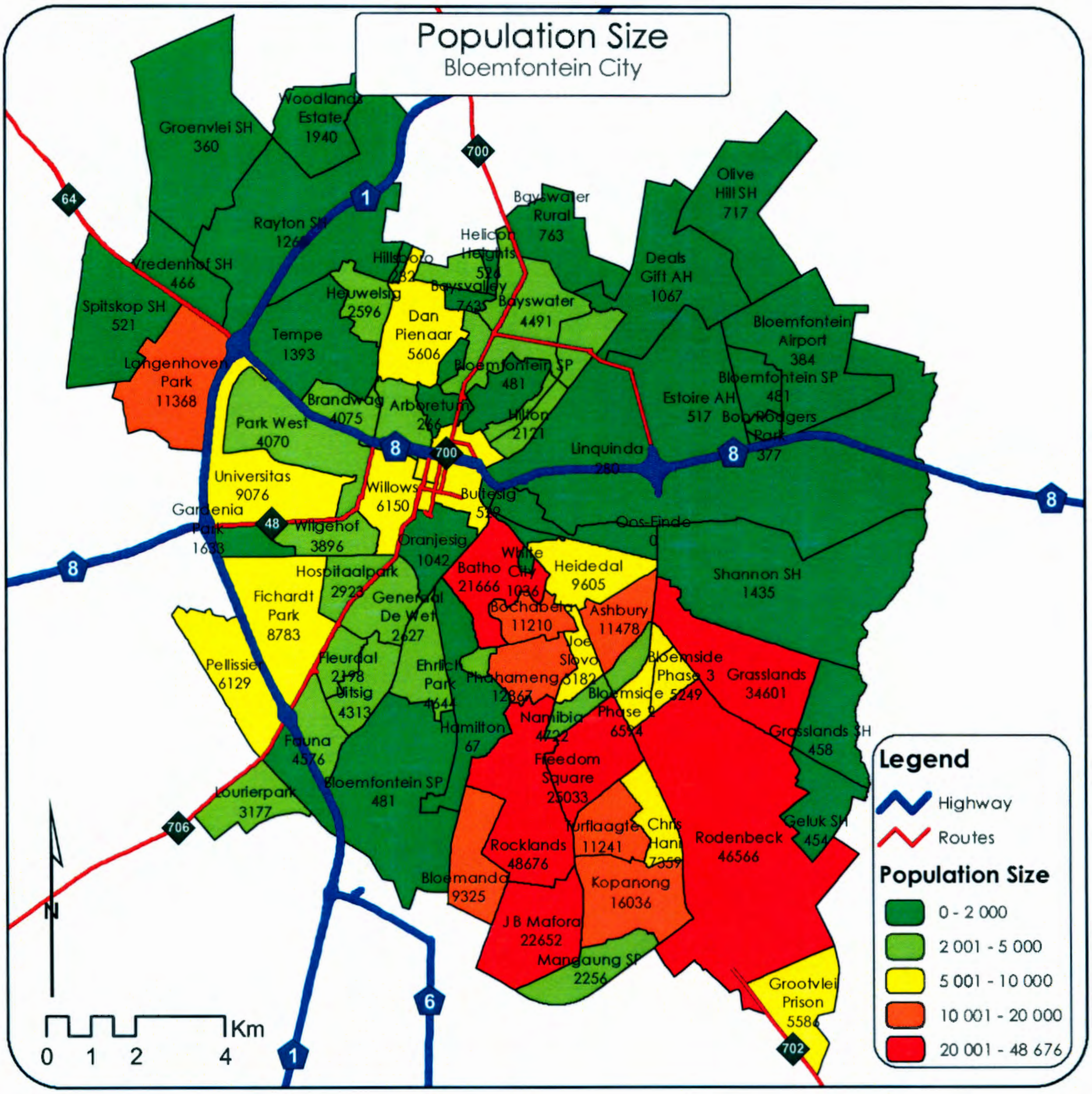


Figure 1 Census data in map form

Interpretation of census data in map format 'paints' a more informative picture, showing spatial relationships, areas of interest, which are otherwise difficult to see in the conventional graphs.

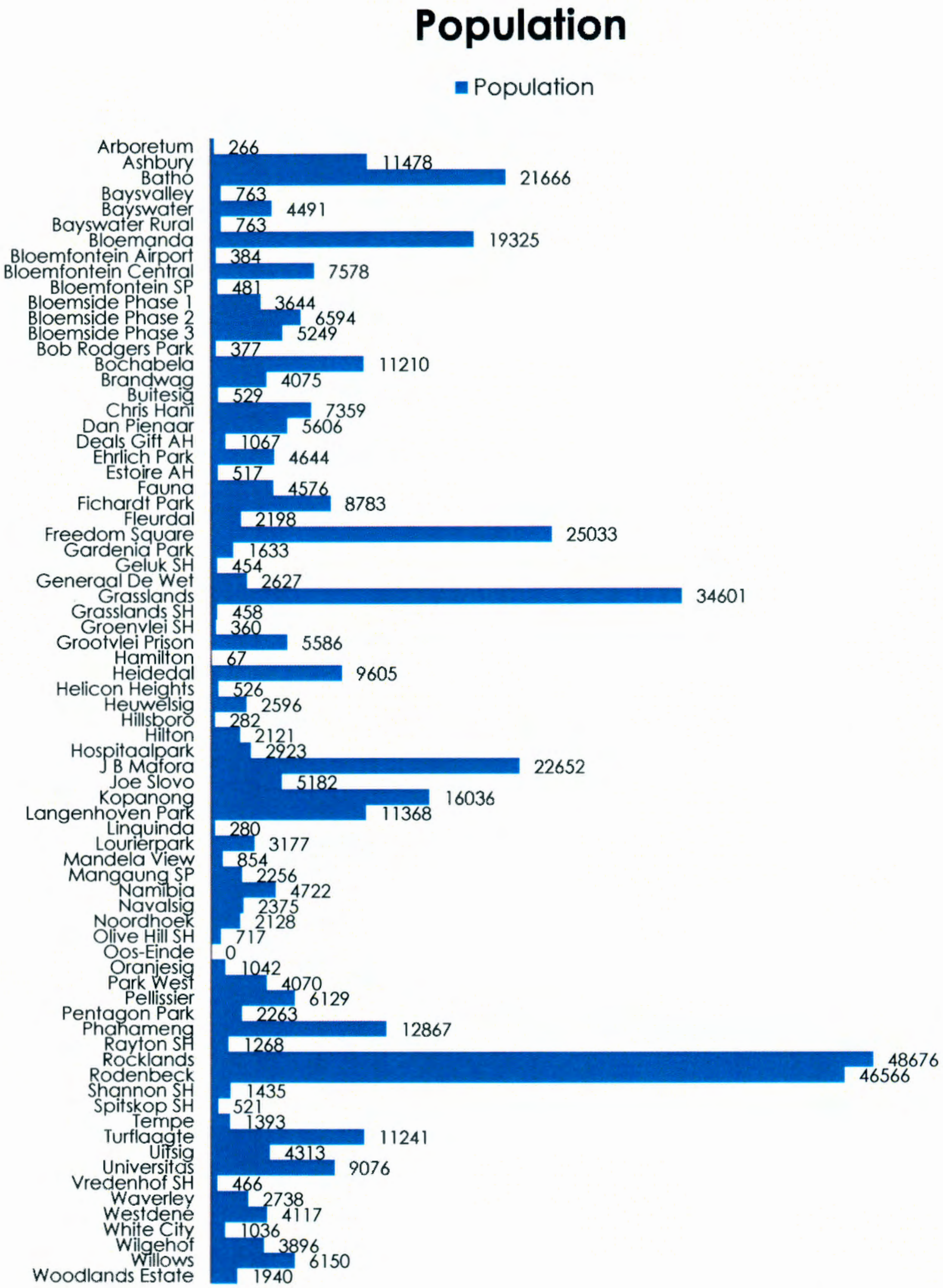


Figure 2 Graph depicting population size of Bloemfontein city (Census, 2011)

Although the same information is encapsulated in two different formats, the map version speaks more vividly to the user's understanding than the graph itself.

1.2. RESEARCH FOCUS

This study is intended to evaluate integrative use between census and GIS. The geographic correlation between the two has been largely omitted in planning facilitation. However, despite the wealth of information hidden within census data, extracting the "gold" requires GIS. Having identified this apparent underutilisation of census data in GIS, this project intends to clarify some of the misconceptions and emphasising the notable benefits derived from census in GIS for decision makers. By taking these two standalone components, census and GIS, this project serves as a critical evaluation between the two by addressing several key aspects of integration.

1.2.1. Problem, Question and Aim

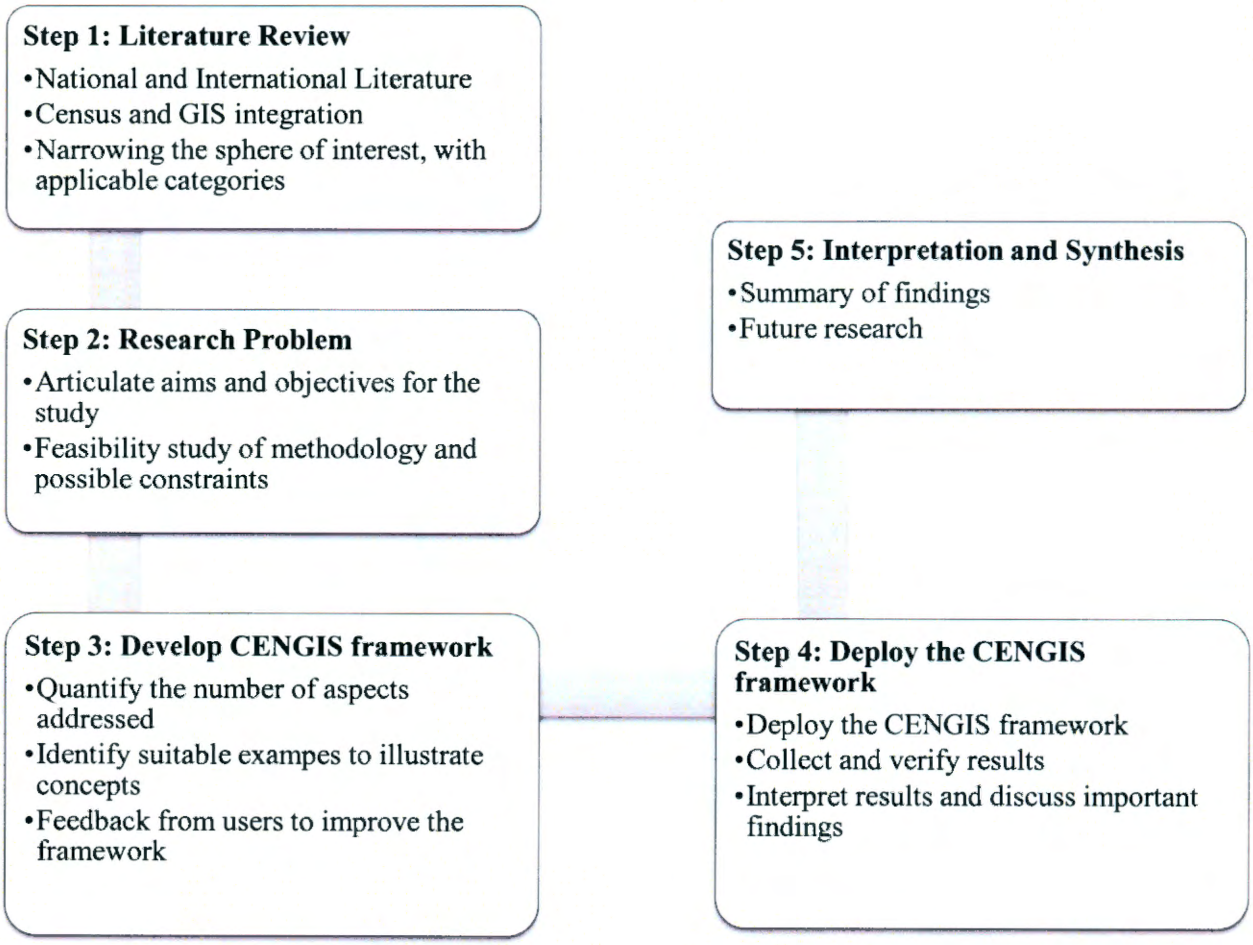
The underutilisation of census data in planning support is evident by the sporadic use of census in GIS. Despite the wealth of information made available to the public free of charge, optimal utilisation of resources has been widely neglected in general. Having identified the apparent gap between census and GIS, this project will serve as a critical **evaluation on the integrative use between Census and GIS for planning support**. To answer this question comprehensively several aspects of integration need to be evaluated. A custom framework called CENGIS (derived from Census and GIS) needs to be developed for assessment purposes. The overarching aim of this research project is to evaluate integrative use between census and GIS for planning support in light of the CENGIS framework. This would clarify the misconceptions between the two and highlight the powerful relationship.

1.2.2. Research Context

In terms of study area, the majority of examples used in the CENGIS framework are purposely confined to the author's home town, Bloemfontein see Figure 1 and 2. Deploying the CENGIS framework in a place people can relate to will improve overall correctness of answers. However, this study is not intended to be geographically bound, but rather a general assessment on integrative use between census and GIS. The CENGIS framework serves as a generic evaluator that can be used in conjunction with different examples, depending on the audience. Another constraint introduced through the CENGIS framework is the fact that it only addresses some of the more predominant integrative principles without extensively looking at minor ones.

To evaluate integrative use one needs to systematically define what and how you measure. All aspects included in the CENGIS framework are derived from the review in chapter 2. There exist myriads of other aspects that serve as integrative indicators of which only the most predominant types would be included between census and GIS. What is important to underline is that the CENGIS framework is designed as a general guideline to evaluate integrative use between census and GIS and can be extended for future research.

1.2.3. Research Methodology



Step 1: Literature Review
- National and International Literature
- Census and GIS integration
- Narrowing the sphere of interest, with applicable categories

Step 2: Research Problem
- Articulate aims and objectives for the study
- Feasibility study of methodology and possible constraints

Step 3: Develop CENGIS framework
- Quantify the number of aspects addressed
- Identify suitable exampes to illustrate concepts
- Feedback from users to improve the framework

Step 4: Deploy the CENGIS framework
- Deploy the CENGIS framework
- Collect and verify results
- Interpret results and discuss important findings

Step 5: Interpretation and Synthesis
- Summary of findings
- Future research

Figure 3 Process flow for the research project

1.3. CHAPTER OUTLINE

Chapter one focused on giving a brief introduction on census and GIS by outlining the apparent gap with regard to integrative use. Furthermore, it introduced the scope of the study, which is to evaluate integrative use through a custom framework between census and GIS. The following chapters will unfold the research aim systematically. Chapter 2 is composed of a literature review on census and GIS, focusing particularly on integrative aspects between the two for decision support. Chapter 3 introduce the CENGIS framework, which is derived from the aspects reviewed in chapter two. Each aspect on integrative use is explained to the reader. Chapter 4 takes the results collected from the CENGIS framework by discussing each aspect on integrative use as derived from the results obtained. Chapter 5 concludes the research by a brief review on the projects aims as mentioned in chapter one and gives a broad summary of the key findings on integration as derived from the CENGIS framework. The project ends with a general summary of the project with recommendations for future research.

CHAPTER 2: UNPACKING CENSUS AND GIS INTEGRATION

2.1. INTRODUCTION

This chapter serves as a twofold review. Firstly, an overview of census from a historical perspective within the South African context as it happened in 1996, 2001 and 2011. In addition to being decennially conducted, census remains one of the most ideal sources of information planning support. Secondly, integration between census and GIS is made possible through its strong relation to geography. The GIS revolution has greatly enhanced the use of census when hand-drawn areal units in 1996 were converted for the first time into their digital counterparts for GIS analysis. A review on key areas of integration between census and GIS is covered in this chapter, focusing on aspects of tabulation, dynamic map making, representation, time series, network analysis and spatial aggregation just to mention a few.

2.2. CENSUS AND GIS OVERVIEW

Census activities in South Africa are conducted under regulation of the Statistics Act No. 6 of 1999. This act ensures that census activities are independent of political interference. It gives statistician-generals the right to collect information they deem necessary for the production and dissemination of official statistics. It is agreed upon within section 17 of the act that the Statistics body of South Africa (StatsSA) will not disclose any information related to an individual, household, business, or any other organisation to protect the confidentiality of all participants. Information is aggregated to minimise the risk of disclosing anyone's identity (Government, 1999, pp. 20-21). The purpose of official statistics articulated in the Statistics Act is to assist planning facilitation for organs or state, businesses and other public or private organisations in planning, decision-making and monitoring of governmental policy (Government, 1999, p. 6).

South Africa conducted fragmented population counts dating as far back as the 18th century. After apartheid South Africa conducted censuses in 1996, 2001 and 2011 respectively. Census information is intended to evaluate the performance of Governmental programs and policies (StatsSAa, 2011, p. 5). Taking 2011 as an example, planning started already in 2003, with pilot studies conducted in 2008 and 2009. The country was subdivided into enumerated areas (EAs), which are roughly composed of 150 households each. Nationally there were 103 567 EAs and 160 000 staff members for census 2011. An estimate of 15 million questionnaires were distributed and processed through scanning to extract information. Post-enumeration studies were conducted to minimise extravagant inconsistencies (StatsSAd, 2011, pp. 1-2).

A census is intended to sample *100%* of the population, whereas surveys only sample a portion thereof. One could readily infer that a 100% count is far better than only a proportional sample. This sets census apart in terms of accuracy, reliability and usability (Peters & MacDonald, 2004, p. 3) for planning, decision-making, monitoring and assessing of policies (Government, 1999, p. 4). However, despite the aspired 100% claim, a 10% undercount is allowed (StatsSAc, 2011, p. 12), which can be adjusted by means of a nationwide post-enumeration survey (StatsSA, 1996, p. 99). Undercount figures can differ significantly depending on gender, age and geographic location (StatsSAb, 2011, p. 13). Despite the undercount, census information remains the most comprehensive baseline for planning in the country (Peters & MacDonald, 2004, p. 4).

Active introduction of GIS in South Africa came in the mid-1980s (Jobson *et al.*, 1986, p. 59), mostly spearheaded by the Stellenbosch Department of Geography who remained the forerunner since 1975 with its expertise in geographical information technology (GIT), especially in the area of cartography, GIS and satellite remote sensing. During the 1990s due to the technological revolutions, Stellenbosch introduced its own independent GIS laboratory, nowadays known as the Centre of Geographical Analysis. H.L. Zietsman is in its own right the founding father of GIS in South Africa pioneering the work in the early 1970s (Liederman, 2015). Since the 1980s trained staff, geographical datasets, private companies and software developers have steadily emerged to make GIS a means for development in South Africa (MacDevette, 1993, pp. 18-19). Datasets in South Africa range from demographics, education, soil, climate, electrification and infrastructure amongst others. Since the 1990s GIS became part of mainstream planning for Government in relation to census, water management, agriculture, environmental management, health care, forestry etc. In terms of the private sector GIS is used extensively in siting a franchise, logistic operation and mining (MacDevette *et al.*, 1999, p. 914).

2.3.CENSUS AND GIS INTEGRATION

2.3.1. General aspects on integration

The emergence of GIS over the past few decades became the most powerful contributor to spatial planning. GIS is revolutionising all planning related activities (Chapin, 2003, p. 1). Since the 1990s GIS became more widely adopted (Felke, 2014, p. 1), featuring in numerous journals (WeiWei & WeiDong, 2015, p. 1). Utilisation of this technology has been limited, partly due to the late introduction of GIS into educational curriculums since 2003 (Felke, 2014, p. 1). Another reason is attributed to GIS's quantitative orientation, which is not suitable for qualitative research. This trend is slowly changing as GIS progressively moulds into a more versatile technology (WeiWei & WeiDong, 2015, p. 1). The rapid expansion of Geographic GIS into socio-economic sciences is a proof of this. Furthermore GIS enables evidence based decisions in critical areas for intervention such as poverty reduction (HSRC, 2011, p. 1).

Since the 1990s governmental adoption of GIS has grown steadily, with more and more municipalities including data analysis into their core workflows of planning (WeiWei & WeiDong, 2015, p. 1). As part of the United Nations development plan, they helped 40 of the poorest countries in the world to gain access to GIS technology for strategic planning. Globally, census has been one of the main areas that has benefited from the adoption of GIS technology (HSRC, 2011, p. 9). The census mapping systems were already utilised by Japan in 1991 and Israel in 1995, which enabled census data to be georeferenced even to the extent of a dwelling. Another program launched in 1997 for Africa, the GeoSpace program, established National Statistical Offices (NSOs) in 15 countries to provide census mapping solutions (HSRC, 2011, p. 10).

GIS mapping in South Africa is a relatively new introduction, especially in relation to census. For example, prior to 1996 EAs were hand-drawn; it was only in 2001 that the EAs were captured digitally. This dynamic transition formed a strong underlying basis for data capturing that could be referenced and queried geographically. The 2011 census excelled at using GIS in the census workflows throughout the planning and pre-enumeration phase. Satellite imagery of France called Satellite Pour I'Observation de le Terre (SPOT 5) of 2008 was used as a reference to draw the EA boundaries for the 2011 census. In addition to the digital EAs, SPOT 5 imagery facilitates in the capturing of Dwelling Frame Units (DFUs) dataset of the entire country (HSRC, 2011, p. 12).

The importance of the geographic frame for census has been elevated extensively since 1996 (Lehohla, 2005, p. 4). Statistical representation of census attributes across space cannot be overemphasised in terms of application power. Decision makers need to know *where* to focus in terms of investment and development (Lehohla, 2005, p. 3). The term statistical geography has become popular since 2001, with different geographic layers made available to the public. Introduction of the Small Areas Layer (SAL) in 2011 improved the accuracy of spatial analysis exponentially. The next revolution would be to move from EAs to Dwelling Frame Units (DFUs) which captures statistics on micro level; producing more reliable statistics (Lehohla, 2005, p. 4).

2.4.1. Elementary aspects on integration

Census data is collected on the basis of individual households. StatsSA ensures the confidentiality of participants by taking the appropriate steps to ensure that tabulated data will not reveal the identity of individual participants. To ensure confidentiality, census data is aggregated over a particular geographical area and averaged (Peters & MacDonald, 2004, p. 22). It starts at dwelling frame points and aggregates into enumerated areas, small areas, suburbs, towns, municipalities, districts and ultimately, provinces (StatsSAd, 2011, p. 5). Software such as SuperCROSS allows diverse tabulation methodologies where users can recode values according to selection for subsets of new labels. Several calculations can be performed on tabulated data, such as column and row totals, percentiles, pareto, variance, asymmetry and skewness (SuperCROSS, 2012, p. 19).

Other than tabulation, census data needs to be displayed through a GIS. From the 1980s GIS focused primarily on two key issues, of which one was automated map making (Burrough, 2001, p. 361). Census lends itself toward extensive cartographic output. England was one of the first countries to use automated map production with census data in 1981, effectively transforming nearly 3.2 million numbers into 580 statistical areas of choropleth shaded maps (Browne &

Fielding, 1987, p. 82). One of the main concerns for statistical representation in map form is homogeneous regions which cannot reflect extreme heterogeneity of variables adequately, as observed on the ground. To keep sample population size relatively even, the size of the delineated area would grow bigger in less populated areas and smaller in urban areas (Browne & Fielding, 1987, p. 83). The importance of scale is another factor that influences representation. This graphic conflict with regard to cartographic representation of census data can be addressed through generalisation to solve the representational conflict induced by scale (Ware *et al.*, 2003, p. 296). Manual map generalisation is intrinsically still a cartographer's work. Until now automated census mapping was still being questioned for reliability, with ongoing research being conducted (Steiniger, 2007, p. i) for identifying rules, which is translated into generalisation processes and algorithms to deal with each map representation scenario (Steiniger, 2007, p. 6).

During map making the most time-consuming task is annotation. Labelling of geographic entities takes time and automation seldom does justice to the representation of the data (Freeman, 2005, p. 287). Usability of maps depends on clearly annotated features; although it seems simple the task is indeed complicated. Labelling should clearly articulate the spatial relationships clearly (Freeman, 2005, p. 289). Labelling area features in census requires consideration on the shape and extent of each feature. Placement of labels should ideally fit inside the area for recognition. To automate map production, text-placement strategies that adhere to cartographic norms and standards need to be implemented (Freeman, 2005, pp. 290-291).

Displaying census data visually needs to be done in a manner that is cartographically acceptable (Burrough, 2001, p. 363). When large amounts of data can be displayed graphically, spatial patterns and relationships should be clearly articulated (Koua & Kraak, 2004, p. 1). Statistical representation of data, such as census, is a powerful analytical tool for decision makers. Despite textual and numerical analyses, governmental policy and planning rely extensively on visualisation of census data. Usefulness of different cartographic depictions of the data needs to be evaluated and adjusted based on the intended use (Manan & Hashim, 2010, p. 367). Change detection is often clearly visible through spatial representation. Visualisation can be done in numerous ways, one of which being different colour tones (Manan & Hashim, 2010, p. 373). GIS is the most reliable medium with which to visualise census data because of its ability to directly link aspatial data (census data) to spatial data (census boundaries) (Manan & Hashim, 2010, p. 376).

According to Monmonier (1991), all geographic representation contains some form of "lie". For example, entities are represented by symbols that are always larger than their real world footprint. The mere fact that a spherical globe needs to be portrayed on a two-dimensional surface (i.e. map) gives room to distortion, which will always represent a selective and incomplete view of reality. However, the degree of misrepresentation varies from negligible to seriously wrong representations. A cartographer's skill is essentially to know where to "draw the line" in terms of the information they want to convey (Monmonier, 1991, p. 1). The mere fact that you can produce infinite variations of the same map using the same data should make users aware that cartographic representations are biased. Not to mention the political influence on shaping public opinion through maps by suppressing contradictory information and using dramatic symbolism (Monmonier, 1991, p. 86).

Knowing the three factors of representation - classification, generalisation and symbolisation - is of critical importance. Classification can produce an infinite number of varieties; it is inherently a creative process and nothing else. There is no clear and absolute method on classification of data (Dodge *et al.*, 2011). Classification introduces order and coherence in the data. Both the purpose and method used need to be evaluated for their constraints. The need to choose appropriate methods for the intended purpose is of utmost importance. Apart from the classification scheme (equal, defined, exponential, manual, quantile, natural or standard deviation), the number of intervals are equally important. Too many intervals may limit distinguishability of data. The choice of symbols placed over a choropleth surface that varies in size depending on the chosen attribute give a good illustration of data variance; however, if extreme values exist smaller symbols may be "swallowed" by bigger ones. Symbols can, however, lead to difficulty in interpretation if the audience is not skilled in cartographic representation (Chainey & Ratcliffe, 2013). Another useful representative means is dots inside a census unit (polygon) that represent the value of dot count within the boundary. Colour variation is seldom necessary for dot density for population estimates (Elangovan, 2006, p. 108).

2.4.2. Intermediate aspects on integration

Census data are collected on individual household level, but available in aggregated format (HSRC, 2011, p. 16), which summarises the samples and averages each across the enumerate-area (Peters & MacDonald, 2004, p. 21) . To ensure confidentiality individual entities need to be aggregated before dissemination (Reidl *et al.*, 2006, p. 900). Aggregation does not portray social activity accurately, and should only be used as very indirect indicators of behaviour. Furthermore, detail is further obscured when normalisation is applied. Census representations elaborate more on the shape and size of the enumerated area that of people actually living and working in them (Reidl *et al.*, 2006, p. 906). Depending on the level of spatial aggregation, disparities can be hidden, for example, population growth is seen on a higher level of aggregation, yet the underlying lower level shows numerous areas of population decrease (Paez & Scott, 2004, p. 58). Aggregation bias can be adjusted by means of a matrix transform, such as correlation and regression analysis (Paez & Scott, 2004, p. 59). What spatial aggregation inevitably causes is a disregard of heterogeneity of underlying samples. The mere fact that census geographies are made of spatial units shows that different areal units will produce different results during analysis (Dumedah *et al.*, 2008, p. 48). According to Openshaw (1984), no sound alternatives to managing aggregated data in a statistically sound framework. The scale and shape of the areal units influences any spatial analysis. It is recommended to compare results from different spatial resolutions to clarify the data (Jacobs-Crisioni *et al.*, 2014, pp. 52-53). It is indeed difficult to predict aggregated elements of coarser resolution, since they follow a stochastic pattern. The shape effect exists due to irregular delineation of spatial geographies that cannot fully account the heterogeneity of the underlying population (Jacobs-Crisioni *et al.*, 2014, p. 53).

Using aggregated spatial data with pre-defined areal units such as census creates a well-known issue called Modifiable Areal Unit Problem (MAUP). Studies have been conducted on the MAUP from the 1930s, but only became of real concern since the 1960s and 1970s. Despite the research conducted results remain vague on how the MAUP influences univariate, bivariate and multivariate statistics (Dark & Bram, 2007, p. 472). The boundaries are the source of the MAUP (Reidl *et al.*, 2006, p. 900). Several analytical techniques are affected by the MAUP such as regression and relation analysis, spatial interaction and location-allocation modelling (Paez &

Scott, 2004, p. 58). Boundaries can be infinitely modified making the MAUP unavoidable. This arbitrary subdivision of areal units for the purpose of aggregating data is known as the MAUP (Jacobs-Crisioni *et al.*, 2014, p. 48) (Manley *et al.*, 2006, p. 144). The direct result is variation in derived answers if different areal units are used. Both scale and zone are inherently related to the MAUP. The irregular size of spatial areal units in census geographies makes the MAUP unavoidable (Dumedah *et al.*, 2008, p. 48). Outcomes are always dependent on scale and shape aggregation. Despite the extensive literature on the MAUP, no clearly defined solution has come of date yet (Jahanshiri, *et al.*, 2015, p. 47). Where data gets aggregated into different sizes or shapes, the aggregation problem occurs. The zonation effect is caused by the grouping of smaller areal units into larger ones (Dark & Bram, 2007, p. 472). To address the scale problem the use of an optimal zoning system is recommended to create homogeneous units. Despite the effort to minimise scale variability in analysis, the results still remain biased (Dumedah *et al.*, 2008, pp. 48-49).

The main concern with census is that data gets collected on non-modifiable entities (households) and aggregated into modifiable units (census boundaries) for reporting. It is not possible to create ideal census geographies that take all spatial scales and processing into account (Manley *et al.*, 2006, p. 159). Misrepresentation is inevitable. One way of minimising this modification and producing more homogeneous zones of data would be to down-size the areal units. This effect is shown where an 800-unit dataset showed a 10% increase in the elderly population cause a $308 decrease in family income; however, with 25-unit dataset a 10% increase produces a $2,654 decrease in family earnings (Prouse *et al.*, 2014, p. 66). This is quite a significant margin of error. The MAUP is especially problematic in demographic studies such as census when choropleth maps are used to visualise data. Thematic mapping is known to grossly misrepresent the "ground truth" of social and economic variables. Just the mere fact of an abrupt change when moving from one boundary to the next illustrates the shortcoming of zone based statistical representations (Reidl *et al.*, 2006, p. 900).

According to Openshaw (1984) the effect of the MAUP could be limited in the census by identifying the appropriate scale for spatial analysis for display. However to work around the MAUP is possible if the individual counted entities are analysed apart from aggregation (Dark &

Bram, 2007, p. 477). The implications of using census data depicted in choropleth cartography and thematic mapping has a significant effect on policy. Census geographies are often politically labelled based, on the assumption that the representation is accurate, which results in intensity either being over or underestimated (Reidl *et al.*, 2006, p. 901). Census data has long been used to formulate public policy for public fund distribution; however, the fundamental flaw associated with such use is that policy makers assume that census areal units are fit for the intended purpose. For example, identification of poverty hotspots is not arbitrarily possible with census, because the geography of poverty has little or no correlation with census areal units. Poor people can be found randomly in areas seen as rich; thus census gives only a distorted view of reality (Reidl *et al.*, 2006, p. 902).

Apart from the MAUP, another concern with decennial census data is time. It is recommended that the census intervals be changed from every 10 years to a more continuous measurement. Using decennial data for trend analysis is not effective because most of the important variability is simply ignored (Salvo & Lobo, 2006, p. 226). In South Africa this gap between the census of 2001 and 2011 was breached with a Community Survey in 2007, with the next one planned in 2016. These types of sub-census programs provide data on municipal level but not on the small census geographies as recorded in the full decennial census every 10 years (Radebe, 2015). There is, however, a positive use of historical census data. Firstly, it allows for meaningful comparison because data is georeferenced. Secondly, data can be visualised and animated. Lastly GIS assists in spatial analysis of coordinate locations of the census features (Gregory & Healey, 2007, p. 639). Because census data is collected spatially, this component makes historical analysis of census optimal for spatial comparison or trend analysis. Data can be joined back to the former boundaries and captured digitally in GIS for temporal analysis (Gregory & Healey, 2007, p. 640).

Having historical data enables the users to layer different time periods and study relationship across different categories. Real insight into local patterns of distributions, such as race, can be determined using historic census data (Gordon, 2011, p. 10). Some hurdles encountered through historical GIS are the reliability of names and numbers used between census dates. Spatial and attribute precision are two factors that influence the comparability of different census datasets

(Southall, 2011, pp. 150-151). If high variability of census boundaries occurs at sublevel, such as enumerated areas, data can always be analysed for spatial temporal analysis using higher-order data, such as municipal boundaries (Masser *et al.*, 1996, p. 91). Census units are subject to boundary shifts, which will acquire additional techniques to ensure continuity and quality of time series within census data (Nyerges *et al.*, 2011, p. 38).

2.4.3. Advance aspects on integration

Decomposition of population distribution estimates is a common problem. Several methods of decomposition have been developed for census. As mentioned by Wu (2008), for various reasons people might need to estimate population not based on census boundaries. Areas might be smaller or even irregular in shape, such as a population living within a flood risk area, or number of people within a certain distance of some transport network, i.e. road (Wu *et al.*, 2008, p. 122). By means of raster representation of census vector boundaries can be converted using pixels, representing the original value within the zone (Spiekermann & Wegener, 1999, p. 1). Methodologies used to decompose census vector data is real weighting, pycnophylactic interpolation and dissymmetric mapping. Weighted interpolation is essentially the most common form of interpolation which takes a regular grid, intersects it with the underlying census boundary, and assigned the value based on the proportion of the census boundary contained within each cell. However, this method applies the assumption of uniform distribution of population within the demarcated census zone. Gridded population sets are quite common, such as the Gridded Population of the World (Sheckhar & Xiong, 2008, p. 882). Zones need not necessarily be connected to be summarised (Frank, 2005, p. 202). Zonal statistics essentially summarise the data from a underlying raster based on an overlying zone. Various statistics can be calculated for each zone where the user can specify which operation to use, such as mean, median, max, min, standard deviation, variance, count or sum (Bahgat, 2015, p. 136).

Apart from zonal statistics, threshold and capacity estimates are another crucial planning tool. Provision of social amenities, according to the Council of Scientific and Industrial Research (CSIR) provide accepted norms and standards for travel distance to social amenities. These services and amenities are classified according to population density, which in turn determine the acceptable distance and coverage area (CSIR, 2012, pp. 11,24). Census data is unfortunately the only available means to ascertain these requirements, and GIS offers the means to do so (Gibson *et al.*, 2011, p. 247). The use of georeferenced data enables the calculation of population estimates within a prescribed distance. Geographic access to services can only be done reliably with GIS. To ensure that the distance calculated is not as the "crow flies", but measured according to topography, the network analysis function is employed. Since the distance between two points is always longer than a straight line, it requires network analysis to give reliable estimates of population estimates within a specified distance of each facility. The network model is the most popular conceptual model to represent a network i.e. roads within a GIS environment. Networks are composed of nodal points and connector polylines. Nodes are one-dimensional entities and polylines are two-dimensional entities. This ensures the topological integrity of the modelled network. Relations of nodes and polylines are stored in a database; this is to ensure the right attributes' associations with each entity, such as speed, elevation, road type, etc. (Fischer, 2006, p. 45). The service area function in network analysis calculates the linear distance road-wise from predefined locations. Service areas can be constructed from individual points or areas. The only requirement to generate a service area is a predefined location, a threshold distance and an underlying network topology. The accuracy of a service area in network analysis depend on the quality of the modelled roadways, directions, connectivity and barriers (Oh & Jeong, 2007, pp. 28-30).

Lastly, zonal representations of statistical data take all attributes within the zone and distribute it uniformly throughout the zone. However topological relationships and complex socio-economic activities are oftentimes ignored which leads to serious methodological problems during analysis (Openshaw, 1984, p. 1). The so-called "strait-jacket" assigned to zones captivate it under the inherent weaknesses attributed to zone-based analysis. Spiekermann and Wegener (1999) refer to this phenomenon as the 'tyranny of zones'. A combination of vector and raster representations can be used in a disaggregating model to overcome the disadvantages of zones. Interpolation can disaggregate zonal data for micro-scale analysis (Spiekermann & Wegener, 1999, pp. 2-3). To facilitate the process, disaggregated data is required. If no micro scale spatial data is available GIS can be used to generate probabilistic disaggregated spatial data based on zone data. To disaggregate zone-based data, such as census areal units, the land use within the zone needs to be taken into consideration. A combination of raster and vector representation using disaggregated spatial data, such as land parcels or transport network, allows for a powerful reorganisation of data on micro scale. Generating artificial sub-block areas and using it for estimating population within an overlying zone is relatively accurate (Wu *et al.*, 2008, p. 121). As the number of sub-blocks increases, so does the margin of error. Estimation of population size often times does not coincide with census zones. Governments might need to estimate the number of people living in a flood-risk area, which will obviously not correlate with census boundaries. Estimation of population within a custom distance forms a single location, or a corridor renders census zone inadequate for the purpose (Wu *et al.*, 2008, p. 122). Population estimations are generally done in three ways: those done based on census zones, inferred population based on physical or socio-economic variables, or disaggregated census unit populations into sub zones. In the end, detailed land use data will essentially improve disaggregated data reliability when choosing to subdivide data for population estimations.

2.4.CONCLUSION

As discussed, the strong relationship between census and GIS is due to its geographic component. In the section covering a brief overview on census, the extensive coverage of census sampling is unparalleled in comparison with other surveys. It remains one of the most reliable baselines for evidence based decision-making. Firstly, integrative use of census and GIS is quietly causing a revolution in planning, since the government's adoption of GIS in the early 1990s. Conversion of the hand drawn census boundaries into their digital counterparts laid the foundation stone for spatial analysis. Tabulation of data in third-party software greatly improves the use of census in different planning scenarios. In addition census takes full advantage of dynamic map making, reducing the overall time needed on generating informative cartographic answers from census with different spatial representations. Besides the fact that census data is aggregated to hide participants' identities, the availability of the SAL greatly reduces the long standing problem of the MAUP, which is especially prevalent when comparing previous census data with newer ones. Although census data is disseminated in census areal units, GIS can reliably free census data form the tyranny of zone through zonal statistics and disaggregation to greatly improve its usefulness for micro scale analysis.

CHAPTER 3: FORMULATING THE CENGIS FRAMEWORK

3.1. INTRODUCTION

Development of the census and GIS framework, known as CENGIS, takes ten aspects of integration into account. To evaluate integrative use one needs to systematically define what and how you measure. All aspects included in the CENGIS framework are derived from the review in chapter 2. In addition to the ten aspects chosen for evaluation, there exist myriads of other aspects not included in this research project for obvious reasons. The CENGIS framework essentially focuses on the more important integrative uses between census and GIS. What is important to underline is that the CENGIS framework is designed as a general guideline to evaluate integrative use between census and GIS and is not the holy grail of assessment. Besides the given examples, concepts discussed in this section can apply to datasets outside the vicinity of census.

Each aspect of integration is evaluated through a brief definition or description on the concept assessed. Some of the aspects are more common to the average user; whereas other aspects may require additional techniques used by more experience users such as conversion of census data into raster datasets for map algebra. Besides evaluation on integrative use, the CENGIS framework is intended to create, amongst census users, some awareness of the vast possibilities effective integration offers for decision support. Although possibilities are endless, constraints nevertheless needs to be properly addressed in a manner that does not undermine decisional accuracy as seen in the spatial analysis *"crimes"* as mentioned in chapter 2. The CENGIS framework places emphasis on some common pitfalls experienced by census users in the GIS environment, such as spatial aggregation, modifiable area unit problem, disaggregation, comparability, and representation. Hopefully the CENGIS framework can assist users in future to minimise misrepresentation by adhering to accepted norms and standards as prescribed in cartography.

3.2. CENGIS FRAMEWORK

3.2.1. Extraction Standard

The first aspect of integrative use between census and GIS starts with the basic relationship between census and corresponding spatial entities. Most users have access to tabulated data through online access or private standalone software such as SuperCROSS. The frequent use of census data in governmental reports and academic research mostly originates form users simply using pre-tabulated data from an existing source, be it a website or spread-sheet. When assessing the use of census variables, one seldom finds sophisticated tabulations done by the users.

Census variables are geographically referenced through geographic codes depending on the spatial level queried. Defining the geographic extent and level of detail is the initial step before tabulating variables. The following spatial layers are made available for tabulation:

- Provincial (9 features)
- District (52 features)
- Local (234 features)
- Ward (4227 features)
- Main Place (14039 features)
- Suburb (22108 features)
- Small Area (84907 features)

After defining the geographic extent and level of detail as to geographic extent, it is best to use codes, instead of names, as unique identifiers to join the census tabulation back to the spatial layer based on that unique value. An example of a unique identifier in census works as follows: Free State (value: 2), Municipal (value: 210), Main Place (value: 211), Sub-Place (value: 211001), Small Area (211001001). As seen, depending on the spatial extent, the unique identifier would increase if the geographic area becomes smaller and is lower than the spatial hierarchy. This unique identifier serves as the fundamental link between the tabulated data and the census spatial layer. The census data is then simply "joined" based on this unique identifier and used in any GIS software package for display see Figure 4 for illustration. Maps produced from census data oftentimes portray a picture in a much more comprehensive way than the conventional way of displaying census data in graphs, tables and charts.

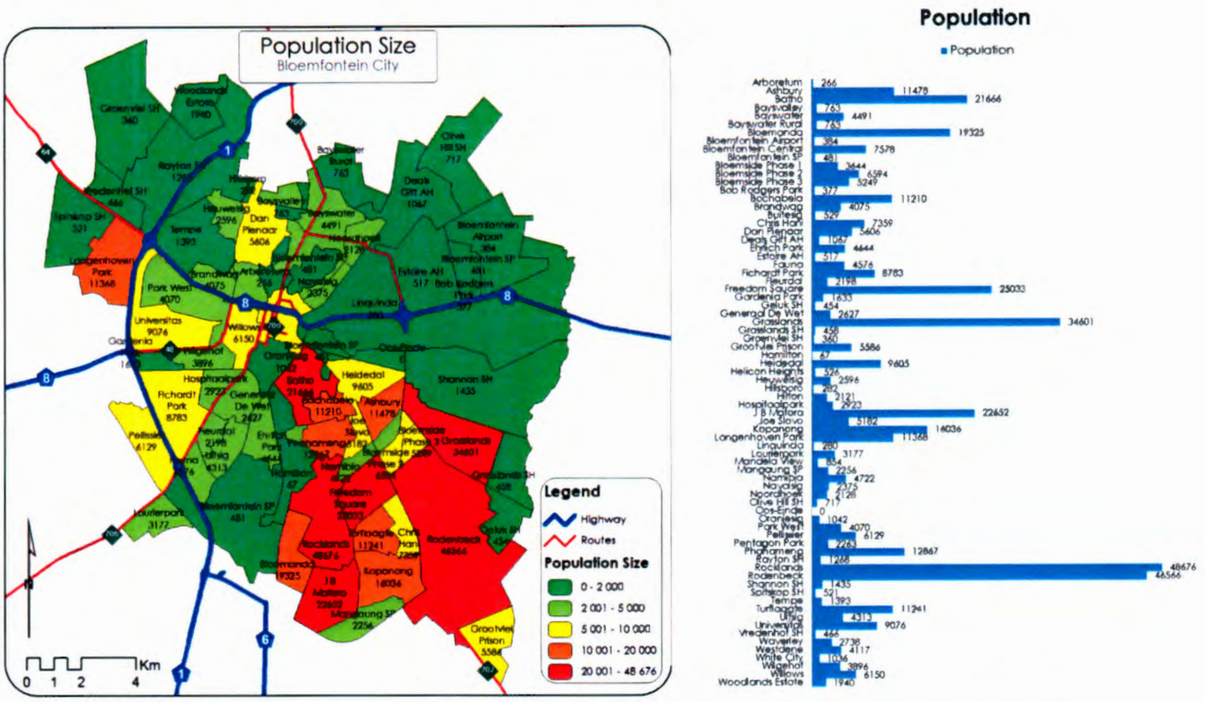


Figure 4 Population size per suburb (left) with the same data in graph format (right)

3.2.2. Extraction Customised

Tabulation through SuperCROSS can be seen as the innovative way to summarise data with queries. Census data is classified according to category such as descriptive stats, dwelling stats, family stats, household services stats etc. Each category contains several tables pertaining to that category to address a vast combination of questions from that specific category. Depending on the level of spatial detail, tabulation can be done from provincial level to small areas which are tinier than suburbs. What tabulation essentially allows the user to do is to build complex queries with multiple criteria, such as: *How many households own a refrigerator, vacuum cleaner, washing machine, computer, motor-car, television, cell phone and have access to the internet within Bloemfontein city*. Querying the census data in innovative ways can essentially answer this question as depicted in Figure 5. Once a census user realises the ease of tabulating custom queries by recoding field values either individually or collectively it opens up a whole new potential for GIS integration. Integrating custom queries from census data into GIS allows for rapid visualisation for decision support. In general the use of custom tabulations from census in GIS is seldom seen or utilised, yet this aspect offers a rich supportive function, especially to those in governmental authorities.

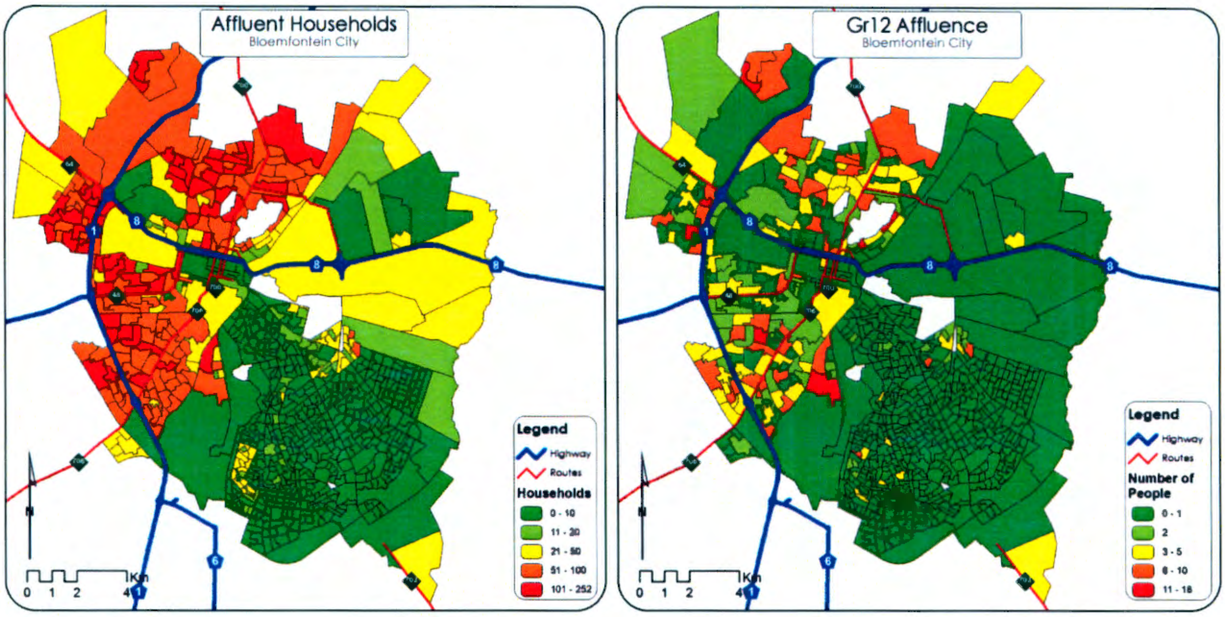


Figure 5 Custom tabulation for affluent households in Bloemfontein (left) and Gr 12 earning more than R25 000 per month (right)

3.2.3. Map Production

After tabulation is completed and joined to the corresponding spatial entities, map production in GIS is relatively straight forward. Over the years GIS developers worked on ways to automate cartographic map production, which require little customisation from the user's side. In South Africa the world's leading propriety software called ESRI continues to dominate the GIS market. ArcGIS includes by default an automate map function called data-driven pages, which essentially loops through a prescribed list of features using a unique identifier. All census spatial boundaries have a unique identifier, making it easy to do automated mapping for all features in census data, regardless of the spatial hierarchy, be it provincial, municipal, ward, suburb, etc.

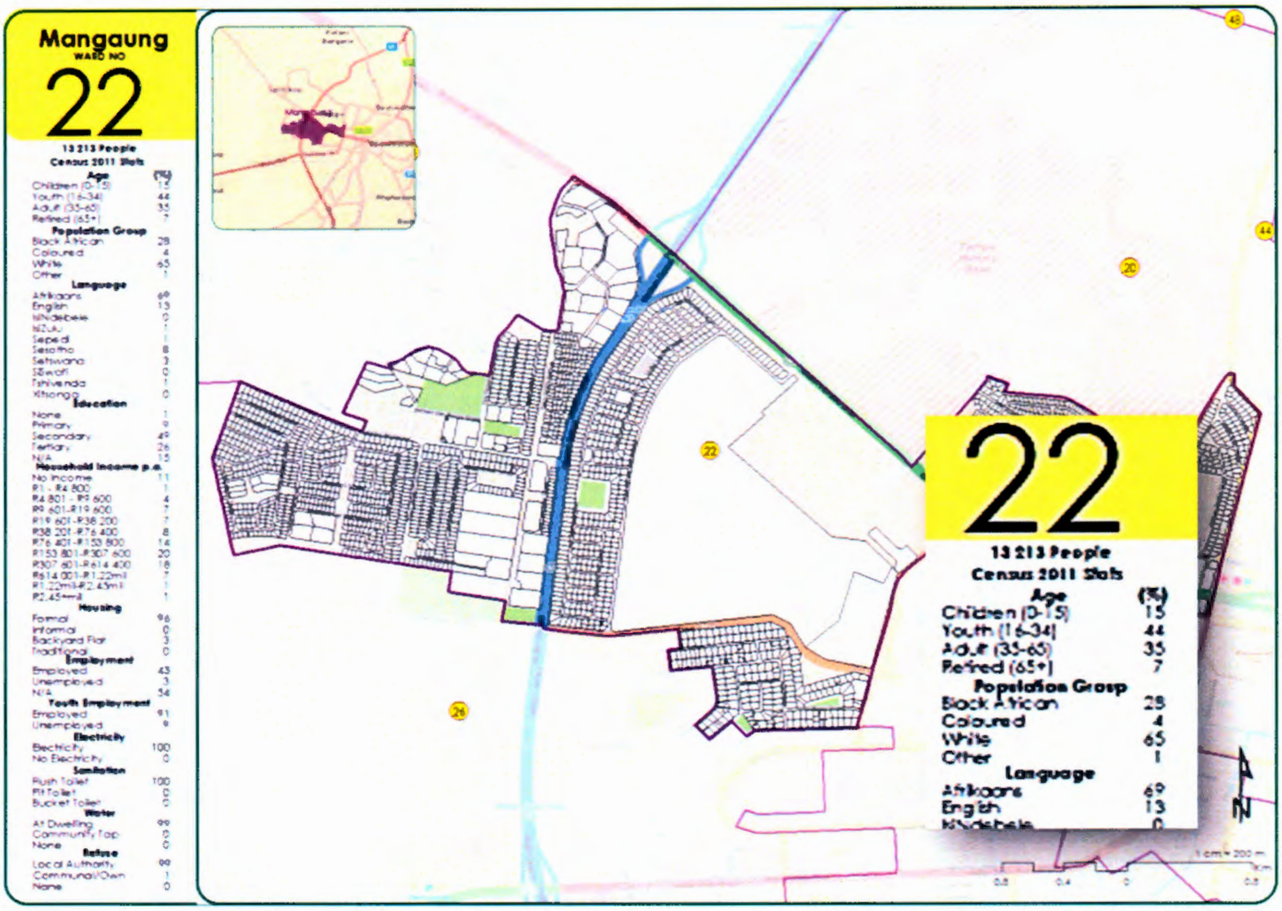


Figure 6 Dynamic ward map generation using census boundaries and stats

If set up correctly automated mapping virtually reduces time exponentially. Map functions include custom displays of an area's extent. Dynamic attributes such as area name, and any other variable associated with that feature can be dynamically updated for each map when generating a series as seen in Figure 6 above. Dynamic attributes enable census users to include vast amounts of information supplementary to the map. For example, creating a ward profile map for every ward in the Free State province would total 317 individual maps. With data driven pages users can have custom attributes assigned with each map such as population count, language percentage, population group, age etc. generating quick and informative maps within a short period of time. Besides the functionalities provided through dynamic map making such as data driven pages, census users seldom utilise this function.

3.2.4. Representation

Cartographic depictions of census data is seldom questioned and regarded as authoritative. As mentioned in chapter two, people are remarkably ignorant about the number of variations a cartographer can generate from the underlying data. Classification of census data is by default univariate and done on the fly. After tabulation of census variables, data needs to be formatted for representation. Classification introduce some form of order into the data that needs to adhere to the intended purpose or use of the data. Using different methods of classification will change the representation accordingly. Among the classification methods used, "Jenks" classification, which is also the default, is the most frequent classifier used in census representation. However, classifiers such as quantile, manual or equal (Figure 7 top left), geometric or standard deviation, might be more suitable, depending on the application. The problem induced with classification is that you can virtually render an infinite number of variations from the same underlying data. For example, depicting poverty in a range from green (not poor) to red (poor), the classification classes one can manipulate to either increase or decrease the visual representation of poor people.

Intervals on the other hand are an abrupt change from one class to another. In general, more intervals would produce more subtle variations in colour, making discernment more difficult to the user. In cartography, going beyond five intervals for a ramp colour is too much (Figure 7 top right), and less than three is not useable. The problem associated with interval count is that there is no concrete guide for choosing the number of intervals. This gives a cartographer room to change the representation of census data by only adding or subtracting intervals.

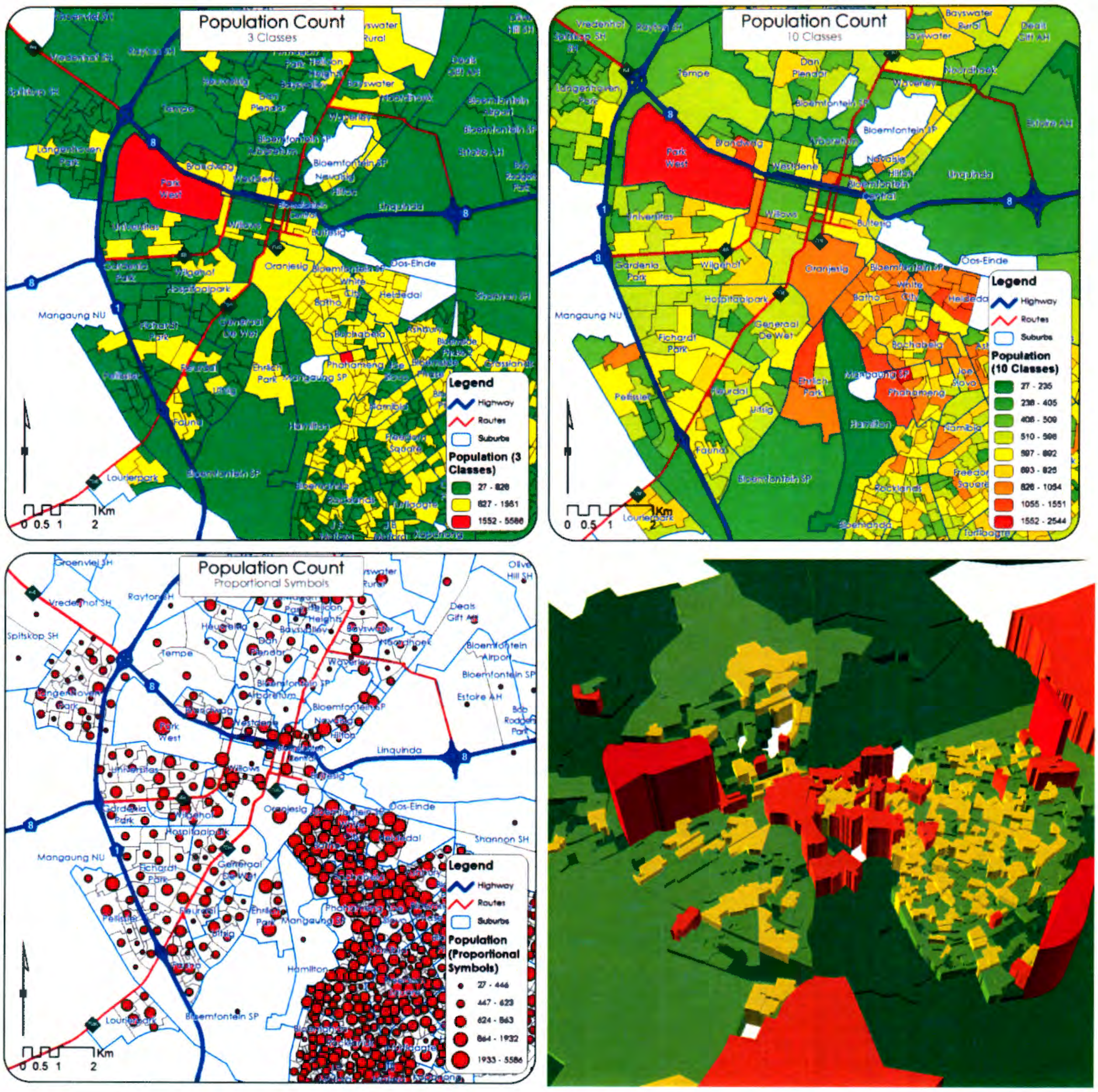


Figure 7 Difference between three classes and ten classes (top), same data with symbols, or in 3D

Apart from colour, census data can also be illustrated with symbols. For example, demographic size can be depicted using circles varying in size on a map to illustrate population density. The advantage of symbols is interpretation. However, just as classification with regards to colour for symbols is subject to the user's bias so is symbols size as seen in Figure 7 bottom left. This inevitably leads to misinterpretation by the user during decision making. Besides symbol size, the number of symbols, such as dot density is another way to illustrate density or census values within a census unit. Census can be represented in numerous ways using different classification methodologies, intervals and symbols; however, the problem with all three is their authoritativeness, which can lead to gross misrepresentation of the actual ground truth. Depending on the application of the census data, the pros and cons of each classification scheme, interval count and symbol type in each scenario one needs to carefully consider. This consideration is often times overlooked, however the direct influence of that representation is enormous.

3.2.5. Spatial Aggregation

Data aggregation is a fundamental principle that influences the use of census data. The nature of census data collection is done per household called geo-referenced dwelling frame points, which are then aggregated into enumerated areas to protect participants' safety and security. Individual dwelling frame points are aggregated hierarchically into enumerated areas, small areas, sub-place, main-place, municipal, district and provincial is done sequentially. Census makes six of these aggregated layers available for dissemination, of which the smallest is called the SAL. The lower layers on the aggregation pyramid are useful for strategic intervention. Moving up the pyramid, higher order entities serve as strategic indicators for decision makers. For example identifying the poorest areas in the Free State, the principle of aggregation can be used to help solve the problem. Starting from strategic: identify the poorest district, then the poorest municipality, then the poorest place, then the poorest suburb, then the poorest small area as illustrated in Figure 8. Using different levels of spatial aggregation greatly facilitates the decision making process.

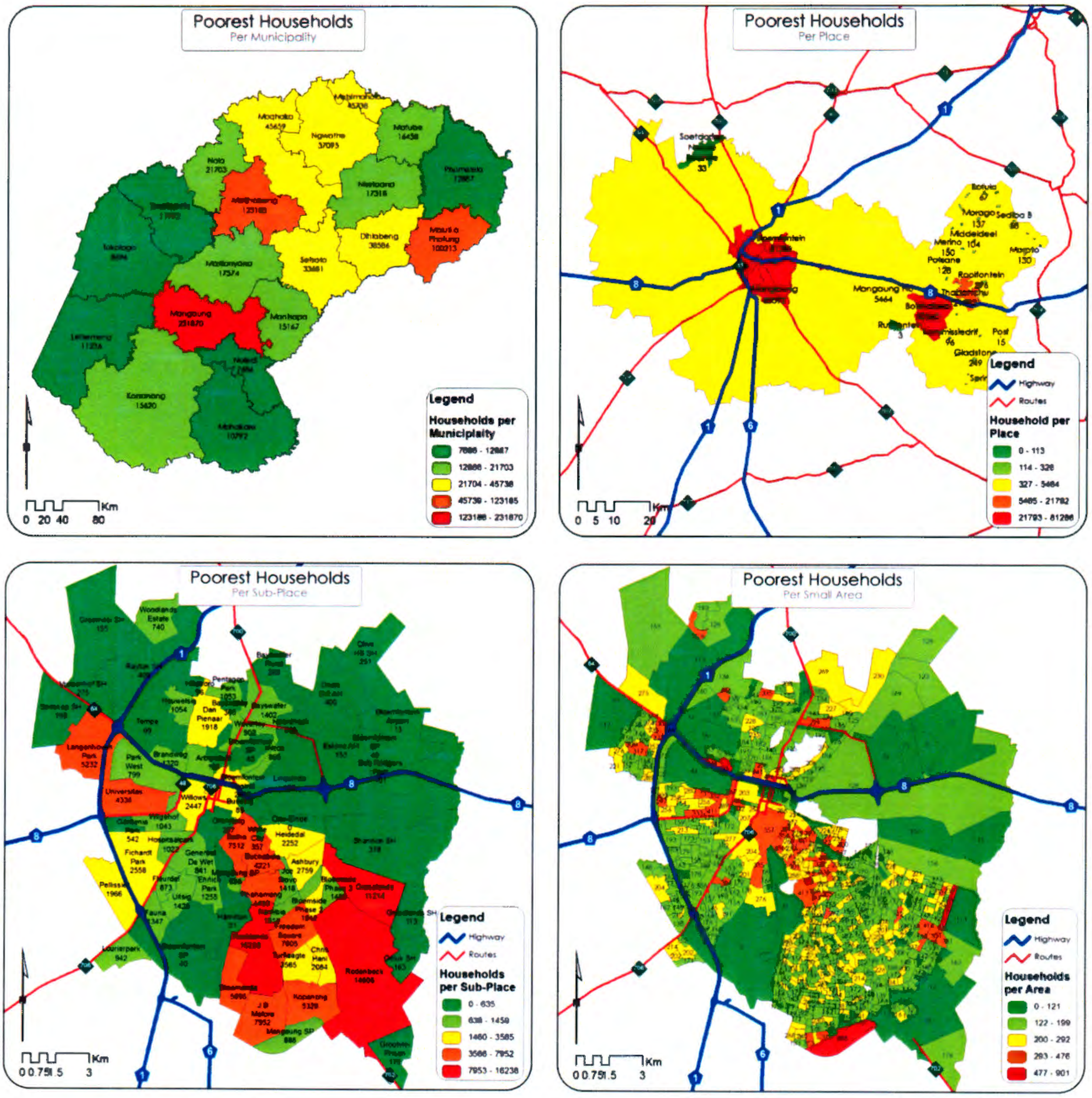


Figure 8 Municipal level (top left), town level (top right), suburb level (bottom left) and small areas (bottom right)

3.2.6. Modifiable Area Unit Problem

Census data unfortunately suffers from a fatal illness diagnosed as the Modifiable Area Unit Problem (MAUP). The fact that census data needs to be aggregated into predefined area units leads to a serious problem in terms of representation. Boundaries are not absolute and can be modified infinitely, including or either excluding certain areas. For example, the number of samples within area A is two and area B is three. However modifying the shared boundary between the two can change the samples within are A to three and area B to two. All census variables are boundary dependent, which means all values are relative to the demarcation chosen. Besides census data, this problem remains yet unsolved in many GIS applications.

Census boundaries are not fixed and often change political transitions. This variability in census is a real concern for the validity in terms of accuracy. For example comparing ward statistics, it is often found that boundaries have been shifted 10 years down the line, defeating any meaningful comparison (Figure 9). The only way to really get rid of the MAUP is to utilise the dwelling frame points, which is, however, a breach in privacy and would not be made available for public due to the privacy constraints (Government, 1999, p. 20).

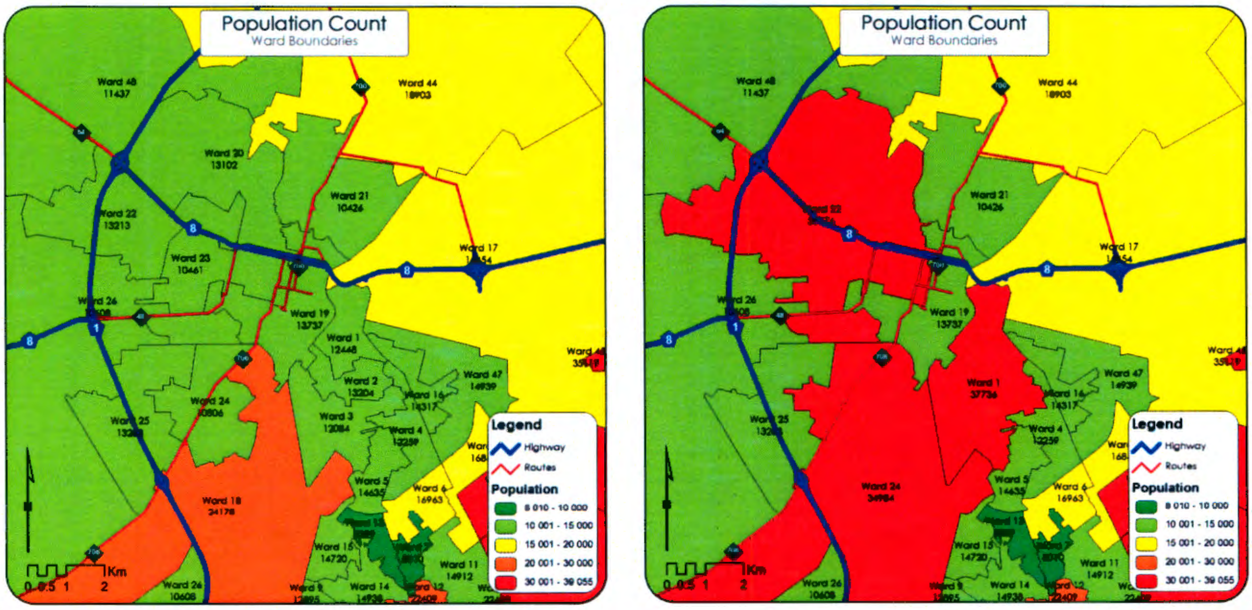


Figure 9 Original boundaries (left), modifying the boundaries cause values to change (right)

Users of census data seldom consider the implications of the MAUP, which is a known weakness of any aggregated dataset. Disregard of this problem has led to many unjust applications and wasteful expenditure of resources especially in governmental decision making. The only way to minimise the MAUP is to decrease the boundary size, however, no matter how small, if aggregation is still applied, the MAUP will always be present.

3.2.7. Time Series

With census the need to compare different datasets from different times has always been very much sought after. Census occurs every ten years and a sub-census is performed every 5 years to minimise the gap on trend analysis. However users seldom used sub-census variables and prefer to compare decennial census data because of its reliability. The sub-census in 2007 distributed only 330 000 questionnaires, whereas the census of 2011 distributed more than 14.5 million. The before and after snapshot is crucial for decision makers to monitor and evaluate progress, using actual numbers derived from census.

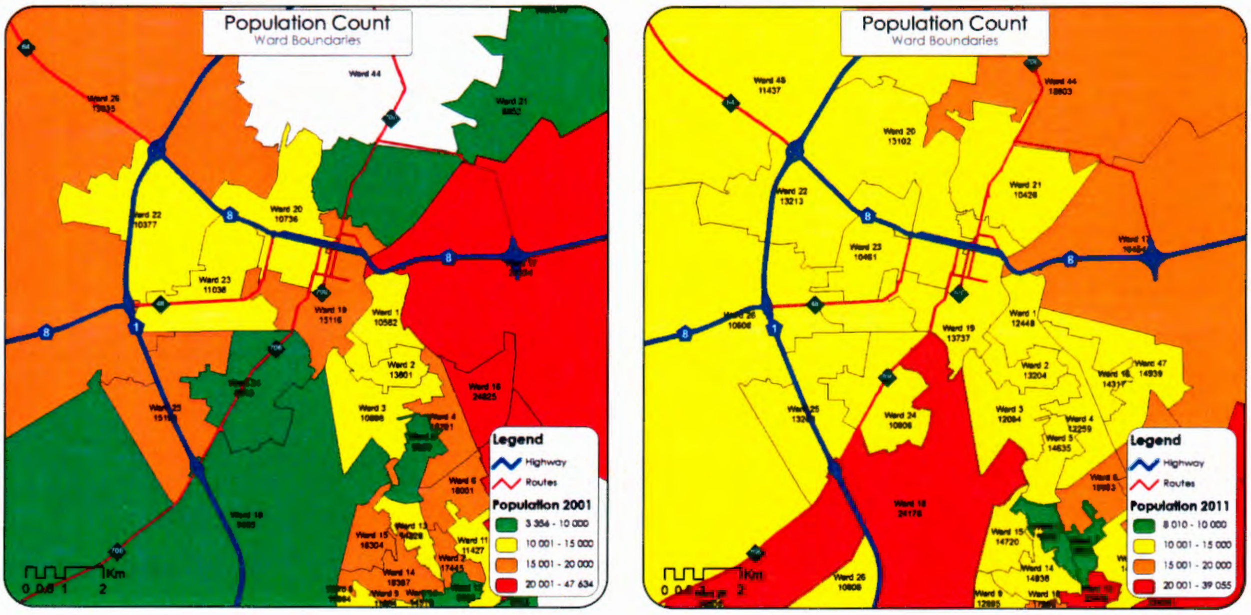


Figure 10 Population count in 2001 (left), population count 2011 (right)

Although census data from different time periods exists, such as 1996, 2001 and 2011, there are a few major concerns that severely hamper its use. Firstly, boundary changes, between census periods of 10 years the demarcated boundaries used in census change quite frequently as seen in Figure 10. Ward boundaries that are politically influenced are particularly vulnerable, as well as administrative boundaries. Comparison of dissimilar boundaries can be misleading as seen in Figure 10. Secondly, spatial resolution of data form census 2001 to 2011 is not the same. As mentioned the SAL only became available in 2011; however, census 2001 only made ward boundaries their lowest level of comparison. Ward boundaries are already quite large in size, which increases the loss of information through aggregation. For most part census comparison is performed on municipal level which is quite sad, not being able to note the finer changes over time due to the spatial level constraint. Lastly, the time lapse between every census is ten years, which is intrinsically too large to do a meaningful comparison. Trend analysis is especially difficult when gaps are that far apart. The fact that official census data is only available for 1996, 2001 and 2011, gives a very limited view on spatio-temporal change. Apart from the limitations, decision makers still find it useful to see cartographic representations from different time periods.

3.2.8. Zonal Statistics

Having vector-based spatial boundaries from census variables limits the type of analysis that can be performed in GIS. Oftentimes users want to estimate population size that does not coincide with the boundaries provided in census. To "escape" the boundary enclosure, GIS enables the conversion of vector boundaries into a raster surface. Raster representation offers significant benefits with a myriad of powerful analysis functions in GIS to enhance the use of census information. To convert census data into raster format, a normal tabulation is done and joined to the corresponding spatial layer. The cell size of the raster is determined, for example 30m x 30m. The area within the boundary of the census areal unit is then divided by the area of the raster cell to calculate the distribution of values within the census areal unit. If done correctly all the raster cells (pixels) within the census areal unit should add up to the original sum census variable within the vector boundary seen in Figure 11. Raster conversion should preferably be done on the lowest level of spatial aggregation to prevent loss of information. The smaller the cell size the more precise would be the calculation performed on the raster surface.

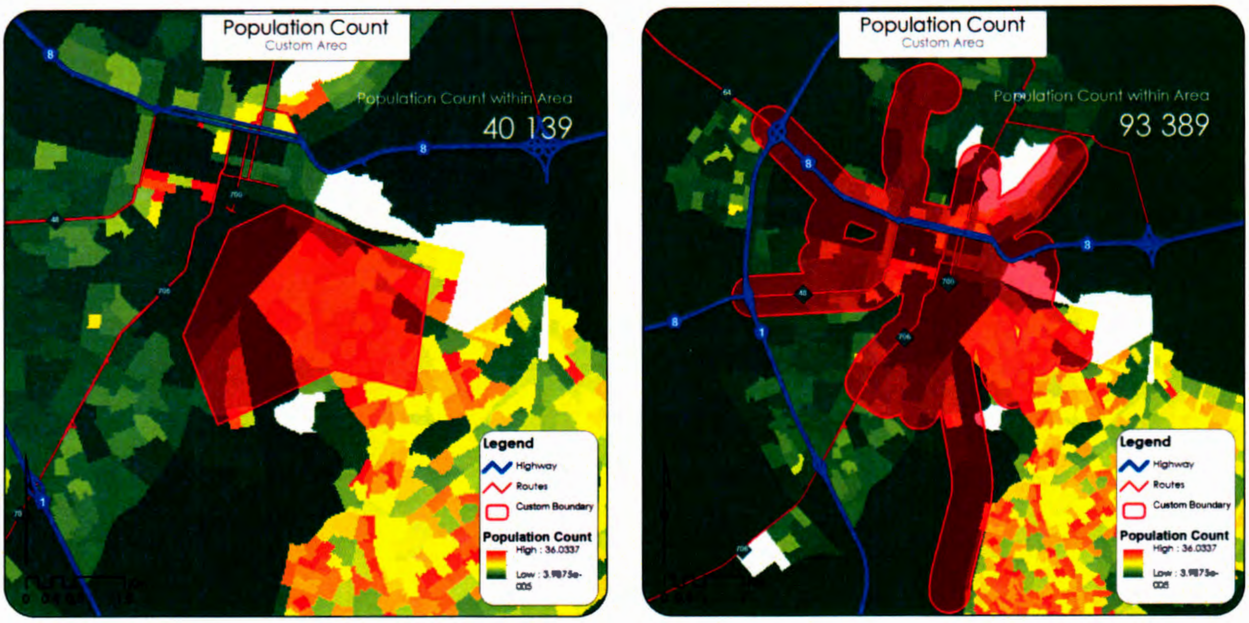


Figure 11 Custom population count of 40 139 within the red boundary (left), custom population count of 93 389 within the red boundary (right) using zonal statistics

The advantage of having census data available that is not bound to delineated census boundaries allows for reliable estimation of population size within a custom areas, such as the number of people staying within 500m of a river, or people within a custom defined area. Zonal statistics is not entirely accurate because cell values within the census area unit is still averaged, which may be far removed from the ground truth; however, for estimation purposes zonal statistics offers a great enhancement for decision makers.

3.2.9. Service Areas

Census data has proven to be particularly useful for estimating population quantities within a prescribed distance. GIS offers census users the ability to estimate a host of threshold and coverage statistics using underlying census data. One particularly useful function provided in GIS is network analysis, which buffers a predefined distance from a chosen location or locations, using the underlying road network as a guide. Travel distance is a common planning principle, especially in governmental service provision. For example, planning new social facilities for a community would require standards in terms of acceptable travel/response distance and the threshold/capacity of the facility see Figure 12.

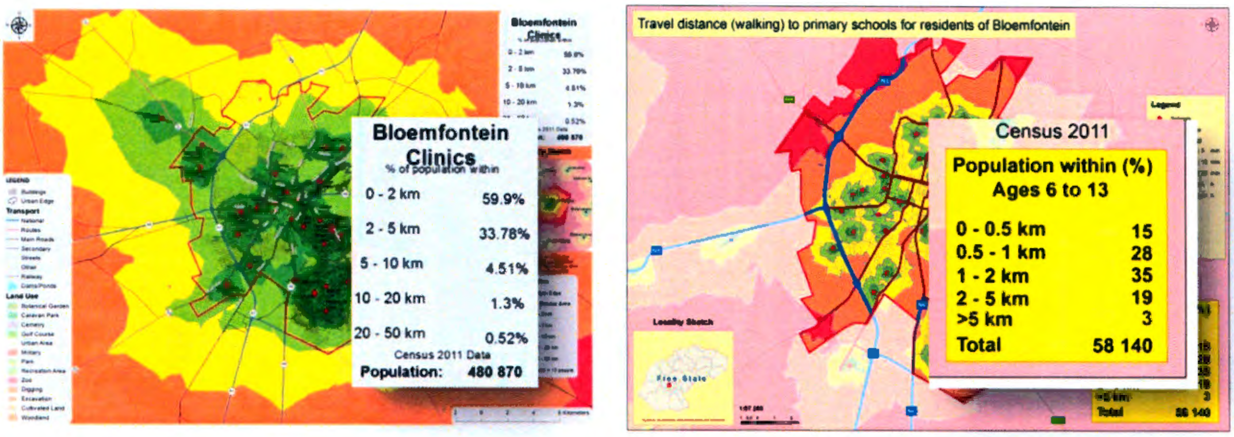
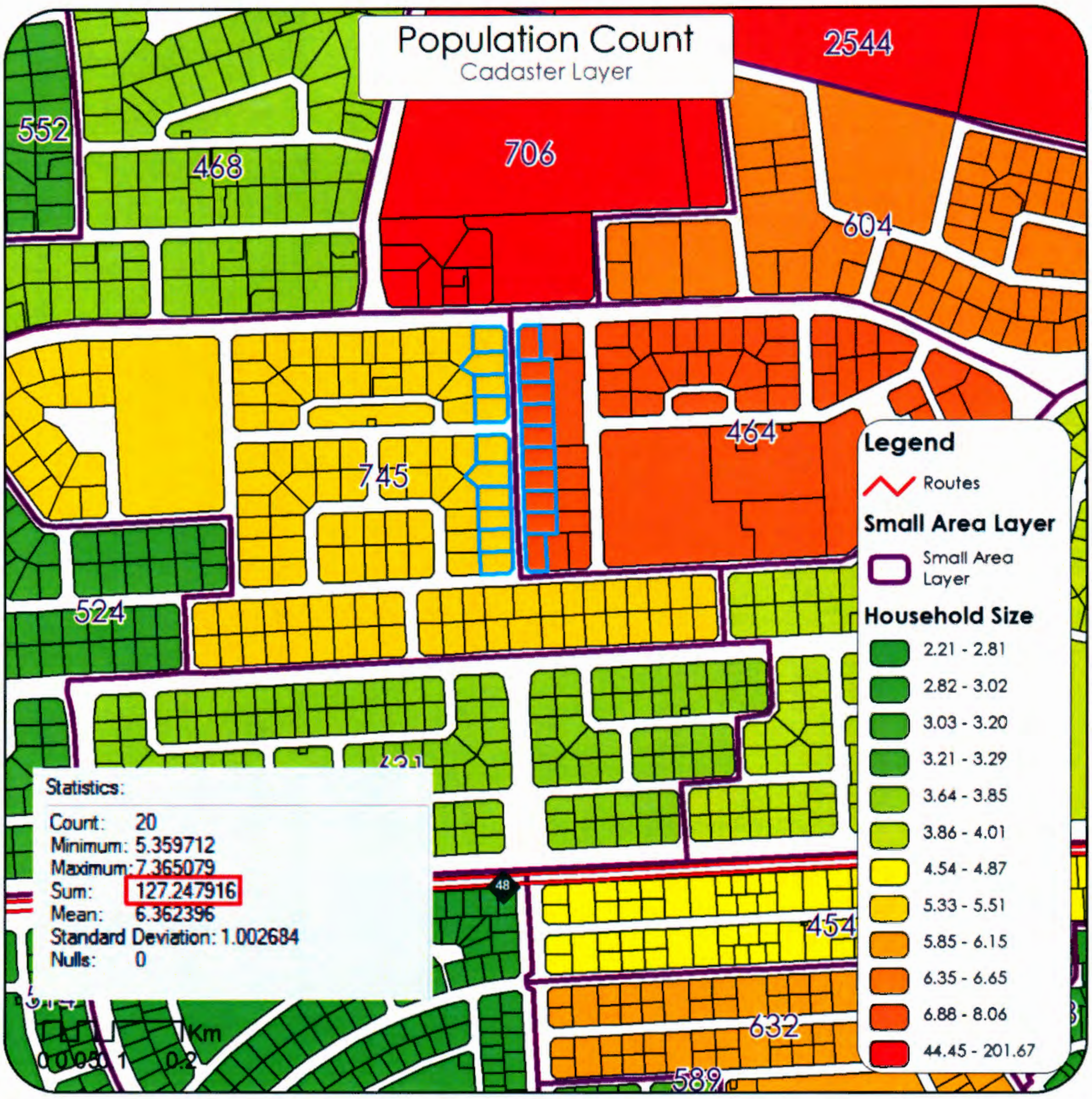


Figure 12 Population estimates from census data using network analysis

After calculating the service area (interval distance form facility) using the underlying network, the census data can be overlaid on top of the service area and joined to the service interval and summarised to compute population estimates. To increase accuracy it is best to convert census data to points to be joined to the service interval to give a more appropriate estimation. This functionality of integration between census and GIS greatly enhanced with the use of census data. In addition to the fact that network analysis and census data is seldom utilised, planning greatly benefits from having reasonable estimates on population size within a prescribed distance of a social facility. Answers can then be represented in cartographic form where totals can be summarised in percentile to make interpretation easier. For example 35% of kids aged 6-13 live within 1-2 km from a primary school.

3.2.10. Disaggregation

As mentioned earlier, aggregation is the main reason for data loss and misrepresentation in census. Due to the fact that census data gets aggregated for the smallest unit called a dwelling frame into enumerated areas, small areas, suburbs, place, municipal, district and province, significant loss of data is present as area size increases. Since 2011 the availability of SAL was added to the census dissemination product, which greatly reduced the area size of sampled units. SAL is nearly the same size as the original enumerated areas. Having the SAL at our disposal, disaggregation can be attempted with reasonable success.



```
Statistics:
Count:    20
Minimum:  5.359712
Maximum:  7.365079
Sum:      127.247916
Mean:     6.362396
Standard Deviation: 1.002684
Nulls:    0
```

Figure 13 Population count using disaggregated census data using surveyed parcels

If data smaller than the SAL exists outside census, it can be used as a guideline to inform the disaggregation process. Depending on the quality of the underlying feature, disaggregation enhances to the usability of census data. Another dataset that is publicly available for users is cadastre boundaries produced by the land surveyor. Since the SAL is smaller in urban areas, the underlying parcel topology could be used to inform disaggregation. Converting parcels to points and joining the point to the census data enables us to distribute the census value within the census area unit equally to all the points within the census boundary as seen in Figure 13. This is an improvement of zonal statistics, which simply ignores the underlying land-use patterns. After disaggregation the results can be joined to the parcel layer, where every parcel can be counted individually. Census data is the most common dataset used in disaggregation for population estimates. With a disaggregated dataset one can, with reasonable care, estimate the total population based on the underlying selected parcels. However, disaggregation is a tedious task and is seldom performed except by more experienced GIS users. In spite of the technical difficulty associated with, disaggregated census data is immensely powerful for decision makers to do estimates based on micro level or sub-census level.

3.3.CONCLUSION

To evaluate integrative use between census and GIS, this project does so by drawing up a custom framework based on ten different aspects. Each aspect is briefly discussed with accompanying examples. The CENGIS framework essentially evaluate integration on three levels; firstly, on active use which serves as an indicator where participants actively utilise the prescribed concepts. Secondly, the frequency of use is evaluated; and lastly, usability for decision makers is accessed. With these three evaluators, each aspect of integration was explained in detail as to why it is included in the CENGIS framework on integrative use. Tabulation is the basic and most elementary form of integration, which can be either standard or more extensive with multiple variables. Dynamic cartographic map generation is another important aspect that greatly enhances integrative use by reducing the time needed to produce informative cartographic products for decision makers. Understanding the nature of cartographic display through classification, interval size and symbol type enhances its usefulness by minimising misrepresentations. Spatial aggregation is indispensible for effective use of census data. Knowing the inherent weakness of spatial aggregation, as articulated in the MAUP, is crucial to consider when using census data for decision making, especially in time series where older census boundaries do not necessarily align with newer ones. Knowing how to escape from the tyranny of census boundaries is essential for population estimates, threshold calculations and micro analysis through disaggregation.

CHAPTER 4: FINDINGS AND DISCUSSIONS ON CENGIS FRAMEWORK

4.1.INTRODUCTION

Chapter 3 is taken as the baseline for chapter four where the findings will be discussed in terms of the CENGIS framework. This chapter focuses on the results collected after users' participation in the CENGIS survey. All ten aspects on integration are discussed individually. Each aspect is preceded by an illustration and brief description to clarify the responses received. Discussions cover the three questions asked in the survey for each aspect of integrative use. Firstly, active use which is derived from a close ended question is depicted as a graph. Secondly, frequency of use is rated on a Likert scale form 1-5, which is averaged to derive the frequency. Lastly, usability for decision makers is also rated on a Likert scale form 1-5, from which the averaged usability score is derived. By using the CENGIS framework with the three evaluators, the overarching aim on evaluating integrative use between census and GIS can be effectively answered.

4.2. CENGIS FRAMEWORK

The CENGIS framework is intended to evaluate 10 aspects on census and GIS integration. The survey can be roughly divided into 10 questions, which cover most of the main concerns, techniques and usage between census and GIS. What is, however, important to note is that the CENGIS framework serves as a preliminary guide toward assessing user awareness of basis and some more advance uses as well as the immediate constraints that all census and GIS users needlessly have to know to improve the decision making process. The CENGIS survey is also intended to evaluate integrative use, specifically among planners. Planning is the primary sector that uses census data in the process of making evidence based decisions through statistical analysis. It is important to keep in mind that when results are evaluated, it reflects the targeted audience. The total number of participants of the CENGIS survey is 23 of which the majority is town planners.

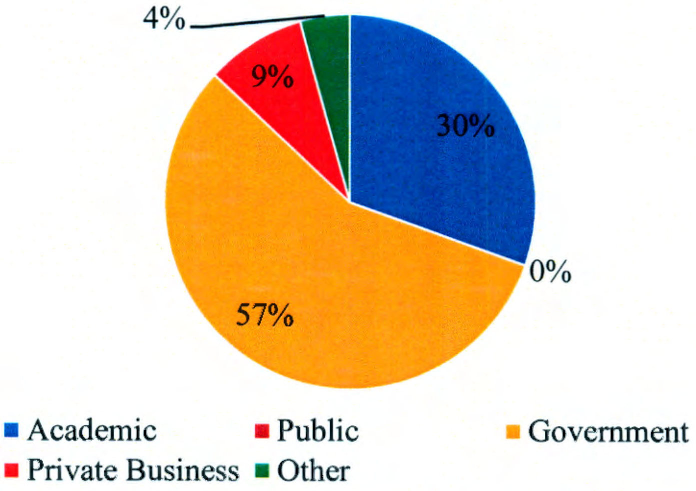


Figure 15 Participation by sector

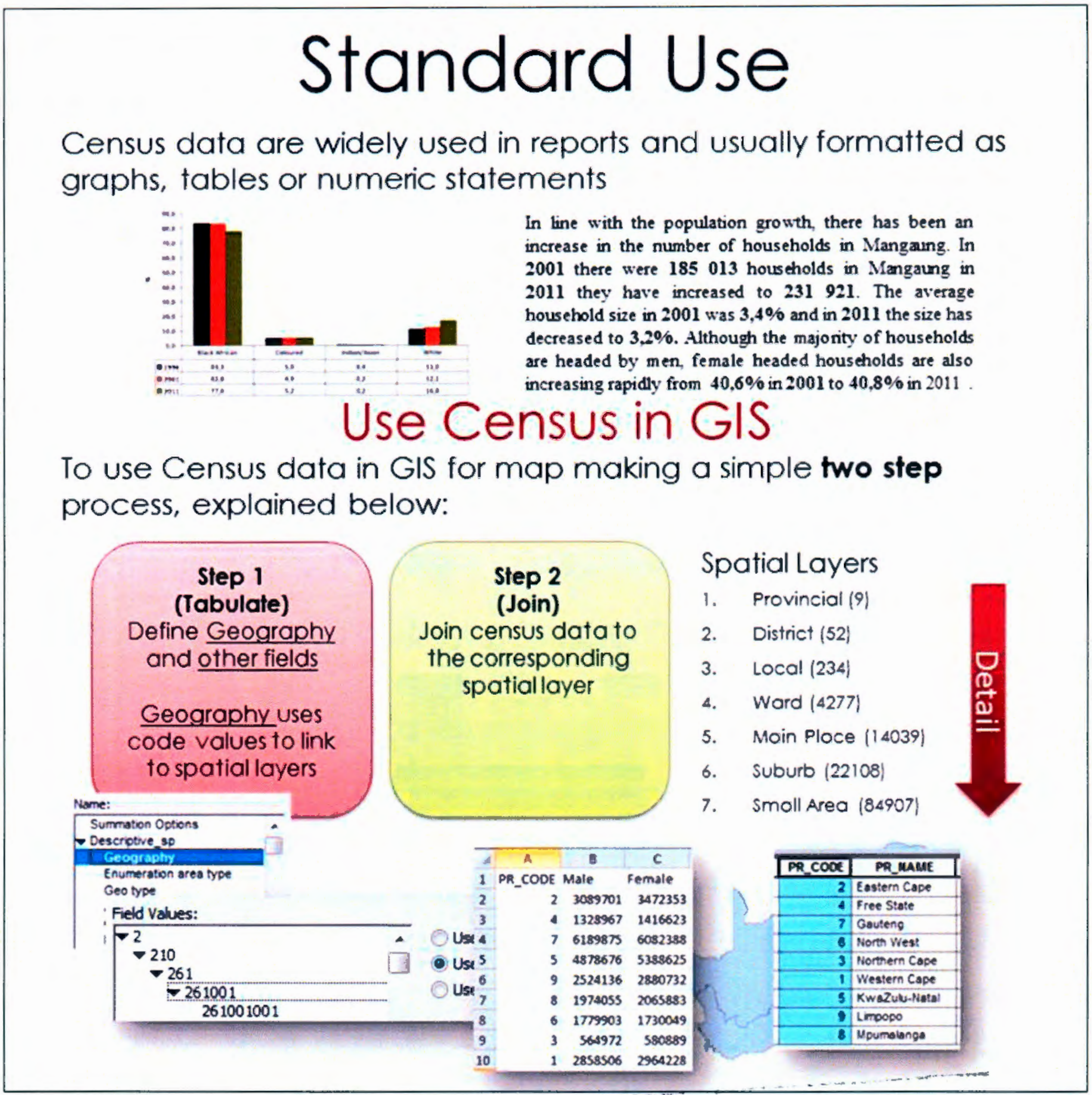
The first question asked to participants was to identify their respective sectors for using census data. According to the results acquired, users of census data tend to be more institutional orientated such as academic or governmental, which is understandable due to the nature of census variables with 87% of the participants. Private institutions do make use of census ,yet on a much smaller scale than their public counterparts.

Table 1 Frequency of using census data in GIS

Frequency	**Never** **(1)**	**Almost Never** **(2)**	**Occasionally** **(3)**	**Often** **(4)**	**Very Often** **(5)**
Score	7	0	8	4	4
Total %	30.4	0	34.8	17.4	17.4
Average Score 3.26					

Secondly, concerning the question on frequency of use it is important to understand just how often people are confronted with the various aspects on integrations as discussed in the survey. As seen, 30.4% of the participants have never utilised census data in GIS, about 34,8% uses it occasionally and 34.8% of the participants are identified as frequent users of census data in GIS. The average score on frequency of use is 3.26, which means the average participant utilises census data only occasionally.

4.2.1. Extraction Standard



Standard Use

Census data are widely used in reports and usually formatted as graphs, tables or numeric statements

In line with the population growth, there has been an increase in the number of households in Mangaung. In 2001 there were 185 013 households in Mangaung in 2011 they have increased to 231 921. The average household size in 2001 was 3,4% and in 2011 the size has decreased to 3,2%. Although the majority of households are headed by men, female headed households are also increasing rapidly from 40,6% in 2001 to 40,8% in 2011 .

Use Census in GIS

To use Census data in GIS for map making a simple **two step** process, explained below:

Step 1
(Tabulate)
Define Geography and other fields

Geography uses code values to link to spatial layers

Step 2
(Join)
Join census data to the corresponding spatial layer

Spatial Layers

1. Provincial (9)
2. District (52)
3. Local (234)
4. Ward (4277)
5. Main Place (14039)
6. Suburb (22108)
7. Small Area (84907)

Name:

Summation Options
Descriptive_sp
Geography
Enumeration area type
Geo type

Field Values:

2
210
261
261001
261001001

	A	B	C
1	PR_CODE	Male	Female
2	2	3089701	3472353
3	4	1328967	1416623
4	7	6189875	6082388
5	5	4878676	5388625
6	9	2524136	2880732
7	8	1974055	2065883
8	6	1779903	1730049
9	3	564972	580889
10	1	2858506	2964228

PR_CODE	PR_NAME
2	Eastern Cape
4	Free State
7	Gauteng
6	North West
3	Northern Cape
1	Western Cape
5	KwaZulu-Natal
9	Limpopo
8	Mpumalanga

Figure 16 Standard use of census data

The first aspect of the CENGIS survey is intended to make users aware of the numerous places that standard census stats are utilised, such as municipal reports or academic research. Despite the extensive use of census data from StatsSA, GIS integration has been largely underutilised due to various reasons, such as access to GIS, ignorance or having difficultly extracting census variables from tabulation software. The basic relationship between censuses and GIS is illustrated in a two-step process. Firstly, defining geography, and adding variables provided such as age, gender, income, education, etc. To facilitate seamless integration into GIS it is recommended that users only use numeric coded values for census areal units as illustrated in figure 16. Secondly, joining the tabulated data to the corresponding spatial layer is a simple GIS function, where the user defines which fields will participate in the join. Tabulation can be done for any of the spatial layers provided with census from Provincial SAL. The main focus is to evaluate general usage of census data in GIS by introducing basic aspects of tabulation and joining.

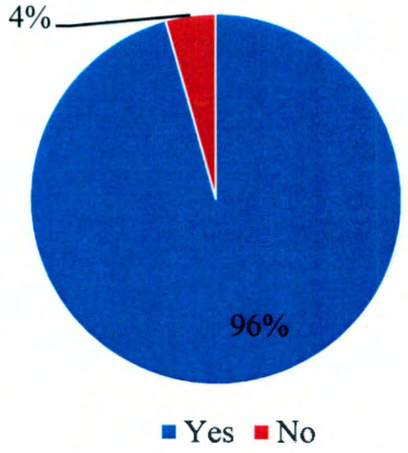


Figure 17 Use of census data in GIS

Standard use of census data in GIS through the two-step process as illustrated in Figure 17 is the most common form of integrative use by far - 96%. The main use of GIS is tabulation done in 3rd party software such as SuperCROSS using coded values and joined to the corresponding spatial representations. Only a very small percentage (4%) has not used census data in GIS.

Table 2 Frequency of using standard census data in GIS

Frequency	Never (1)	Almost Never (2)	Occasionally (3)	Often (4)	Very Often (5)
Score	2	1	11	7	2
Total %	8.7	4.3	47.8	30.4	8.7
Average Score 3.26					

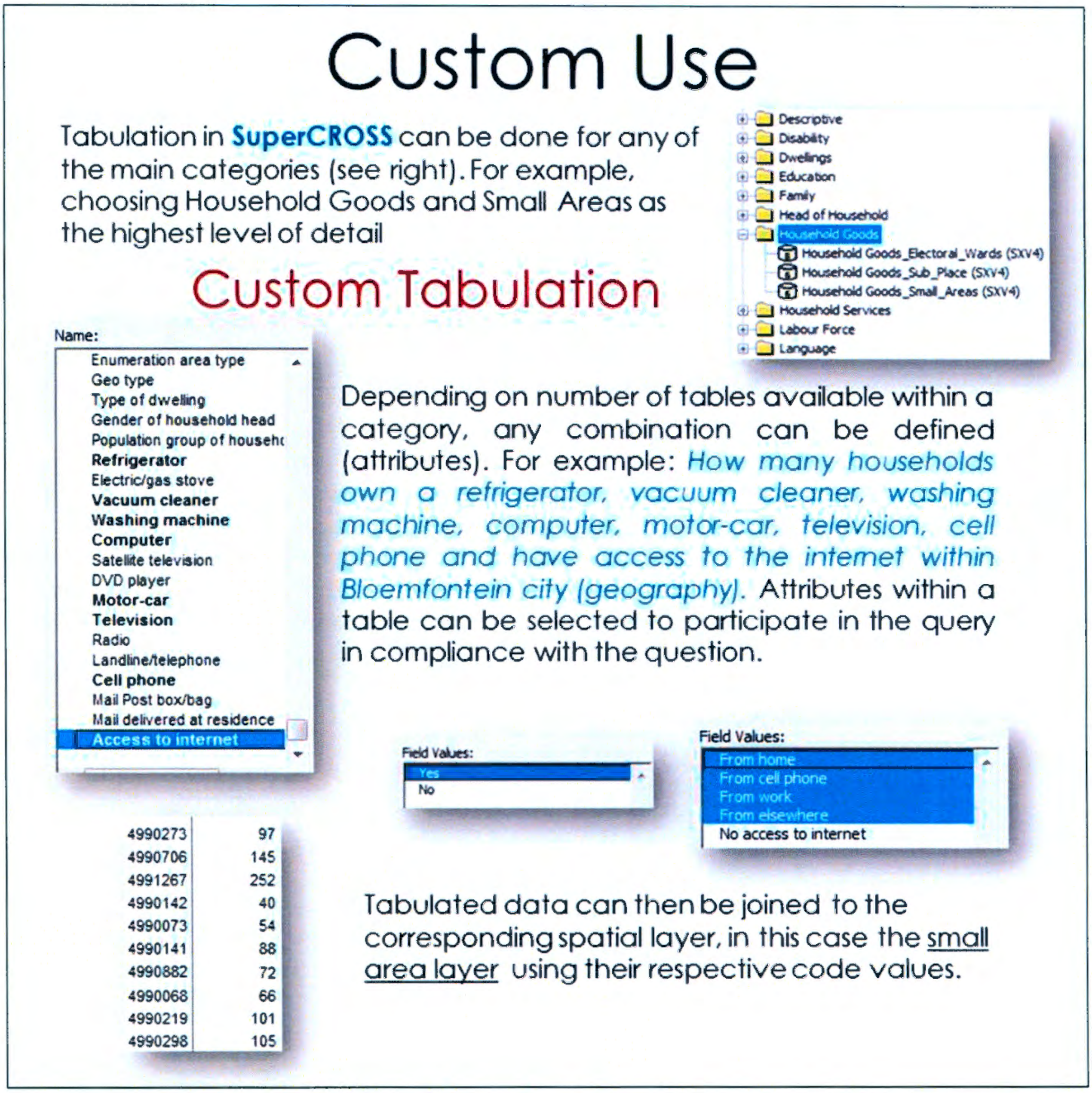
Standard use of census data in GIS has a relative high frequency with a combined score of 3.26 as seen in Table 2, which implies that most users do so occasionally. About 40% of users use standard census data frequently and 13% are non-frequent users of standard census data in GIS. This implies that 87% uses standard census data in general.

Table 3 Usefulness of standard census data in GIS for decision making

Frequency	**Not Useful**	**Barely Useful**	**Fairly Useful**	**Useful**	**Very Useful**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Score	2	2	6	7	9
Total %	8.7	8.7	26.1	30.4	26.1
Average Score 3.56					

Regarding usefulness of standard census data for decision making, the average is 3.56 seen in Table 3, which fall within the "useful" category of the spectrum. More than 80% of users deem standard census data to be useful for decision making. Only 17% ranks it is less useful for decision makers. Suggesting that users deem standard census data useful for decision support.

4.2.2. Extraction Customized



Custom Use

Tabulation in SuperCROSS can be done for any of the main categories (see right). For example, choosing Household Goods and Small Areas as the highest level of detail

Custom Tabulation

Depending on number of tables available within a category, any combination can be defined (attributes). For example: How many households own a refrigerator, vacuum cleaner, washing machine, computer, motor-car, television, cell phone and have access to the internet within Bloemfontein city (geography). Attributes within a table can be selected to participate in the query in compliance with the question.

4990273	97
4990706	145
4991267	252
4990142	40
4990073	54
4990141	88
4990882	72
4990068	66
4990219	101
4990298	105

Tabulated data can then be joined to the corresponding spatial layer, in this case the small area layer using their respective code values.

Figure 18 Custom tabulation process from census data

This question addresses a more advanced use of census data using custom tabulations. Firstly, the user is made aware of the different categories with their respective data tables in the SuperCROSS software package. Selecting an appropriate category, constructing a sentence in the form of a query can then be formulated by identifying participatory fields. Fields can be redefined to have only those values that apply to the original query. Grouping of values is a common practice for simplifying output. Advance tabulation greatly enhances strategic use of census variables in decision making. After tabulation the output can then be joined to the corresponding spatial layer in GIS. Deducting from statistics used in reports by governmental or private parties, very little custom tabulation is present despite the immense opportunities possible when utilising advance tabulation on census. It is purposely defined as custom tabulation because standard tabulation mostly include no field refinement or use of multiple participatory fields.

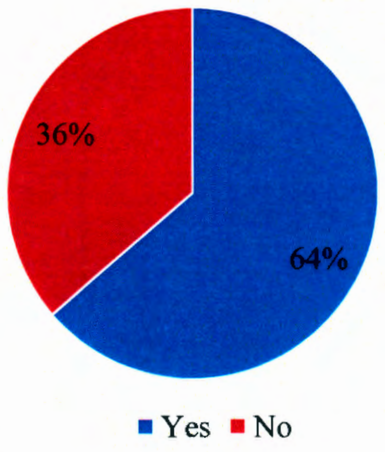


Figure 19 Use of custom queries with census data in GIS

When users do utilise census data the majority (64%) do so with custom queries. Only 36% of users prefer standard census tabulation seen in Figure 19. This shows that using advance queries through multiple fields being added and recoded has a significantly higher use than only standard queries. The use of custom queries has a much broader field of application than mere standard tabulations.

Table 4 Frequency of using custom queries from census

Frequency	Never (1)	Almost Never (2)	Occasionally (3)	Often (4)	Very Often (5)
Score	5	7	7	7	2
Total %	21.7	30.4	30.4	8.7	8.7
Average Score: 2.52					

Frequency of use in GIS using custom tabulated queries as described in the illustration in Table 4 is ranked between almost never to occasionally. Custom queries are significantly more complex to tabulate and need more user discretion on what they want to query which is proven by the fact that 53% almost never use custom queries. Only a small margin uses extensive tabulation with GIS (17%). This is unfortunate because tabulation forms a crucial part in answering complex scenarios so often faced by decision makers.

Table 5 Usefulness of custom queries from census for decision making

Frequency	Not Useful (1)	Barely Useful (2)	Fairly Useful (3)	Useful (4)	Very Useful (5)
Score	3	0	8	4	8
Total %	13	0	34.8	17.4	34.8
Average Score: 3.6					

When asked about usefulness, about 87% regard custom queries to be useful for decision makers in Table 5. Only 13% deems it as not useful. With an average score of 3.6 most users see custom tabulation as useful for decision makers in general. A significant margin of 34.5% regards it is as very useful indeed.

4.2.3. Map Production

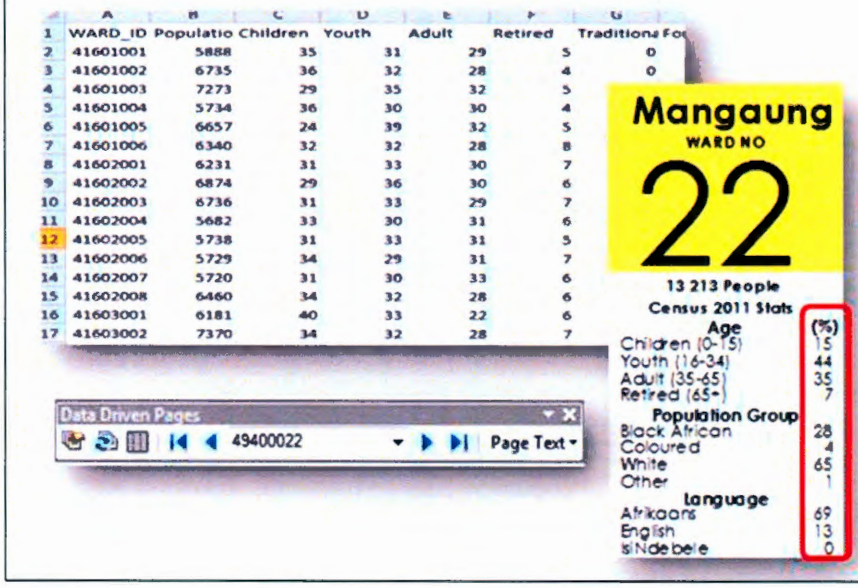
Map Production

Map production with Census data can be done with relative ease. After tabulation in SuperCROSS and joining data to the corresponding spatial layers, dynamic map making is possible.

Data-Driven Pages

Producing ward maps for the Free State with general census information for each ward in the province would be done as follows:

Tabulate the required data in SuperCROSS using either standard or custom queries. Join the data to the corresponding spatial layer (wards). Using the Data Driven Pages function in ArcGIS to produce a series of maps depicting census data can be produced.



	A	B	C	D	E	F	G
1	WARD_ID	Populatio	Children	Youth	Adult	Retired	Traditiona Fo
2	41601001	5888	35	31	29	5	0
3	41601002	6735	36	32	28	4	0
4	41601003	7273	29	35	32	5	
5	41601004	5734	36	30	30	4	
6	41601005	6657	24	39	32	5	
7	41601006	6340	32	32	28	8	
8	41602001	6231	31	33	30	7	
9	41602002	6874	29	36	30	6	
10	41602003	6736	31	33	29	7	
11	41602004	5682	33	30	31	6	
12	41602005	5738	31	33	31	5	
13	41602006	5729	34	29	31	7	
14	41602007	5720	31	30	33	6	
15	41602008	6460	34	32	28	6	
16	41603001	6181	40	33	22	6	
17	41603002	7370	34	32	28	7	

Mangaung
WARD NO
22
13 213 People
Census 2011 Stats

Age	(%)
Children (0-15)	15
Youth (16-34)	44
Adult (35-65)	35
Retired (65+)	7
Population Group	
Black African	28
Coloured	4
White	65
Other	1
Language	
Afrikaans	69
English	13
isiNdebele	0

Dynamic attributes derived from the census data can be added for each map. Attributes are updated automatically for each map.

Figure 20 Dynamic map production using data driven pages with attributes

The importance of automated map making cannot be overstated for census users. What this question aims to address is user awareness to automated mapping, using census data and dynamic map making functions such as data driven pages. Dynamic map making allows users to attach multiple dynamic attributes to a map, which is automatically updated for each map. Census variability can be attached to a dynamic map and displayed on the map as an information panel in either variable or percentage format. Automated map making in census allows for rapid map production using a master template. Attributes can be custom tabulated to answer predefined questions for the area on any of the provided spatial layers. Dynamic map making with census in GIS is one of the primary integrative uses between the two. Other than the significant improvement to GIS-related output that data driven pages adds for decision makers, it remains one of the least utilised functions on integration.

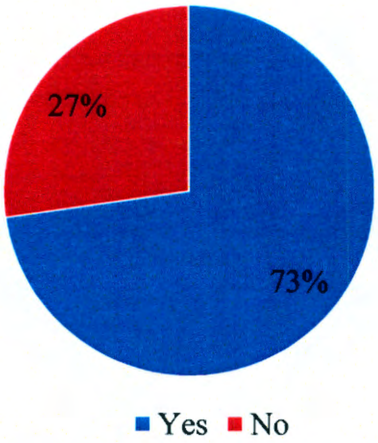


Figure 21 Use of data driven pages with census

Dynamic map making forms an intrinsic part to effective GIS and census integration. With 73% of participants being users of dynamic map making functions, such as data-driven pages, the usefulness of these functions are proven. Only 27% has not used dynamic map making functions before as seen in Figure 21.

Table 6 Frequency of using data driven pages with census data

Frequency	**Never (1)**	**Almost Never (2)**	**Occasionally (3)**	**Often (4)**	**Very Often (5)**
Score	4	4	10	4	1
Total %	17.4	17.4	43.5	17.4	4.3
Average Score 2.73					

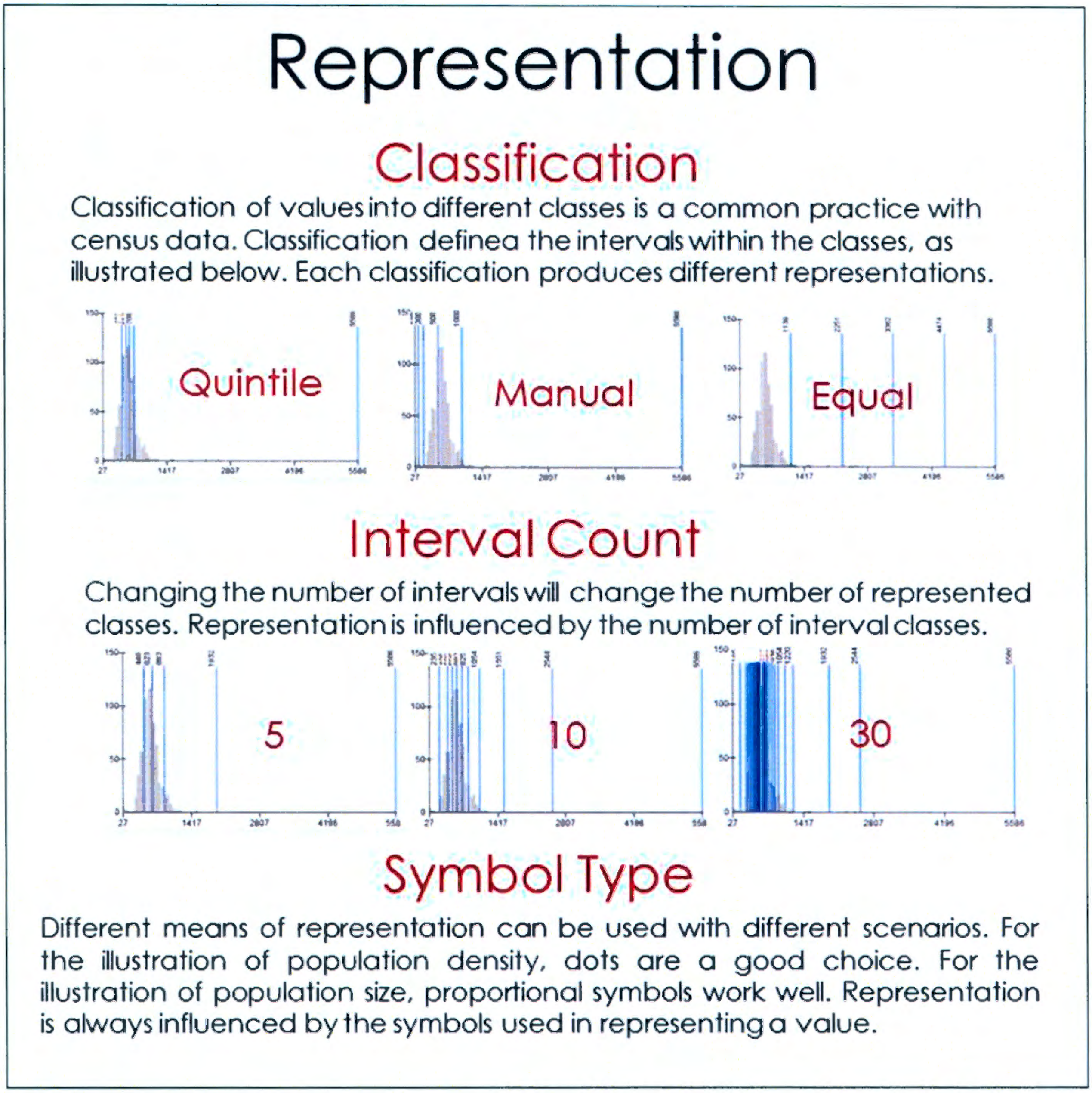
When asked how often users make use of data driven pages as shown in Table 6 (dynamic map making function in ArcGIS), 43.5% do so occasionally. About 21% of users identify themselves as frequent users of this function and 34.8% use it infrequently. The technicality of setting up a good template for dynamic map generation is proven by the fact that only 4.3% uses this function very often.

Table 7 Usefulness of data driven pages using census data for decision making

Frequency	Not Useful (1)	Barely Useful (2)	Fairly Useful (3)	Useful (4)	Very Useful (5)
Score	1	3	7	6	6
Total %	4.3	13	30.4	26.1	26.1
Average Score 3.56					

Usefulness of dynamic map making in ArcGIS for decision makers is rated 3.56 on the usability index in Table 7. This implies that most users deem this function on census and GIS integration for decision makers as useful, with 26.1% deeming it as very useful. The majority of users (82.6%) assent to the usefulness of dynamic map making functions in supporting decision makers. Only 17.3% do not approve of its usability.

4.2.4. Representation



Representation

Classification

Classification of values into different classes is a common practice with census data. Classification definea the intervals within the classes, as illustrated below. Each classification produces different representations.

Interval Count

Changing the number of intervals will change the number of represented classes. Representation is influenced by the number of interval classes.

Symbol Type

Different means of representation can be used with different scenarios. For the illustration of population density, dots are a good choice. For the illustration of population size, proportional symbols work well. Representation is always influenced by the symbols used in representing a value.

Figure 22 The influence of classification, interval count and symbol size on representation

Users of census data seldom question the authoritativeness of cartographic representations. Multiple parameters reside behind these representations that can drastically alter the representation without changing the actual data itself. This question is intended to raise users awareness of the deceptiveness of representation which can be arbitrarily altered by the user. This aspect of the census and GIS integration speaks to the limitation that cartographic depictions unfortunately inherit. By simply changing the classification scheme, interval count or symbol type the same dataset can be styled in an infinite number of ways, each giving a different impression to the user. It is important that census users understand representational constraints and learn how to best apply cartographic principles to prevent giving users wrong impressions. Representation often requires that different depictions of the same data be given to the user to clarify any obscurities that may be hidden in one cartographic display and shown in another.

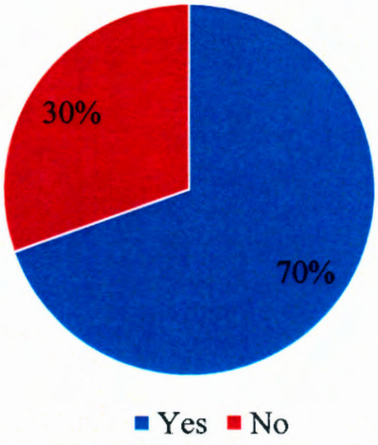


Figure 23 Using different representations of the same census data

Different representations of the same underlying census data is a relatively common practice as seen in Figure 23, with 70% actively making use of it. 30% does not use different representation of the same census data. What's important to realise is that the majority of census users are aware of the inherent cartographic weakness when displaying in GIS.

Table 8 Frequency of using different representations of the same census data

Frequency	Never (1)	Almost Never (2)	Occasionally (3)	Often (4)	Very Often (5)
Score	3	9	5	3	3
Total %	13	39.1	21.7	13	13
Average Score 2.74					

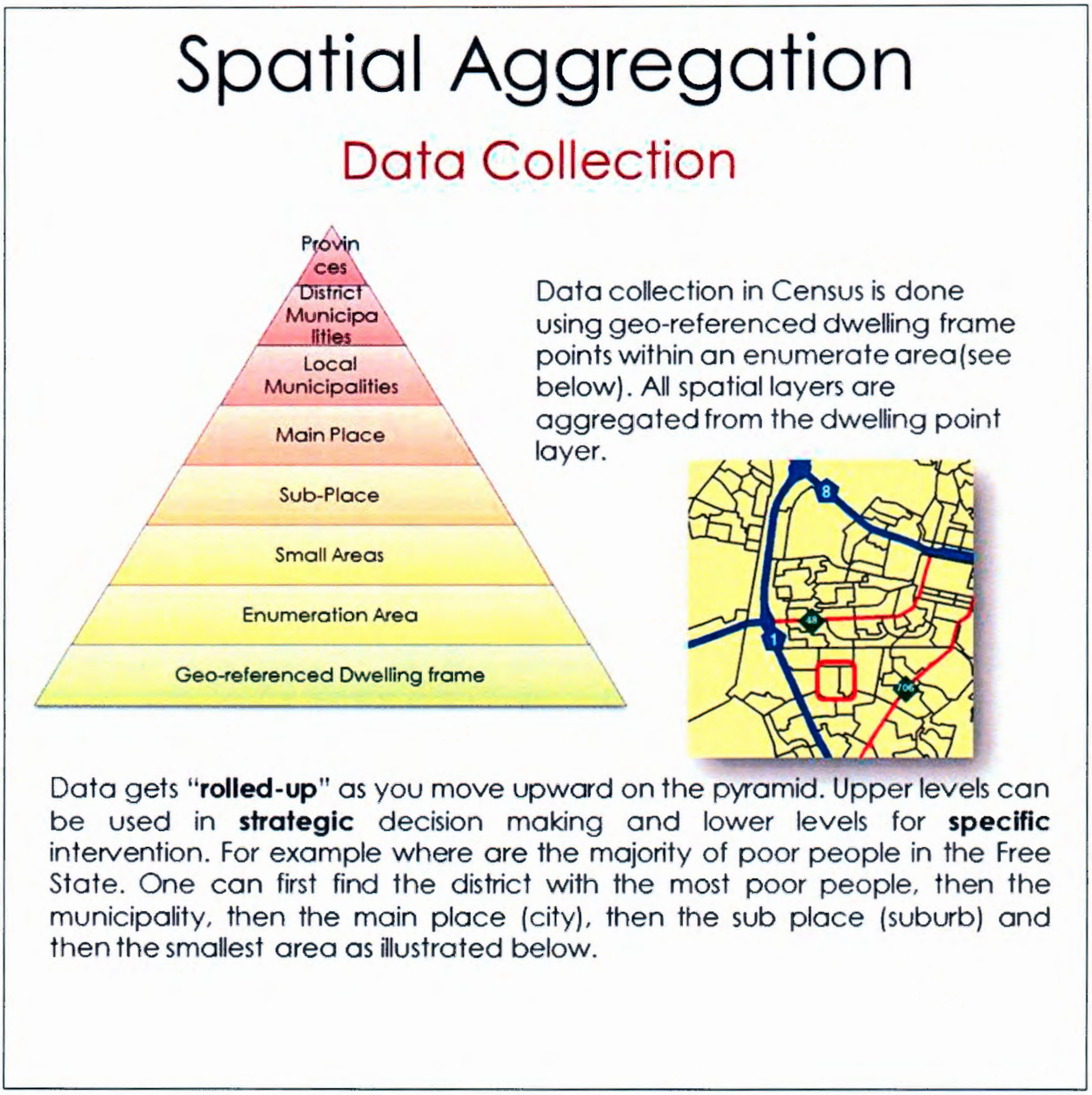
Despite the majority of users consenting to the use of different representations to clarify the underlying variables in Table 8, 39.1% almost never uses different representations and 13% never does. Only 21.7% occasionally utilises different representation and 26% makes more frequent use of different representations. The average score on frequency is 2.74 which is below the "occasionally" category, which testifies to average use of different representations.

Table 9 Usefulness of different representations of the same census data

Frequency	Not Useful (1)	Barely Useful (2)	Fairly Useful (3)	Useful (4)	Very Useful (5)
Score	2	3	5	8	5
Total %	8.7	13	21.7	34.8	21.7
			Average Score 3.48		

Although the frequency of use is relatively low, the usefulness is ranked at 3.48 in Table 9, which is above average. A meagre portion of the participants deem different representations to be useful for decision makers with 56% classifying it above average usefulness. Only a small percentage of users (21.7%) see it as not useful. This result confirms users awareness of the common misconceptions in using census data by making provision for multiple representations of the same underlying data to clarify any disparities.

4.2.5. Spatial Aggregation



Spatial Aggregation

Data Collection

Provinces

District Municipalities

Local Municipalities

Main Place

Sub-Place

Small Areas

Enumeration Area

Geo-referenced Dwelling frame

Data collection in Census is done using geo-referenced dwelling frame points within an enumerate area(see below). All spatial layers are aggregated from the dwelling point layer.

Data gets "**rolled-up**" as you move upward on the pyramid. Upper levels can be used in **strategic** decision making and lower levels for **specific** intervention. For example where are the majority of poor people in the Free State. One can first find the district with the most poor people, then the municipality, then the main place (city), then the sub place (suburb) and then the smallest area as illustrated below.

Figure 24 Spatial aggregation of census data using dwelling frame points

Users have to understand how data gets resampled to hide any sensitive information which resulting in obvious trade-offs and misrepresentations of ground truths. In addition to the fact that users take maps as being authoritative, aggregation will always hide information not seen by the user. It is important for users to realise that sampling points into homogeneous areas known as aggregation always introduces some margin of error to the representation and loss of important information. All census data is aggregated regardless of their spatial level. Understanding the pyramid of aggregation where large quantities of data, such as your geo-referenced dwelling frame points get resampled into enumerated areas, then into small areas, sub-places, main-places, municipalities, district and finally provinces, is an important concept to grasp as it influenced decision making in numerous ways. Users can use multiple levels of aggregation for different purposes; higher order layers can be used for strategic planning and lower order layers for detailed planning. This question is aimed to address users awareness of the aggregation, and how multiple levels of aggregation can facilitate them in making better decisions.

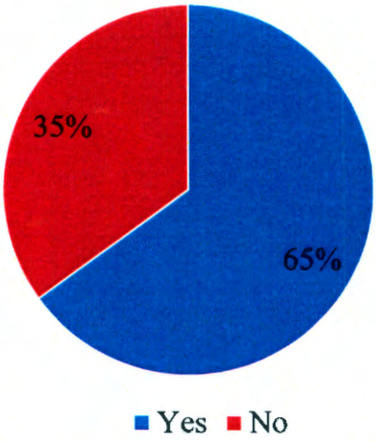


Figure 25 Using multiple spatial levels of census in GIS

Using multiple spatial levels of census is a common practice (65%) depicted in Figure 25. Only 35% does not use multiple spatial levels of census. Spatial aggregation is an important principle to understand, and using multiple spatial levels is the natural outflow of such understanding which is proven by the majority, of users utilising this component of census.

Table 10 Frequency of using multiple spatial levels from census

Frequency	Never (1)	Almost Never (2)	Occasionally (3)	Often (4)	Very Often (5)
Score	5	2	8	6	2
Total %	21.7	8.7	34.8	26.1	8.7

Average Score 2.91

Frequency of using multiple spatial levels by users is rated at 2.91 in Table 10, which is just below the occasionally category on the frequency spectrum. About 30.4% are rated as infrequent users of multiple spatial levels. However, 34.8% do use multiple spatial levels occasionally and 34.8% uses it frequently. Using multiple spatial levels should be a common practice to minimise the effect of aggregation, which is proven by the fact that about 70% of users make use of multiple levels of spatial aggregation.

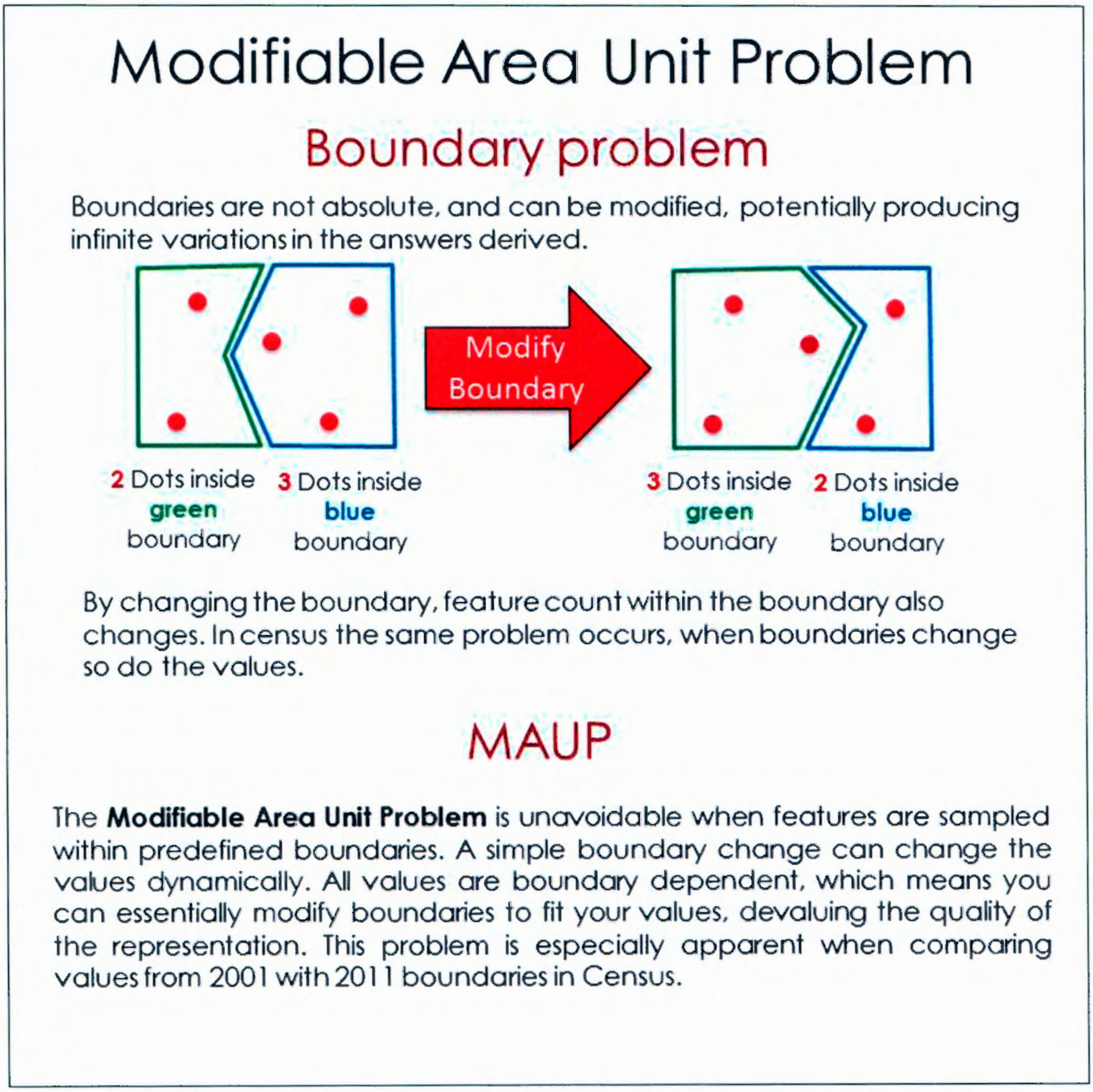
Table 11 Usefulness of multiple spatial levels in census for decision making

Frequency	Not Useful (1)	Barely Useful (2)	Fairly Useful (3)	Useful (4)	Very Useful (5)
Score	2	2	3	8	8
Total %	8.7	8.7	13	34.8	34.8

Average Score 3.78

Multiple spatial levels are rated as very useful by users with an average rating of 3.78 in Table 11, putting it in the "useful" category of the usability spectrum. 69.6% regard it as useful to very useful with only 17% not assenting to its usefulness. Spatial aggregation is proven to be a great benefit for decision makers to incorporate when using census data, with 82.6% consenting to its usefulness.

4.2.6. Modifiable Area Unit Problem



Modifiable Area Unit Problem

Boundary problem

Boundaries are not absolute, and can be modified, potentially producing infinite variations in the answers derived.

Modify Boundary
2 Dots inside green boundary
3 Dots inside blue boundary
3 Dots inside green boundary
2 Dots inside blue boundary

By changing the boundary, feature count within the boundary also changes. In census the same problem occurs, when boundaries change so do the values.

MAUP

The **Modifiable Area Unit Problem** is unavoidable when features are sampled within predefined boundaries. A simple boundary change can change the values dynamically. All values are boundary dependent, which means you can essentially modify boundaries to fit your values, devaluing the quality of the representation. This problem is especially apparent when comparing values from 2001 with 2011 boundaries in Census.

Figure 26 Modifiable area unit problem illustrated through boundary change

The MAUP is an intrinsic problem to all boundary sample datasets, of which census is no exception. This aspect of integration is intended to assess users awareness of the MAUP. Boundaries are infinitely modifiable, which undermines the authoritativeness of census data, since all values depend intrinsically on the extent of the boundaries. Census users have to grasp how misleading census areal units really are when compared to the original sampled area. This variability in census calls for serious consideration on how best to minimise the MAUP for decision makers. Unfortunately, complete eradication of the MAUP is not possible, however minimising its effect can be done to some extent by using smaller area units such as the SAL.

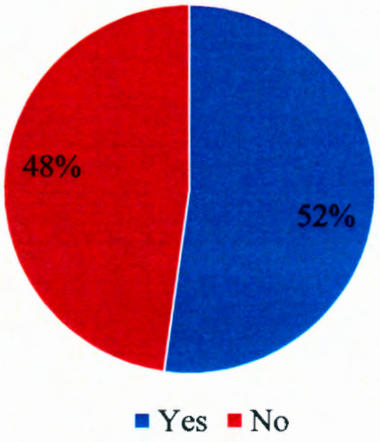


Figure 27 Awareness of the MAUP

Awareness of the MAUP is also a 50|50 problem, with 52% confirming awareness in Figure 27. What is staggering is the remarkable ignorance of most users with 48% only becoming aware at the point of taking the CENGIS survey. This creates a rather grim scenario where almost half of the users simply overlook this inherent weakness caused by using census area units.

Table 12 Frequency of using census boundary data

Frequency	**Never** **(1)**	**Almost Never** **(2)**	**Occasionally** **(3)**	**Often** **(4)**	**Very Often** **(5)**
Score	6	4	6	3	4
Total %	26.1	17.4	26.1	13	17.4
Average Score 2.78					

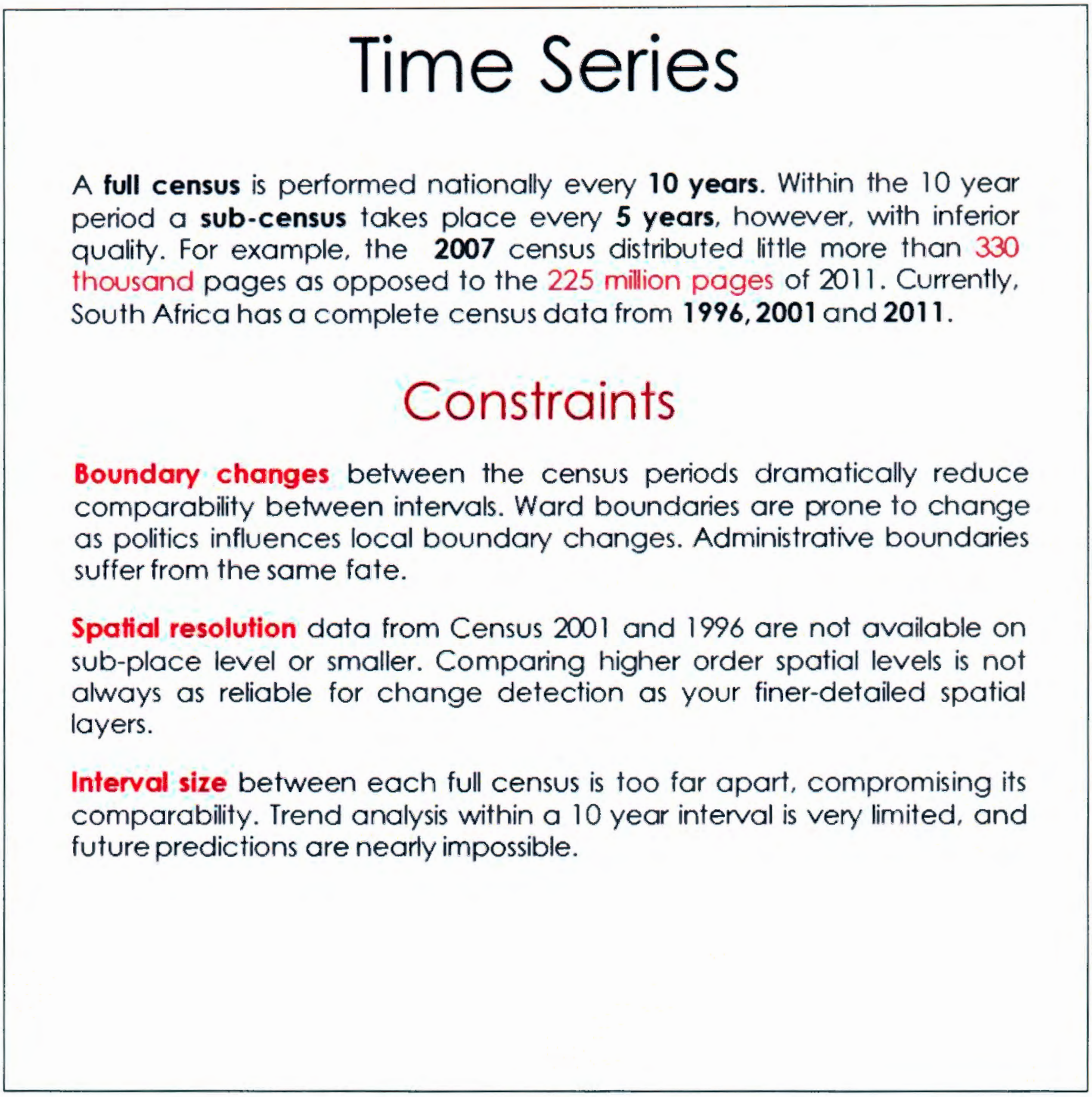
According to the response rendered, the frequency of using census boundary data varies, with 26.1% indicating they never use boundary data (which may be a slight misunderstanding). The average score on frequency of use is 2.78 in Table 12, placing the use of census boundary data below average. Only 30.4% indicates frequent use of census boundary data. Although the MAUP is present in all aggregated boundary confined datasets, users may not have considered the prevalence of the MAUP in using census data proven by the fact that 43.5% seldom use census boundary data.

Table 13 Usefulness of maps containing the MAUP for decision making

Frequency	Not Useful (1)	Barely Useful (2)	Fairly Useful (3)	Useful (4)	Very Useful (5)
Score	3	2	5	6	7
Total %	13	8.7	21.7	26.1	30.4
Average Score 3.52					

Despite the apparent limitations the MAUP introduces in using boundary data for decision makers seen in Table 13 participants nevertheless still deem census data useful (78.2%). Only a small portion considers this to be a fatal flaw for decision makers (21.7%). The average score proves that the usefulness is considered in the "useful" category. Although 48% of the participants for the first time have considered the implications on using census data that contain the MAUP, it still remains a useful source of information for decision makers.

4.2.7. Time Series



Time Series

A **full census** is performed nationally every **10 years**. Within the 10 year period a **sub-census** takes place every **5 years**, however, with inferior quality. For example, the **2007** census distributed little more than 330 thousand pages as opposed to the 225 million pages of 2011. Currently, South Africa has a complete census data from **1996, 2001** and **2011**.

Constraints

Boundary changes between the census periods dramatically reduce comparability between intervals. Ward boundaries are prone to change as politics influences local boundary changes. Administrative boundaries suffer from the same fate.

Spatial resolution data from Census 2001 and 1996 are not available on sub-place level or smaller. Comparing higher order spatial levels is not always as reliable for change detection as your finer-detailed spatial layers.

Interval size between each full census is too far apart, compromising its comparability. Trend analysis within a 10 year interval is very limited, and future predictions are nearly impossible.

Figure 28 Time series of different census datasets with usage constraints

Census data is sampled decennially and subsampled every five years in between. Having historical census data of the same area is vital for analysing trends and doing objective comparisons. Cartographic depictions of census data in different periods have significant value in understanding spatio-temporal change. This aspect on time variability of census data in GIS evaluates the usability of different census periods in GIS, which oftentimes give the wrong impressions. Besides time, comparing dissimilar boundaries is as good as comparing apples with pears. Spatial resolution is another aspect that needs to be considered when using different periods of census data. The censuses of 1996 and 2001 did not make provision for higher resolutions than ward level. The most detailed layer in census 2011 cannot be utilised for comparison purposes; only after 2021 will it be liable for comparison. Decision makers have to know the constraints when comparing census data from different time periods.

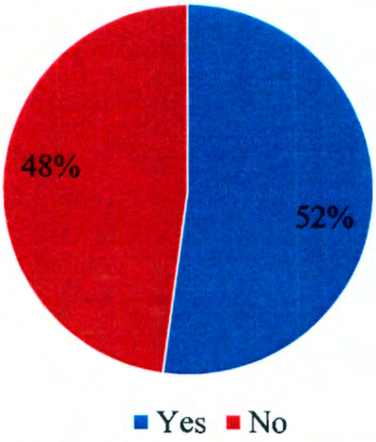


Figure 29 Use of older census data for comparison

In spite of the simple concept on comparing previous census records for time-based evaluation, users indicate that 48% in Figure 29 have never done so before with census data. The reason for this may be that previous census records are not accessible and usable in GIS for comparison. However, 52% indicate their use of previous census data for comparison.

Table 14 Frequency of using older census data for comparison

Frequency	**Never** **(1)**	**Almost Never** **(2)**	**Occasionally** **(3)**	**Often** **(4)**	**Very Often** **(5)**
Score	10	4	5	2	2
Total %	43.5	17.4	21.7	8.7	8.7
Average Score 2.22					

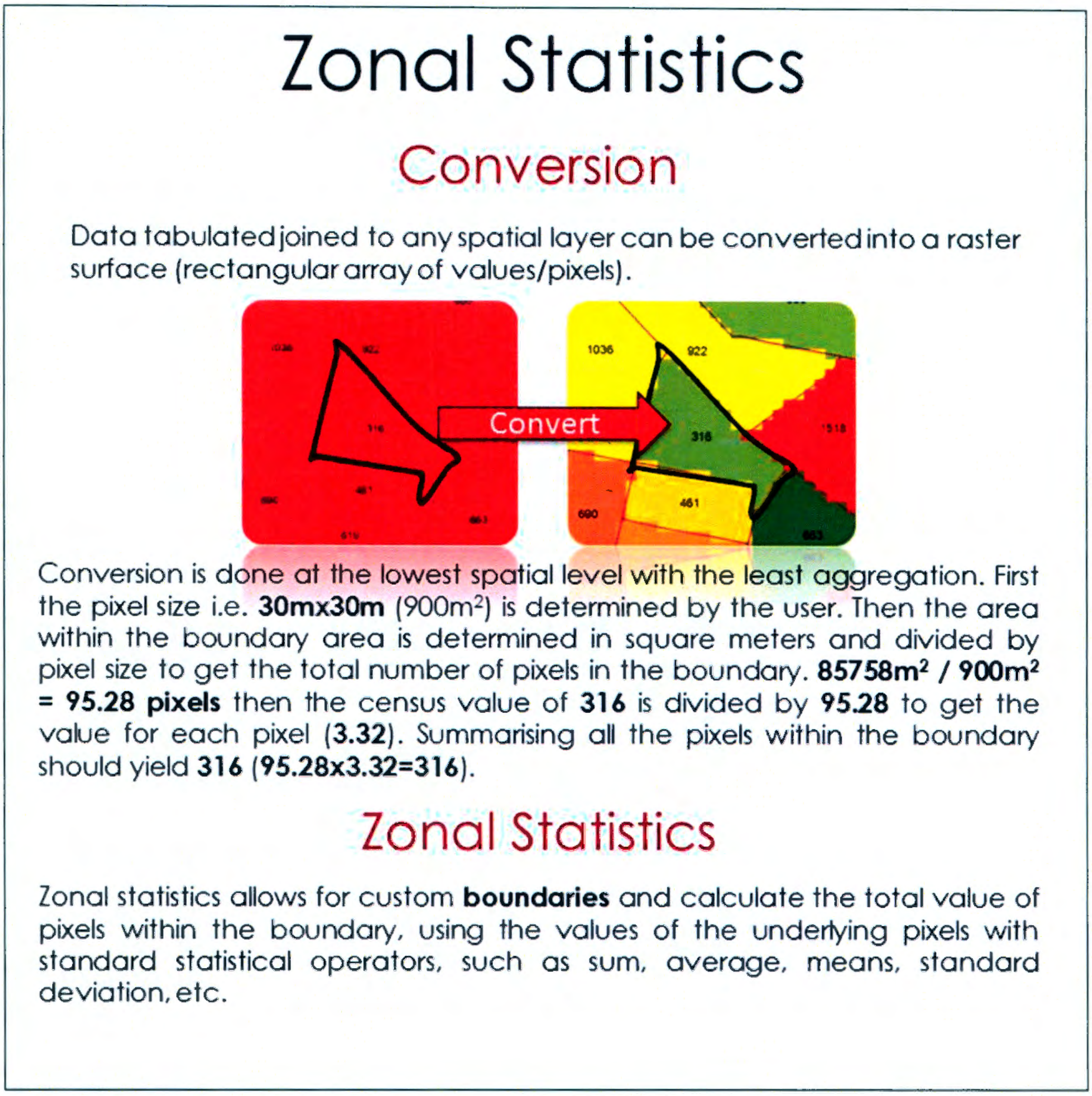
Although comparing previous datasets of census with newer ones is a common practice in analysing differences seen in Table 14, the majority of the participants (43.5%) indicate that they never use previous census data for comparison. With an average score of 2.22 it is classified in the "Almost Never" category. The reason for this apparent underutilisation could be due to the limitations on boundary changes, spatial resolution and time period between censuses as discussed above. Oftentimes tabulation software is too old and not available for the purpose of comparing variables. Only 17.4% uses older census data frequently for comparison purposes.

Table 15 Usefulness of older census data for comparison in decision making

Frequency	Not Useful (1)	Barely Useful (2)	Fairly Useful (3)	Useful (4)	Very Useful (5)
Score	5	2	6	3	7
Total %	21.7	8.7	26.1	13	30.4
			Average Score 3.21		

Users classify older census data above average in terms of usefulness for decision makers as shown in Table 15. 30.4% deems it as very useful and 26.1% as fairly useful. Only 30.4% sees it as not useful. Although the concept is easy to grasp, preparing older census data for comparison is a rigorous task, as indicated by the non-frequent use by users. Overall, 69.5% uses previous census data for comparison in facilitating decision makers.

4.2.8. Zonal Statistics



Zonal Statistics

Conversion

Data tabulatedjoined to any spatial layer can be convertedinto a raster surface (rectangular array of values/pixels).

Conversion is done at the lowest spatial level with the least aggregation. First the pixel size i.e. **30mx30m** (900m^2) is determined by the user. Then the area within the boundary area is determined in square meters and divided by pixel size to get the total number of pixels in the boundary. **85758m^2 / 900m^2 = 95.28 pixels** then the census value of **316** is divided by **95.28** to get the value for each pixel (**3.32**). Summarising all the pixels within the boundary should yield **316** (**95.28x3.32=316**).

Zonal Statistics

Zonal statistics allows for custom **boundaries** and calculate the total value of pixels within the boundary, using the values of the underlying pixels with standard statistical operators, such as sum, average, means, standard deviation, etc.

Figure 30 Zonal statistics conversion from vector to raster format

This question addresses advance integration between census and GIS. For most part users only utilise vector boundaries in census, which unfortunately has their limitations. To overcome this constraint users can convert vector data from census into corresponding raster sets. Raster opens the way for myriads of analysis functions in GIS, such as map algebra and statistical analysis. Due to some mathematical requirements, integration of census data into raster has been very limited and is only utilised by more experienced users. What this question aims to do is to make users aware on how to "free" census data from their "cages" (areal units). After conversion, users can then use their own boundaries and recalculate totals for each using zonal statistics. Understanding the simple mathematical conversion from vector to raster will assist decision makers in answering more difficult questions of population estimates.

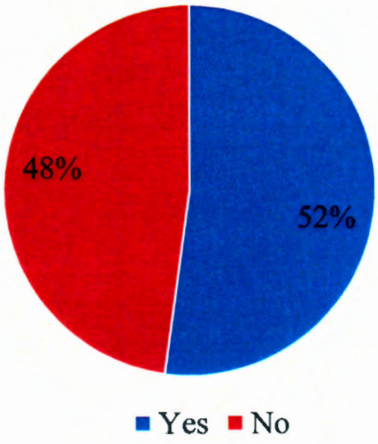


Figure 31 Use of zonal statistics with census data

As seen, the use of zonal statistics is a 50|50 scenario shown in Figure 31. With 48% of the users having no exposure to its use this proves that this form of integrative use between census and GIS is seldom considered. As described, the technicality in converting vector data to raster format may be a hurdle only experience GIS users could overcome proven by the 52% that have utilised this function before.

Table 16 Frequency of using zonal statistics with census

Frequency	Never (1)	Almost Never (2)	Occasionally (3)	Often (4)	Very Often (5)
Score	9	3	8	1	2
Total %	39.1	13	34.8	4.3	8.7
Average Score 2.30					

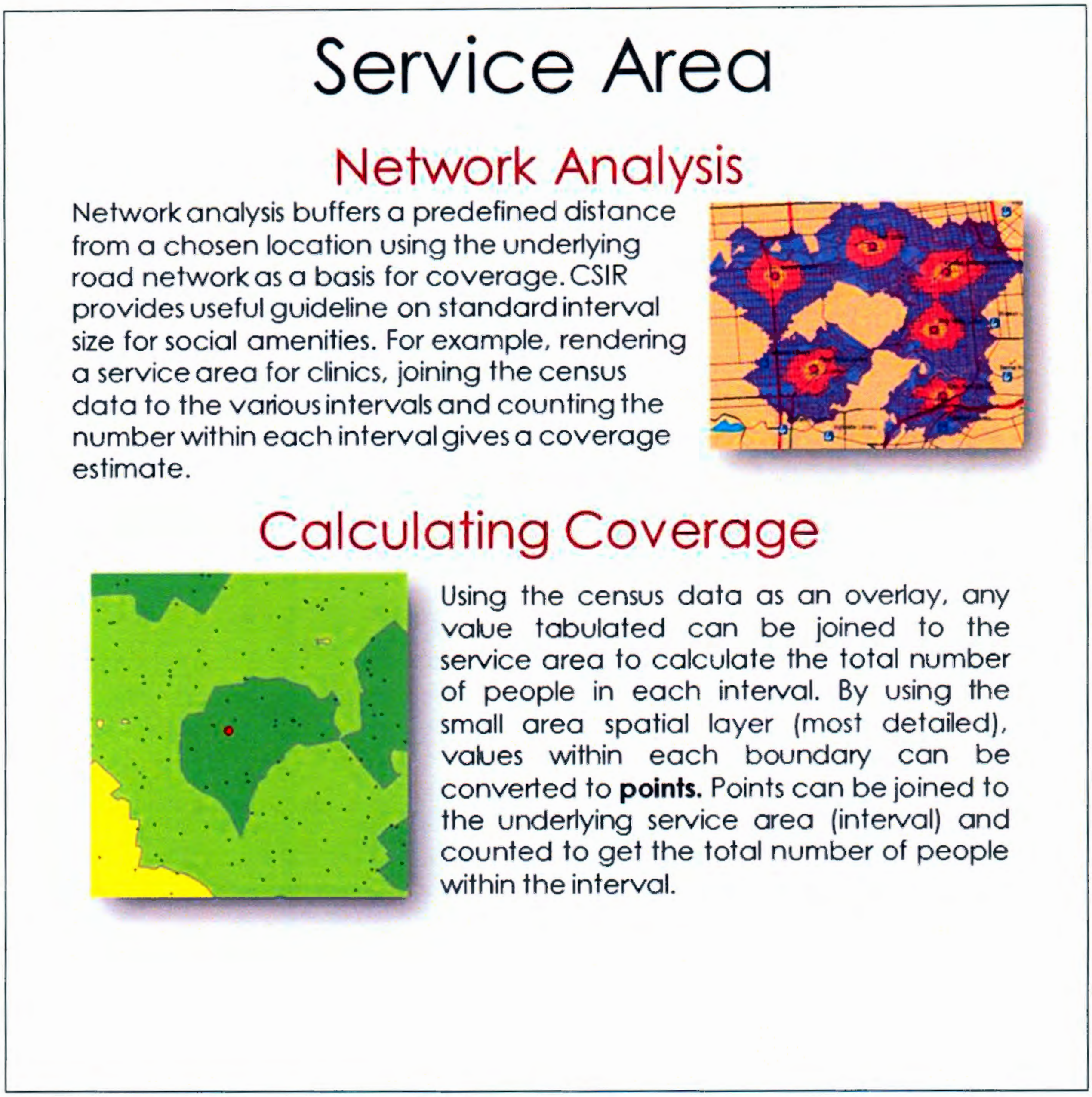
Most of the users in Table 16 identified themselves as not having used zonal statistics before (39%). Very meagre portions use it frequently (13%). This proves the complexity of this procedure of converting census data to raster format for zonal statistics. About 34.8% sees themselves as occasional users. The average score on frequency is rated 2.3, which implies it is one of the less frequent forms of census and GIS integration.

Table 17 Usefulness of zonal statistics on census for decision making

Frequency	Not Useful (1)	Barely Useful (2)	Fairly Useful (3)	Useful (4)	Very Useful (5)
Score	4	2	3	6	8
Total %	17.4	8.7	13	26.1	34.8
Average Score 3.52					

Despite the infrequent use of zonal statistics, users nevertheless deem it as a very useful function with 60.9% classifying it as useful to very useful in Table 17. The average score on zonal statistics is 3.52 which prove that it is above average usefulness for decision makers. 26.1% classifies it as non-useful, due to complexity or not having used it before. Overall evaluation on zonal statistics has proven it to be a useful integration mechanism between census and GIS for decision makers (73.9%).

4.2.9. Service Areas



Service Area

Network Analysis

Network analysis buffers a predefined distance from a chosen location using the underlying road network as a basis for coverage. CSIR provides useful guideline on standard interval size for social amenities. For example, rendering a service area for clinics, joining the census data to the various intervals and counting the number within each interval gives a coverage estimate.

Calculating Coverage

Using the census data as an overlay, any value tabulated can be joined to the service area to calculate the total number of people in each interval. By using the small area spatial layer (most detailed), values within each boundary can be converted to **points.** Points can be joined to the underlying service area (interval) and counted to get the total number of people within the interval.

Figure 32 Service area generation using network analysis

This question focuses on integrative use of census data in network analysis for calculating thresholds and capacity constraints. Integrative use of census and network analysis enables users to answer numerous complex questions related to planning. Knowing how many people live within a prescribed distance from a social facility enables planners to do a high level of strategic planning. However, census allows users to utilise census data in network analysis. Custom usage of census data and network analysis is immensely powerful for decision makers when estimating acceptable travel distance and capacity constraints.

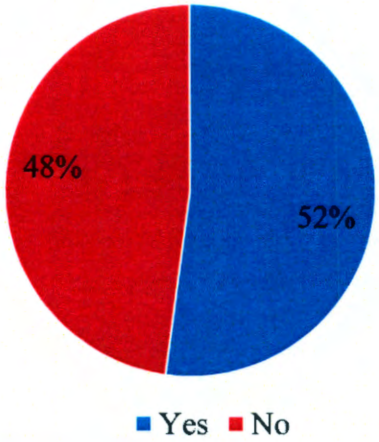


Figure 33 Use of network analysis with census

Network analysis in itself is a rather complex GIS function, yet extremely powerful means to answer complex questions estimated within a prescribe distance. The fact that 48% have not use this function before on census and GIS integration before proves that's its more useful for the more experienced GIS users (52%) seen in Figure 33. Network analysis in combination with census variables can be considered a higher level of integrative use proven by the 50|50 scenario as it was similar than zonal statistics.

Table 18 Frequency of using network analysis with census data

Frequency	**Never (1)**	**Almost Never (2)**	**Occasionally (3)**	**Often (4)**	**Very Often (5)**
Score	8	6	5	2	2
Total %	34.8	26.1	21.7	8.7	8.7
Average Score 2.30					

Deducting from the 48% of users in Table 18 who has never utilised network analysis in combination with census data is proven by the fact that 60.9% of the participants rank its frequency of use below average. This is also confirmed in the 2.30 score which falls within the

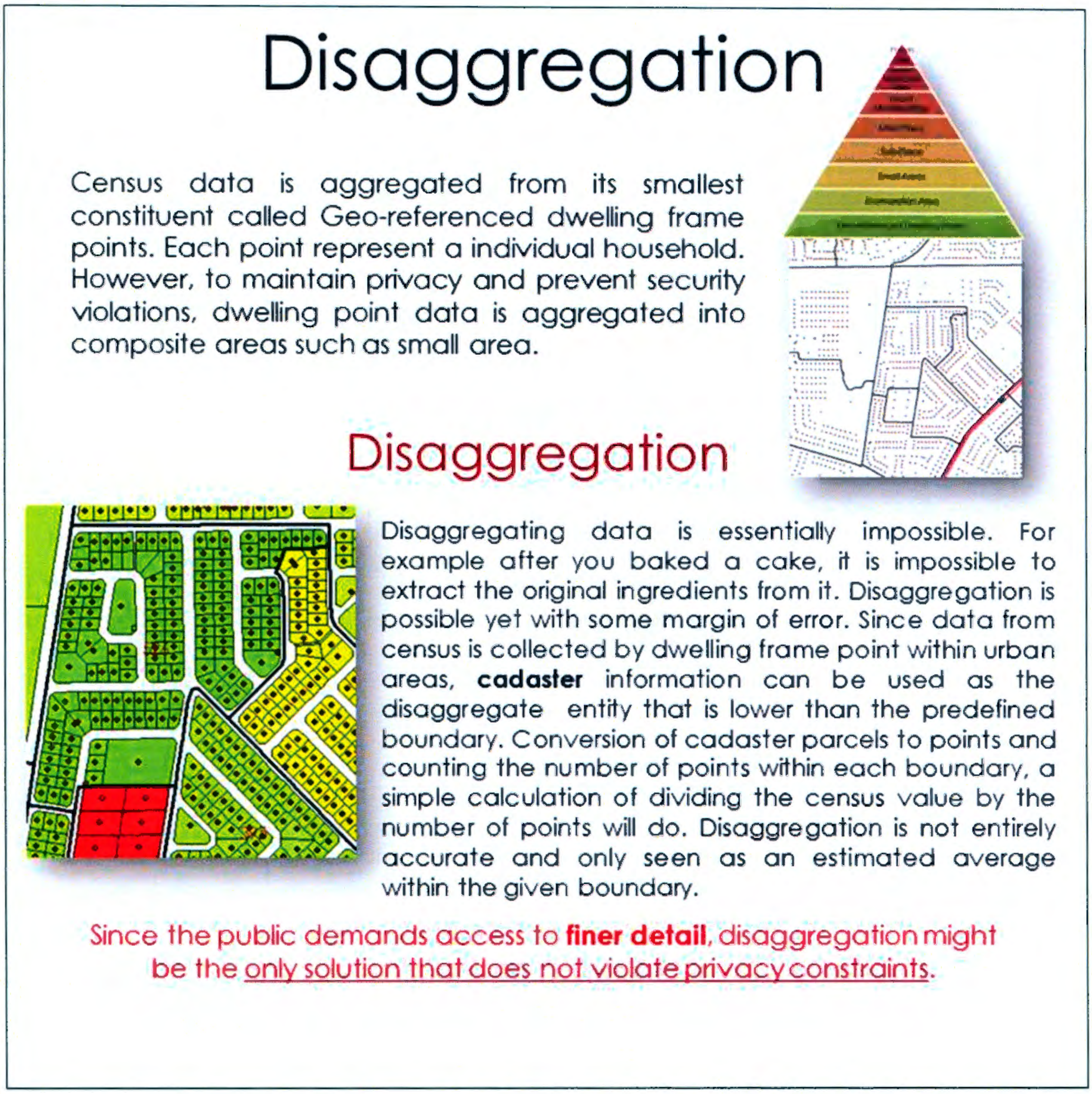
"Almost Never" category. Due to the nature of network analysis and the various steps involved in integrating census successfully with network service areas only a small percentage (17.4%) uses it frequently and 21.7% occasionally.

Table 19 Usefulness of network analysis with census data for decision making

Frequency	Not Useful	Barely Useful	Fairly Useful	Useful	Very Useful
	(1)	(2)	(3)	(4)	(5)
Score	3	4	2	3	11
Total %	13	17.4	8.7	13	47.8
			Average Score 3.65		

In spite of the technical difficulties to integrate census data into network analysis, users in Table 19 nevertheless deem it as an extraordinarily useful, integrative use, with 47.8% classifying it as very useful indeed. Only 30.4% classified it as not useful. Network analysis integration into census data is extremely valuable information for decision makers for estimation purposes. This is proven by the 69.5% that ranks it as useful on the usability index. Oftentimes more advance levels of integration, such as network analysis and zonal statistics offer decision makers very valuable information, but it is infrequently used due to the technicalities involved in generating such answers, shown by the large difference in frequency and usability scores.

4.2.10. Disaggregation



Disaggregation

Census data is aggregated from its smallest constituent called Geo-referenced dwelling frame points. Each point represent a individual household. However, to maintain privacy and prevent security violations, dwelling point data is aggregated into composite areas such as small area.

Disaggregation

Disaggregating data is essentially impossible. For example after you baked a cake, it is impossible to extract the original ingredients from it. Disaggregation is possible yet with some margin of error. Since data from census is collected by dwelling frame point within urban areas, **cadaster** information can be used as the disaggregate entity that is lower than the predefined boundary. Conversion of cadaster parcels to points and counting the number of points within each boundary, a simple calculation of dividing the census value by the number of points will do. Disaggregation is not entirely accurate and only seen as an estimated average within the given boundary.

Since the public demands access to **finer detail**, disaggregation might be the only solution that does not violate privacy constraints.

Figure 34 Disaggregation of census data for micro analysis

The last aspect addressed in the CENGIS framework focused on disaggregation. What census users seldom realise is that census is the most used dataset in disaggregation purposes. Micro-scale analysis is extremely powerful in strategic planning, even more powerful than the most detailed layer in census such as the SAL. Because census data is originally collected from points and aggregated into enumerated areas, users seldom realise just how different the areal unit is to the distribution of samples on the ground. Disaggregation of census data is based on reliable underlying land use patterns. Due to privacy constraints, census data cannot be made available on point level, which should be ideal. Disaggregation essentially allows for higher quality datasets without violating privacy constraints.

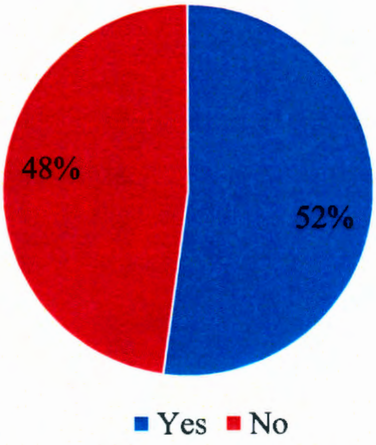


Figure 35 Use of disaggregated census data

Disaggregated use of census data is considered a very high level of integrative use between census and GIS, proven by the fact that 48% has not used disaggregation with census before shown in Figure 35. Although census is the most frequent source considered for disaggregation, only 52% has done so with census data before. The tremendous technical requirements in generating disaggregated data may be considered too complex for novice GIS users as illustrated in the 50|50 scenario.

Table 20 Frequency of using disaggregated census data

Frequency	Never (1)	Almost Never (2)	Occasionally (3)	Often (4)	Very Often (5)
Score	13	3	5	1	1
Total %	56.5	13	21.7	4.3	4.3
			Average Score 1.87		

According to the frequency of use shown in Table 20, disaggregated data scored by far the lowest, with 1.87, which falls within the "Almost Never" category. The reason for this is due to the complexity of disaggregating census data into reliable underlying ground use patterns. This is proven by only 8.6% of users using it frequently. Only 21.7% use it occasionally. The fact that 56.5% never use disaggregated census data is another indicator on the level of integration it relates to between census and GIS.

Table 21 Usefulness of disaggregated census data for decision making

Frequency	Not Useful (1)	Barely Useful (2)	Fairly Useful (3)	Useful (4)	Very Useful (5)
Score	5	1	3	5	9
Total %	21.7	4.3	13	21.7	39.1
Average Score 3.52					

In spite of the technical hurdle in generating disaggregated data, participants in Table 21 nevertheless indicated its usefulness, with 39.1% classifying it as very useful. The above average score of 3.52 ranks disaggregated census data as useful for decision makers in general. However, 26% considered it as non-useful. This question on integration proves yet again that the level of integration is more advance when the difference between the frequency of use score and the usability score is large. Disaggregated census data has proven to be the most strategic level of integrative use in the CENGIS framework.

4.3.CONCLUSION

The final results of the CENGIS framework are tabulated in the following table. By using the frequency and usability scores as guidelines to rank the complexity of the different question on integration, a hierarchical ranking from easy (top) to advance (bottom) is produced. Depending on the margin between the frequency score and the usability score, the greater the difference the more complex the integrative concept proves to be. The first column of the table assesses active use of the given concept – meaning the percentage of those who answer "yes". The second column features the frequency score between 1-5 on the Likert scale. The third column similarly features the usability score for decision makers between 1-5 on the Likert scale. The last column computes the difference between the usability and frequency score to rank the concepts in terms of integration complexity between census and GIS.

Table 22 CENGIS matrix on integrative use between census and GIS

Integration Aspect	Active Use (%)	Frequency of Use	Usability for Decision Making	Complexity
1. Standard Use	96	3.26	3.56	0.3
2. Representation	70	2.74	3.48	0.74
3. MAUP	52	2.78	3.52	0.74
4. Map Production	74	2.73	3.56	0.83
5. Spatial Aggregation	65	2.91	3.78	0.87
6. Time Series	52	2.22	3.21	0.99
7. Custom Use	64	2.52	3.60	1.08
8. Zonal Statistics	48	2.30	3.52	1.22
9. Network Analysis	52	2.30	3.65	1.35
10. Disaggregation	52	1.87	3.52	1.65
Total	62.5	2.56	3.54	0.98

Firstly, noting the ordering of the integrative aspects in Table 22 from 1-10, the more advance concepts on integration such as zonal statistics, network analysis and disaggregation one can clearly identify. As a general rule, technicality transposes into a higher level of complexity. Deducting from the large gap between the frequency of use and the usability for decision makers, those concepts that remain very useful yet are seldom used, are for more experienced GIS users. This is also illustrated by the active use percentage that declines as complexity increases; this means an average of 62.5% has at some point utilised these integrative aspects between census and GIS. Also note how the frequency of use declines from 1-10 on Table 22 which means a higher level of integration is less frequently used than simpler and more novice ones such as standard tabulation and representation. The average score on frequency of use is 2.56, which is much lower than the usability ranking. In spite of the decline in frequency, the usability score of each integrative aspect remains relatively high with an average of 3.54 out of 5. In conclusion the CENGIS framework on integrative use evaluates frequency of use to be 39% and usability to be 63.5%, respectively. This implies that integrative use depending on frequency and usability for census and GIS is 51.2%. This figure is, of course, relative to the aspects covered within the CENGIS framework and would of necessity change by either including more aspects on integration or changing existing ones.

CHAPTER 5: CONCLUSION AND FUTURE RESEARCH

5.1. REVISITING THE AIMS AND OBJECTIVES

To evaluate integrative use between census and GIS, a predefined framework based on different aspects of integration was necessary. The technical aspects on the side of GIS and the wealth of information embedded within census can be integrated to facilitate strategic decision making. The overarching aim was to evaluate the integrative use between census and GIS for planning support. To answer this question a comprehensive framework covering key aspects on integration was necessary to evaluate the relationship between census and GIS systematically. Through the use of the CENGIS framework on integration, each aspect can be graded by the participants and scored in terms of frequency of use and usability for decision makers. Taking these variables in account, an ultimate integrative score can be computed in response to the answer on integrative use between census and GIS.

5.2. SUMMARY OF FINDINGS

5.2.1. In Relation to the census and GIS integration

Despite the revolution of GIS caused in planning related activities, underutilisation of this technology is evident (Chapin, 2003, p. 1). This is confirmed by the 39.1% score on frequency of use in the CENGIS framework. Apart from the underutilisation, the predominant use of census data is found in the governmental sphere as of the 1990s (WeiWei & WeiDong, 2015, p. 1). This is confirmed in the 57% share in terms of sector for planning participants in the CENGIS framework. Since 2001 the geographic component of census has been steadily improved, especially with the availability of the small areas layer, which is in close proximity to the enumerated area layer. The most detailed layer in census SAL is extensively used to demonstrate different integrative aspects, such as aggregation, disaggregation and zonal statistics.

With regard to elementary aspects on integration, geo-referenced data can now be seamlessly joined in GIS through geographically defined queries. As seen in the usability score from the custom query aspect, 65% deem advance tabulation to be the most useful for decision makers. According to Burrough (2001) one of the key issues that GIS primarily focuses on was automated map making. Sadly, however, according to the CENGIS framework, only 47% uses this function frequently, despite the fact that census lends itself toward extensive automated mapping. Automatic Labelling and dynamic attributes greatly reduce the time needed to generate decent quality maps (Freeman, 2005, pp. 290-291). This is proven by the 64% who regard it as a useful resource for decision makers. According to Monmonier (1991) underlying data oftentimes requires different cartographic depictions for clarity, however only 43% uses different representations of the same data when making cartographic depictions of census data, which is vital to avoid intentionally misleading decision makers.

Spatial aggregation causes underlying heterogenic patters to be disregarded, which is inevitable in the case of census areal units (Dumedah, Schuurman, & Yang, 2008, p. 48). With 70% uses different spatial levels of aggregation to support decision making, which effectively address distortions due to scale (Jacobs-Crisioni *et al.*, pp. 52-53). When aggregating point data into predefined area units, the Modifiable Areal Unit Problem (MAUP) cannot be ignored; however, it is a worrying prospect on census and GIS integration with only 52% of the participants claiming to be aware of the MAUP. The arbitrary subdivision of areal units in aggregation is due to the geographic delineation of census boundaries (Jacobs-Crisioni *et al.*, 2014, p. 48). Despite the fact that its unavoidable, participants still classify maps with the MAUP as 63% useful for decision makers (Dumedah *et al.*, 2008, pp. 48-49). The use of more homogeneous zones, such as the small areas layer, significantly reduces the margin of misrepresentation for maps containing the MAUP (Reidl *et al.*, 2006, p. 900). Lastly, the aspect of comparing previous census datasets with newer ones is not recommended due to the decennial interval size (Salvo & Lobo, 2006, p. 226). This is confirmed by the infrequent use of older census data, with a mere 31%. The spatial resolution of previous census data is often not recorded on the same level of detail as newer census datasets (Radebe, 2015). The high variability of census boundary changes at sublevels further undermine the usefulness; however, users still rank it as 55% useful (Masser *et al.*, 1996, p. 91).

Lastly, on advance integrative use, according to Wu (2008) decomposition of census data for population estimates is useful for calculation of areas either smaller or irregular to the census area units. Converting census units into their corresponding raster representations (pixels) is one of the most common decomposition techniques (Spiekermann & Wegener, 1999, p. 1). However complexity still calls for some more comprehensive experience which is confirmed by the fact that only 48% have used zonal statistics in the past, with a 30% score on frequency of use. However usability is rated at 64%. Determination of population thresholds and coverage using network analysis in conjunction with census data is the only viable approach to answer the prescribed standard set forth under CSIR for social facilities (CSIR, 2012, pp. 11,24). However, despite its usefulness only 52% uses this function with a 30% score on frequency of use. In spite of the technicalities involved (Oh & Jeong, 2007, pp. 28-30), network analysis remains a very useful form of integration for decision makers, with a 64% score in terms of usefulness. Lastly to escape the tyranny of zones through disaggregation, as mentioned by Spiekermann and Wegener (1999), is a great improvement for micro scale analysis. The usefulness of disaggregated data is proven by the 39.1% of participants that rank it as very useful. However with the difficulties involved in generating disaggregated census data only 52% have attempted, with a 22% score in as to of frequency of use. Disaggregation is the highest form of integrative use in this study, shown by the large gap between frequency of use and usability.

5.3. CONCLUSION

Census has come a long way with the first official census conducted under Servius Tellius of Rome. Historically, census transformed the military and political outlook from a mere barbarian horde to a populous capable of collective action. Today census has progressed to a highly sophisticated enumeration profile covering all aspects of social life and economic activity. Administrative action relies heavily on census information for evidence-based decision making. The geographic component of census seamlessly integrates into modern day Geographic Information Systems (GIS) to enhance spatial decision making as never before. The combination between the ancient origins of census 500 B.C. and the modernistic technological breakthrough of GIS 1950 A.D. serves as an excellent example of how ancient and new technology can seamlessly integrate due to geography. Strategic planning has come to the fore with census and GIS to facilitate broad-scale-planning initiatives, especially in government. However, despite the wealth of information in census that can be extracted with GIS, integration between the two has been predominantly underutilised. Thus this project is intended the evaluate integrative use between census and GIS by means of a custom framework called CENGIS that addresses ten aspects on integration.

The fact that census data is conducted decennially and independent of political interference, and it remains one of the primary sources in planning support, decision making and monitoring of governmental policies. South Africa has a long track record of disparate census activities dating as far back as the 18th century. Today it is still the most comprehensive data collection attempt in South Africa with 15 million surveys distributed in 2011. The fact that census is intended to counts *100%* of the population makes sets it as superior to all other sampling methodologies. Despite the significant benefit census data offers to decision makers when integrated with GIS it remains largely underutilised, and only started gaining momentum as of 2003 when introduced to the education curriculum, and more interdisciplinary sciences started to incorporate it. Census is one of the main sources that can take full advantage of GIS analysis. Since 1996 enumerated areas were captured digitally and linked spatially forming the basis of integration between census and GIS.

Since the introduction of the Small Areas Layer (SAL) in 2011, GIS analysis has grown exponentially. The availability of tabulation software such as SuperCROSS allows users to generate custom queries to be used in GIS for cartographic displays. Dynamic map making in GIS greatly reduces the time needed to generate high-quality and informative cartographic displays complemented with dynamic attributes derived from census. Excellent generalisation capabilities such as automated map annotation in modern day GIS software dramatically reduces the time needed to composite high-quality cartographic outputs. Different cartographic depictions of the same underlying census data using different classification schemes and interval size dynamically enhance usability. Although census data is collected from individual dwellings, data gets aggregated into census areal units to keep participants identities hidden, which unfortunately give rise to modifiable area unit problem (MAUP). The effect of this is often time exhibited when comparing older census data with newer ones. To escape this effect the hard boundary constraint conversion of census areal units to corresponding raster representation can be sampled by using zonal statistics for reliable estimations. Similarly disaggregating census data by using reliable underlying land use patterns has proven to be a significant enhancement for decision makers who wish to use census data on micro-scale analysis.

To evaluate integrative use between census and GIS, a custom framework was developed called CENGIS, which takes ten aspects of integration into account based on the literature in chapter two. Although these ten aspects are not the alpha and omega, they serve as a general guide on integrative use between census and GIS. Every aspect was evaluated using an illustrative explanation with supportive examples, from which the participants can answer the question on active use, frequency of use and usability for decision makers. The ten aspects covered ranges from basic to more advance concepts concerning integration between census and GIS. Firstly looking at tabulation, the automated cartographic map generation, the concept of multiple levels of spatial aggregation, using census data with custom boundaries through zonal statistics, looking at representation through classification schemes and intervals size, addressing the modifiable area unit problem, comparing previous census data with newer ones for time series analysis, using census for threshold estimates and disaggregating census data for micro-scale analysis.

Since the overarching aim of this research projects is to evaluate integrative use between census and GIS, the CENGIS framework provides the means to answer various aspects on integration. The governmental sector is the primary user of census data for planning purposes, followed by academic institutions. Active use was evaluated using a close-ended question. Turns out, according to all ten aspects on integrative use, that only 62.5% actively utilised census data according to the aspects prescribed in the CENGIS survey. Frequency and usability were assessed using a Likert scale that ranges form 1-5. With regards to frequency wise integrative use stands at 39% and overall usability for decision makers stands at 63.5%. According to the CENGIS framework the overall integrative use between census and GIS is 51.2% which confirms the original assumption that census remains largely underutilised by decision makers.

5.3.1. Future Research

The CENGIS framework focused on ten primary aspects of integrative use between census and GIS. However, in addition to these ten aspects, other aspects can be evaluated in future studies by either expanding on the existing framework or creating a new one. Evaluation of integrative use can be optimized by having a more comprehensive set of indicators, in addition to frequency and usability. Besides indicators, this project focused solely on those in planning-related professions. Choosing a different audience will produce different results. Also, to evaluate integrative use more scientifically refinement of scales is another aspect that can be considered for future research. According to the responses received, future research can also include integrative use between census data and statistical applications such as Statistical Analysis Software (SAS) in unison with GIS extensions, from which models can be derived on regression and forecasting analysis. Others have specified that the main concern with using census data was spatial resolution. Conducting a comparative study based on enumerated areas vs. small areas is proposed.

BIBLIOGRAPHY

Bahgat, K. (2015). *Python Geospatial Development Essentials.* UK: Packt Publishing Ltd.

Browne, T. J., & Fielding, A. J. (1987). Automating map production of the 1981 Census data for Brighton and Hove, England. *The Visual Computer, 3*, 82-87.

Burrough, P. A. (2001). GIS and geostatistics: Essential partners for spatial analysis. *Environmental and Ecological Statistics, 8*, 361-377.

Chainey, S., & Ratcliffe, J. (2013). *GIS and Crime Mapping.* New York: John Wiley & Sons.

Chapin, T. S. (2003). Revolutionizing the core: GIS in the planning curriculum. *Environment and Planning B: Planning and Design, 30*(1), 565-573.

CSIR. (2012). *CSIR guidelines for the provision of social facilities in South African settlements.* Council for Scientific and Industrial Research, Pretoria.

Dark, S. J., & Bram, D. (2007). The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography, 31*(5), 471-479.

Dodge, M., Kitchin, R., & Perkins, C. (2011). *The Map Reader: Theories of Mapping Practice and Cartographic Representation.* New York: John Wiley & Sons.

Dumedah, G., Schuurman, N., & Yang, W. (2008). Minimizing effects of scale distortion for spatially grouped census data using rough sets. *Journal of Geographical Systems, 10*, 49-69.

Elangovan, K. (2006). *GIS: Fundamentals, Applications and Implementations.* New Dehli: New India Publishing.

Felke, T. P. (2014). Building Capacity for the Use of Geographic Information Systems (GIS) in Social Work Planning, Practice, and Research. *Journal of Technology in Human Services, 32*(1), 81-92.

Fischer, M. M. (2006). GIS and Network Analysis. In M. M. Fischer, *Spatial Analysis and GeoComputation: Selected Essays* (pp. 43-60). Springer Science & Business Media.

Frank, A. U. (2005). Map Algebra Extended with Functors for Temporal Data. In J. Akoka, *Perspectives in Conceptual Modeling: ER 2005 Workshop AOIS, BP-UML, CoMoGIS, ECOMO, and QoIS* (pp. 194-207). Klagenfurt: Springer Science & Business Media.

Freeman, H. (2005). Automated cartographic text placement. *Pattern Recognition Letters, 26*, 287-297.

Gibson, J., Deng, X., Boe-Gibson, G., Rozelle, S., & Huang, J. (2011). Which households are most distant from health centers in rural China? Evidence from a GIS network analysis. *GeoJournal, 76*, 245-255.

Gordon, C. (2011). Lost in Space, Or Confessions of an Accidental Geographer. *International Journal of Humanities and Arts Computing, 5*(1), 1-22.

Government. (1999). *Statistics Act No 6 of 1999.* Pretoria: Government Printer.

Gregory, I. N., & Healey, R. G. (2007). Historical GIS: structuring, mapping and analysing geographies of the past. *Progress in Human Geography, 31*(5), 638-653.

HSRC. (2011). *The Applications of GIS in Census Data.* Pretoria: Human Sciences Research Council.

Jacobs-Crisioni, C., Rietveld, P., & Koomen, E. (2014). The impact of spatial aggregation on urban development analyses. *Applied Geography, 47*, 46-56.

Jahanshiri, E., Shariff, A. R., Amiri, F., Soom, M. A., Wayayokb, A., Buyonga, T., et al. (2015). Spatial soil analysis using geostatistical analysis and map Algebra. *Arabic Journal of Geoscience, 8*, 9775-9788.

Jobson, T. A., Rooyen, M. M., Reynecke , C. D., Biljon, W. R., & Scheepers, C. F. (1986). *The KANDELAAR geographic information and map compilation system: functional description.* Pretoria: CSIR.

Kakembo, V., & van Niekerk, S. (2014). The integration of GIS into demographic surveying of informal settlements: The case of Nelson Mandela Bay Municipality, South Africa. *Habitat International, 44*, 451-460.

Koua, L. E., & Kraak, M.-J. (2004). Geovisualization to support the exploration of large health and demographic survey data. *International Journal of Health, 3*(12), 1-13.

Lehohla, P. (2005). *Statistics needs Geography; Geography needs Statistics.* Pretoria: StatsSA.

Liederman, C. (2015, January 01). *History.* Retrieved January 13, 2016, from Department of Geography & Environmental Studies: http://www0.sun.ac.za/geography/about-us/history/

MacDevette, D. R. (1993). The status of advanced information technology applications for natural resource management in southern Africa. *Proceedings of conference on 'Application of Advanced Information Technologies: Effective Management of Natural Resources*, (pp. 9-17). Spokane.

MacDevette, D. R., Fincham, R. J., & Forsyth, G. G. (1999). The rebuilding of a country: the role of GIS in South Africa. In P. A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind, *Geographical Information Systems: Principles and Technical Issues* (pp. 913-924). John Wiley & Sons.

Manan, M. S., & Hashim, N. R. (2010). GIS Visualization of Population Censuses in Peninsular Malaysia: A Case Study of Jempol, Negeri Sembilan, 1947-2000. *Pertanika Journal of Social Science and Humanities, 18*(2), 367-378.

Manley, D., Flowerdew, R., & Steel, D. (2006). Scales, levels and processes: Studying spatial patterns of British census variables. *Computers, Environment and Urban Systems, 31*, 143-160.

Masser, I., Campbell, H., & Graglia, M. (1996). *GIS Diffusion: The Adoption And Use Of Geographical Information Systems In Local Government in Europe.* Boca Raton, US: CRC Press.

McGrath, G., & Sebert, L. (1999). Mapping a Northern Land: The Survey of Canada 1947–1994. In G. A. Toomey, & R. Tomilson, *GIS and LIS in Canada* (pp. 1947-1994). McGill-Queen's University Press.

Monmonier, M. (1991). *How the Lie with Maps.* Chicago: University of Chicago.

Nyerges, T., Couclelis, H., & McMaster, R. (2011). *The SAGE Handbook of GIS and Society.* Thousand Oaks: SAGE.

Oh, K., & Jeong, S. (2007). Assessing the spatial distribution of urban parks using GIS. *Landscape and Urban Planning, 82*, 25-32.

Openshaw, S. (1984). The Modifiable Areal Unit Problem. *Concepts and Techniques in Modern Geography, 38*, 40.

Paez, A., & Scott, D. M. (2004). Spatial statistics for urban analysis: A review of techniques with examples. *GeoJournal, 61*, 53-67.

Peters, A. H., & MacDonald, H. I. (2004). The census: An introduction. In A. H. Peters, & H. I. MacDonald, *Unlocking the Census with GIS* (pp. 3-37). Sandiago: ESRI Press.

Prouse, V., Ramos, H., Gran, J. L., & Radice, M. (2014). How and when Scale Matters: The Modifiable Areal Unit Problem and Income Inequality in Halifax. *Canadian Journal of Urban Research, 23*(1), 61-82.

Radebe, J. (2015, May 6). *Statistics South Africa Dept Budget Vote 2015/16.* Retrieved November 2, 2015, from South African Government: http://www.gov.za/speeches/minister-jeff-radebe-statistics-south-africa-dept-budget-vote-201516-7-may-2015-0000

Reidl, A., Kainz, W., & Elmes, G. A. (2006). *Progress in Spatial Data Handling.* Heidelberg: Springer-Verlag Berlin Heidelberg.

Salvo, J. J., & Lobo, A. P. (2006). Moving from a decennial census to a continuous measurement survey: factors affecting nonresponse at the neighborhood level. *Population Research Policy Review, 25*, 225-241.

Sheckhar, S., & Xiong, H. (2008). *Encyclopedia of GIS.* New York: Springer Science and Business.

Southall, H. (2011). Rebuilding the Great Britain Historical GIS, Part 1. *Historical Methods, 44*(3), 149-159.

Spiekermann, K., & Wegener, M. (1999). Freedom from the Tyranny of Zones: Towards New GIS-Based Spatial Models. In S. Fotheringham, & M. Wegener, *Spatial Models and GIS: New and Potential Models* (pp. 45-61). London: Taylor & Francis.

StatsSA. (1996). *The People of South Africa Population Census, 1996.* Pretoria: StatsSA.

StatsSAa. (2011). *Census Information Guide.* Pretoria: StatsSA.

StatsSAb. (2011). *How the count was done.* Pretoria: StatsSA.

StatsSAc. (2011). *Population Census 2011: Strategy.* Pretoria: StatsSA.

StatsSAd. (2011). *Statistical release (Revised).* Pretoria: StatsSA.

Steiniger, S. (2007). *Enabling Pattern-Aware Automated Map Generalization.* Zurich: Universit of Zurich.

SuperCROSS. (2012). Space Time Research. *User Guide*. Melbourne, Australia: Space-Time Research Pty Ltd.

Tenney, F. (1930). Roman Census Statistics from 508 to 225 B.C. *American Journal of Philology, 51*, 313-324.

Ware, J. M., Wilson, I. D., & Ware, J. A. (2003). A knowledge based genetic algorithm approach to automating cartographic generalisation. *Knowledge-Based Systems, 16*, 295-303.

WeiWei, L., & WeiDong, L. (2015). GIS : Advancement on Spatial Intelligence Applications in Government. *The Open Cybernetics & Systemics Journal, 9*, 587-593.

Wu, S.-s., Wang, L., & Qiu, X. (2008). Incorporating GIS Building Data and Census Housing Statistics for Sub-Block-Level Population Estimation. *The Professional Geographer, 60*(1), 121-135.

APPENDIX A

View the CENGIS survey online at:

tinyurl.com/cengis2015