Evaluating Data Classification Methods for Choropleth Maps in South Africa:

A Usability Study

by

Lourens Fourie Snyman

A thesis submitted in partial fulfilment of the requirements for the degree

Doctor of Philosophy

in the Department of Geography, Geoinformatics and Meteorology

at the

University of Pretoria

Faculty of Humanities

Supervisor:

Serena Coetzee

12 July 2025

DECLARATION

I, Lourens Fourie Snyman, declare that the dissertation/thesis, which I hereby submit for the degree Doctor of Philosophy at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution. Grammarly was used to verify spelling and grammar.

SIGNATURE: 

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to thank and acknowledge the following people for their insights, love and support.

Thank you to my wife Daleen, for all your love, support and encouragement. My supervisor Serena Coetzee, for your advice, guidance, and patience.

Thank you to all the respondents who participated in the user study. Without your valuable inputs, this thesis would not have been possible.

May we never stop exploring!

ABBREVIATIONS AND ACRONYMS

ADCM	Absolute Deviations from the Class Medians
DPSA	Department of Public Service and Administration
GADF	Goodness of Absolute Deviation Fit
GGM	Geography, Geoinformatics and Meteorology
GIS	Geographic Information System
GVF	Goodness of Variance Fit
IEC	Electoral Commission of South Africa
IT	Information Technology
PPMCC	Pearson Product Moment Correlation Coefficient
sig.	Significance
SPSS	Statistical Package for Social Science

ABSTRACT

Today, location is entrenched in many organisations in both the public and private sectors. Organisations, both locally and internationally, are realising the importance of location – and therefore maps – but do they know how to visually communicate this spatial knowledge?

Geospatial data visualisation techniques have evolved rapidly over the past decade. Today, most geographic information system (GIS) software has a plethora of built-in spatial analysis and visualisation techniques that enable users to quickly and effortlessly visualise spatial patterns in data. Choropleth maps are among the oldest and still one of the most frequently used techniques for visualising quantitative data in a GIS. The challenge with using choropleth maps in South Africa is selecting a data classification method that effectively displays unequal and dispersed population densities.

The aim of this research was to assess the suitability of different data classification methods for effectively visualising population demand using choropleth maps in South Africa. The research focused on geographic accessibility as a use case where choropleth maps are used to visualise population demand, allowing decision makers to identify over- or underserved areas for the provisioning of service centres. This was achieved with a user study. The user study included the design of an online questionnaire featuring map interpretation questions specifically related to geographic accessibility. Subsequently, the results from the user study were compared to a recommended mathematical equation that measures the error between class breaks, in a data classification method.

The user study shows that respondents were more likely to provide correct answers when presented with maps using the quantiles and natural breaks (Jenks) data classification methods, suggesting that these methods are easier to interpret and analyse for understanding population distribution in South Africa. A goodness of variance fit calculation that measures the error between class breaks delivered somewhat different results. Based on these calculations, natural breaks (Jenks) and geometric interval were considered the optimal data classification methods, while logarithmic scale and quantiles were ranked lowest.

Based on the results of both the user study and error calculations, a more comprehensive view of the use of data classification methods was obtained. This research emphasises the importance of including human interpretation when assessing methods or techniques used to represent spatial phenomena.

1. INTRODUCTION AND BACKGROUND INFORMATION

1.1 Overview

"Use a picture. It's worth a thousand words" (Arthur Brisbane, 1911). But then, "If a picture is worth a thousand words, then a map must be worth a million" (Patrick Abbott, 2009).[1] Pictures such as maps or charts, as powerful as they are, can also mislead or confuse the audience. Communicating information clearly and effectively is greatly influenced by the type and complexity of the data, as well as the intended target audience.

Today, location is deeply ingrained in many organisations, both in the public and private sectors. Many use location data and perform some form of data analysis. Regardless of their core business, companies and organisations, both locally and internationally, are realising the importance of location and, therefore, maps. They are increasingly becoming aware of the possibilities of spatial thinking and geographic data analysis, which can enable them to increase revenue, save money, maintain a competitive edge in their field of expertise, potentially enhance client or customer satisfaction, and improve overall service delivery for citizens. But do they know how to visually communicate this spatial knowledge?

Data visualisation, or rather geospatial data visualisation techniques, has evolved rapidly over the past decade. Today, any geographic information system (GIS) software application offers a plethora of built-in spatial analysis and visualisation techniques, including choropleth maps, kernel density estimation heat maps, firefly maps, dot maps, graduated symbols, point density or isochrones, among others. These enable users to visualise spatial patterns in data quickly and effortlessly. With the increased processing capability of desktop computers, high-speed internet connections, cloud computing (Zhang et al., 2019), and parallel processing (Zhao et al., 2016), analysing and displaying large volumes of geographic data has become easier and faster, making it accessible to anyone.

Current proprietary and open-source GIS software development companies and communities are continually simplifying their applications, allowing non-GIS professionals to execute advanced spatial queries and visualisation techniques with just a few clicks. User interfaces are designed to guide end users, regardless of their GIS background, through logical steps to analyse geospatial data and prepare maps. Whether it is to create a topographic or choropleth map or to visualise travel distances with isochrones to optimise service locations, the software guides the user through the analysis and visualisation in a stepwise approach. Although these steps enable non-GIS users to create maps, the software cannot guarantee that the story is

[1] https://blog.education.nationalgeographic.org/2009/11/20/patrick_abbott-_a_map_is_worth_a_million_words/

told effectively and that the visual message is communicated. “Because anyone with the right software and an internet connection can now make and publish maps, mapmakers can also easily lie to themselves and others” (Monmonier, 2018).

Due to its history, the geographic distribution of the South African population presents unique challenges for visualising spatial patterns on a map for effective decision-making. The Group Areas Act 41 of 1950 “prohibited the multiracial use or occupation of urban land” (Strauss, 2019). Segregated zones were created in urban areas, allowing only certain racial groups to live and work there. This resulted in the establishment of densely populated townships and informal settlements outside city centres and suburban living spaces. Mostly poor people reside in these areas, and access to basic services and public service centres is limited and insufficient. Connecting the population to public services requires decision makers to understand where the population demand is for these services in relation to public service facilities (Snyman & Coetzee, 2024). Population demand refers to the “number of people who may need the services” (Ma et al., 2018). To determine this, supply (service centres or facilities) and demand (population) maps are needed to visualise potential gaps and shortfalls. Generally, choropleth maps have been proven effective in visualising the population distribution when expressed in densities or ratios (Tyner, 2014).

Choropleth maps are one of the oldest and most frequently used techniques for visualising quantitative data in a GIS (Tyner, 2014). Slocum et al. (2014) noted that a choropleth map is “the most commonly used (and abused) method of thematic mapping”. A choropleth map categorises observations into several classes, either manually or based on a data classification method. Today, GIS software offers a range of data classification methods for creating choropleth maps, allowing users to easily select and apply a method from a drop-down menu. Each data classification method has its advantages and disadvantages, which can vary based on factors such as the geographic scale and spatial distribution of the data. Kraak et al. (2021) pointed out that classification could potentially increase uncertainty since patterns derived from a classification are influenced by the positioning of class breaks. Choosing the wrong or inappropriate data classification method to visualise data with choropleth maps can potentially distort spatial patterns, causing misleading representations and a potential oversimplification of information (Brewer, 2006; Evans, 1977; Monmonier, 2018). However, the target audience, such as policymakers, who will be interpreting the maps should also be considered (Tyner, 2014).

The challenge with data classification methods for choropleth maps in South Africa is selecting a classification method that effectively displays the country’s unequal and dispersed population densities (demand). It should emphasise not only the city centres and their

surroundings but also secondary or tertiary populated areas, such as townships and informal settlements, that are segregated from the city centre.

1.2 Problem Statement

South Africa is characterised by an uneven population distribution, with unequal access to service facilities. A World Bank report on poverty and inequality describes South Africa as one of the most unequal countries in the world, stating that “inequality has increased since the end of apartheid in 1994” (World Bank, 2018). Approximately 67% of the country’s population lives in urban spaces, with an estimated urban growth rate of 1.97% (UN-Habitat, n.d.). This affects service delivery, resulting in an ever-increasing unemployment rate and rising levels of crime. Understanding and managing the growing population’s demand for current and available services requires a geographic visualisation method that depicts demand and supply, allowing decision makers to identify optimal locations for service facilities. Choropleth maps are one of the oldest and most frequently used geospatial data visualisation techniques for analysing and visualising population densities (Monmonier, 1993; Tyner, 2014).

A major challenge associated with choropleth maps is choosing a suitable data classification method (Slocum et al., 2014) that displays and communicates the data distribution clearly and effectively. Using the wrong data classification method can distort spatial patterns, providing a misleading representation of population densities and identifying wrong or inappropriate locations for positioning service centres. This, in turn, can lead to poor service delivery and frustrated citizens. Additionally, the question arises as to whether a target audience, such as decision makers or map users, can use choropleth maps that depict population demand to identify areas with inadequate access to services, as well as those that are overserved.

1.3 Research Questions

The following research questions about data classification methods for choropleth maps were examined:

1. Which data classification methods are frequently used to visualise statistical data with choropleth maps?
2. Which measurement techniques are used to evaluate the suitability and effectiveness of these data classification methods?
3. For a geographic accessibility analysis of service centres in South Africa, which data classification method(s) for choropleth maps depicting population demand are best interpreted by map users or decision makers, and which are also statistically proven to be effective?

1.4 Research Aim and Objectives

This research aimed to assess the suitability of different data classification methods for effectively visualising population demand in South Africa using choropleth maps. The research focused on geographic accessibility as a use case in which choropleth maps were used to visualise population demand, allowing decision makers to identify areas that are either over- or underserved in the provision of service centres.

The following primary objectives were defined to achieve the research aim:

- **Objective 1** (Research Question 1) – Review literature on data classification methods for visualising demographic data with choropleth maps.
- **Objective 2** (Research Question 2) – Identify methods for evaluating the effectiveness of data classification methods for choropleth maps.
- **Objective 3** (Research Question 3) – Identify the most suitable data classification method(s) for visualising population demand for decision makers in geographic accessibility studies in South Africa.
- **Objective 4** (Research Question 3) – Statistically assess the effectiveness of data classification methods depicting population distribution in South Africa.

1.5 Contributions and Significance of the Research

The contributions of this research include:

- A comprehensive review of data classification methods for choropleth maps and an evaluation of their effectiveness.
- An empirical assessment of the effectiveness of the various classification methods for visualising the unique population distribution in South Africa, based on a user study (see Chapter 4).
- An empirical calculation and evaluation of the classification methods using the mathematical equations recommended in the literature (Chapter 5).
- An example and guidance for other countries seeking to assess the suitability of choropleth maps for visualising population distribution.

The research makes a significant contribution to cartography, as the effectiveness of data classification methods for choropleth maps has not been previously evaluated for South African data intended for a South African target audience. This study is unique in that its use case focuses on data visualisation for geographic accessibility to service centres in South Africa, where settlement typologies are highly diverse and dynamic. Furthermore, the concept of geographic accessibility analysis is closely related to various sustainable development

goals (SDGs) which were adopted by all United Nations Member States in 2015. Adequate access to infrastructure (SDG 9 - Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation), medical and health services (SDG 3 - Ensure healthy lives and promote well-being for all at all ages) or green spaces (SDG 11 - Make cities and human settlements inclusive, safe, resilient and sustainable) promote health and well-being for all citizens.

The results of this research will be valuable for professionals who prepare choropleth maps for various applications, such as the Electoral Commission of South Africa to illustrate voter registration patterns and election results, as well as Statistics South Africa to present census results. The findings from this research could also be beneficial for other countries with similar population distribution characteristics and dynamics.

Lastly, the research benefits the broader academic community. The following abstracts were published in conference proceedings:

- Snyman, L., Coetzee, S., & Rautenbach, V. 2023. Evaluating data classification methods for choropleth maps to visualise geographic accessibility in South Africa: A usability study. *Abstracts of the International Cartographic Association*, 6, 241. https://doi.org/10.5194/ica-abs-6-241-2023
- Snyman, L., Coetzee, S., & Rautenbach, V. 2024. Assessing the suitability of data classification methods for choropleth maps depicting population distribution in South Africa. *Abstracts of the International Cartographic Association*, 7, 160. https://doi.org/10.5194/ica-abs-7-160-2024

1.6 Structure of this Dissertation

The remainder of the dissertation is structured as described below.

Chapter 2 addresses Research Questions 1 and 2. Firstly, the chapter describes cartographic maps and geospatial data visualisation, followed by a review of choropleth maps with a specific focus on the underlying data classification methods used to visualise statistical data. Secondly, various techniques for measuring and assessing the effectiveness of data classification methods for choropleth maps are discussed. Thirdly, this is followed by a brief historical overview of population distribution in South Africa, which describes the country's diverse population patterns and unequal access to service facilities. Lastly, the chapter explores the concept of geographic accessibility, which serves as the use case for this research, highlighting the key factors relevant to the optimal positioning of service facilities to ensure effective service delivery.

Chapter 3 first describes the materials and methods used for choosing appropriate or suitable study areas, geographic units, and data classification methods that will be evaluated for this research. Secondly, it outlines a series of choropleth maps depicting population demand.

Chapters 4 and 5, which address Research Question 3, evaluate the effectiveness of the selected data classification methods, initially through a user study, which is described in Chapter 4. The user study included the design of an online questionnaire, in which participants' interpretations of choropleth maps were assessed based on real-world scenarios. Additionally, a second approach is to calculate an accuracy score, as recommended in the literature. The accuracy score calculation measures the error between class breaks for each data classification method as described in Chapter 5.

Chapter 6 concludes with a summary of the findings and recommendations for using data classification methods to effectively visualise population demand in South Africa using choropleth maps. Also, limitations and shortcomings are highlighted, and potential topics for future research are identified and discussed.

The study includes the following supporting information:

- **Appendix A** is the approved ethics document.
- **Appendix B** includes a test for data normality.
- **Appendix C** includes the Qualtrics questionnaire, which was designed for the user study. Results are described and discussed in Chapter 4.
- **Appendix D** comprises the R Code that was used to measure the error between class breaks for each selected data classification method. Results are described and discussed in Chapter 5.

2. LITERATURE REVIEW

2.1 Introduction

The purpose of this chapter is to address Research Objectives 1 and 2, as defined in Section 1.4. Firstly, the chapter describes cartographic maps and geospatial data visualisation, followed by a review of choropleth maps with a specific focus on the underlying data classification methods used to visualise statistical data.

Secondly, various techniques for measuring and assessing the effectiveness of data classification methods for choropleth maps are discussed. This is followed by a brief historical overview of population distribution in South Africa, which describes the country's diverse population patterns and unequal access to service facilities.

The final section of the chapter explores the concept of geographic accessibility, which serves as the use case for this research, highlighting the key factors relevant to the optimal positioning of service facilities for effective service delivery.

2.2 Background: Cartographic Maps and Geospatial Data Visualisation

Kraak and Ormeling (2020) highlighted that geospatial data has a "specific location in space"; hence, these locations can be visualised and analysed on a map. They also argued that "It is not possible to get an overview of an area in any way other than by consulting a map" (Kraak & Ormeling, 2011). Some of the earliest examples of visualisation were by means of geometric diagrams, "in tables of the positions of stars and other celestial bodies, and in the making of maps to aid in navigation and exploration" (Friendly, 2008).

Maps, in particular, are subjective by nature and are described as partially displaying objective realities, while also partially showing subjective elements (Wright, 1942). Maps are further a distortion of reality based on three elements, namely map scale, map projection, and map symbolisation (Monmonier, 2018). Monmonier (2018) further commented in his book *How to Lie with Maps* that "Not only is it easy to lie with maps, it's essential". Thus, in general, a scientist or cartographer needs to manipulate spatially referenced data, including place boundaries, lines, point features, and raster imagery. This allows them to highlight and accentuate specific spatial elements or phenomena relevant to their research while distorting or hiding irrelevant data in an attempt to avoid clutter.

The data displayed on a map, along with the chosen method of visual presentation, convey information, whether it is a topographic orientation map or a statistical map depicting population densities. In most cases, these maps are designed for interpretation and analysis

by a target audience, enabling them to gain knowledge about a specific subject. This, in turn, empowers them to solve a particular problem. In the end, knowledge – sometimes referred to as insights – is transformed into wisdom. This flow from data to information, then to knowledge, and finally to wisdom has received significant attention over the years (Baskarada & Koronios, 2013; Bellinger et al., 2004; Hey, 2004). Bellinger et al. (2004) described information as "data that has been given meaning". For example, population density data that is visualised using a choropleth map depicts high- and low-density patterns, providing insight to those analysing population patterns. They further explain that knowledge is "the collection of information, such that it's intent is useful". Lastly, Bellinger et al. (2004) concluded that wisdom is a "uniquely human state" from which we as humans can derive a sense of understanding, "which there has previously been no understanding".

Several factors influence geospatial data visualisations, including (a) the level of measurement or measurement scale, such as nominal, ordinal, interval, or ratio (Kraak & Ormeling, 2020); (b) data resolution, also referred to as granularity (Slocum et al., 2014), or geographic units (Boscoe & Pickle, 2003), such as census blocks, wards, districts or hexagons; and (c) the data visualisation technique. GIS Geography (2019) highlights a number of these techniques. They include, among others, firefly maps, dot maps, graduated symbols, vector direction maps, distributive flow, density-equalizing cartograms, Voronoi diagrams, choropleth maps, heat maps (point density), isochrones, dasymetric maps, space-time cubes, topographic maps, contours (isolines), non-contiguous cartogram, Dorling cartogram, surface maps, schematics, and network flows.

Another important aspect to consider, especially when static maps are produced as map image files, is the cartographic scale. Cartographic scale refers to the relationship between the "map and earth distances" (Slocum et al., 2014). For example, a 1:50 000 cartographic scale indicates that one unit of distance measured on the map corresponds to 50 000 units of distance in reality. These factors play a vital role in the overall map designing process.

As expected, geovisualisation is not without its problems. As O'Sullivan and Unwin (2014) pointed out, the abundance of "graphic variables" and the ability to design "colourful, dynamic displays" can create problems. Also, "just because technology allows you to do something clever doesn't mean that you should" (O'Sullivan & Unwin, 2014). In 1942, Wright observed that because maps are created by people and not produced automatically, they are "influenced by human shortcomings". Hogräfer et al. (2020) described the concept of map-like visualisations as a technique in which abstract data are presented or incorporated into cartographic maps. In their overview of map-like visualisations, the problem of choosing the appropriate technique for visualising a specific problem or task was emphasised. They argued

that excessive schematisation of a visualisation or excessive map imitation could “confuse” the map reader, resulting in “misinterpretation of map symbolism” (Hogräfer et al., 2020).

2.3 Choropleth Maps

One of the oldest and still one of the most frequently used techniques for visualising quantitative data in a GIS is choropleth maps (Tyner, 2014). The term ‘choropleth’ originates from the Greek words *choros*, meaning area, and *plethos*, meaning value (Kraak & Ormeling, 2011). Slocum et al. (2014) noted that a choropleth map is “the most commonly used (and abused) method of thematic mapping”. A choropleth map is a simple, easy-to-use and easy-to-read technique for classifying and visualising data for statistical areas (Shaito & Elmasri, 2021; Tyner, 2014), such as enumerator areas, wards, municipalities or districts. Choropleth maps are also referred to as thematic maps. De Smith et al. (2018) defined choropleth maps as maps that display information about an area based on a particular theme, using techniques such as colour shading, categorised into different classes. Juergens (2020) stated that “an essential purpose of choropleth maps is the visual perception of spatial patterns”.

Generally, there are three types of choropleth maps: a simple choropleth map, a dasymetric choropleth, and an unclassed choropleth (Tyner, 2014). A simple choropleth map uses different colours or patterns (Tyner, 2014) to group spatial entities with similar characteristics, enabling users to recognise spatial patterns in the data. Simple choropleth maps are used frequently and are appropriate for analysing and visualising population densities and/or other intensity data measures (Monmonier, 1993). Dasymetric choropleth maps are more challenging to create, which is why they are used less often. The fundamental principle of a dasymetric choropleth map is to exclude areas within geographic units that are irrelevant for visualisation, such as uninhabited mountains, lakes, or industrial areas. Thereafter, densities are calculated based on the remaining areas relevant to the analysis, resulting in a more accurate distribution of the population (Barrozo et al., 2016). Unlike simple and dasymetric choropleth maps, an unclassed choropleth map does not categorise geographic units into distinct classes, so the data are not generalised. Instead, each geographic unit is displayed with a different pattern or colour intensity. Therefore, the map legend only displays low and high values based on the colour density.

To create simple choropleth maps, Robinson (1995) identified three elements, including the size and shape of areas or polygons, the number of classes, and the class limits. These are discussed in more details in the sections that follow.

2.3.1 Size and Shape of Polygons

The size and shape of polygons are frequently referred to as geographic units. A geographic unit refers to the spatial resolution or granularity of the data, such as enumerator areas, census tracts, wards, or municipalities. These units can also include polygons of equal size, such as hexagons or grid blocks. One important factor that is often overlooked is the availability of data for a usable and fit-for-purpose geographic unit. Data may not necessarily be available at a suitable resolution for effectively visualising spatial patterns using a choropleth map. Boscoe and Pickle (2003) defined the following ideal characteristics of geographic units, among other factors:

- “A high degree of resolution,
- homogeneity of population size and land area,
- minimum population thresholds and land area thresholds,
- compactness of shape,
- audience familiarity, and
- functional relevance”.

Population data derived from national census surveys are typically aggregated into specific geographic units, such as enumerator areas, wards, or municipal boundaries. It is generally recommended to visualise population data as ratios, such as population per square kilometre or rate, instead of actual values (Tyner, 2014). In a South African context, other data aggregation types include crime statistics by police precinct and information on voter registration and party support by voting district.

2.3.2 Number of Classes

Kraak and Ormeling (2020) mentioned that “it is a good cartographic practice to conveniently arrange the data before displaying them. This process is called classification”, which is indeed a form of data generalisation. Five to seven classes are generally accepted throughout the literature (Brewer, 2015; Kraak & Ormeling, 2011; O’Sullivan & Unwin, 2010). Furthermore, Peterson (2015) mentioned the five-shade rule: “the human eye can only distinguish between five shades of the same colour”. Additionally, fewer than five class intervals are not recommended, as the level of detail will be lost or severely limited. The opposite is true when more than nine class intervals are used. The display becomes too detailed, and “key differences between classes are difficult to see“ (De Smith et al., 2018).

2.3.3 Class Limits

When choosing a specific number of classes, the choropleth map designer should also consider the upper and lower limits of each class. According to Robinson (1995), “no aspect of choropleth mapping has received more space in the cartographic literature than methods to determine class limits”. Over the years, numerous data classification algorithms have been developed and integrated into GIS software applications. Since each data classification method has its own advantages and disadvantages, choosing an appropriate method should be considered carefully. One of the key aspects to consider is the overall distribution of the data. “Different frequency distributions suggest different class interval systems” (Evans, 1977). Some methods, such as standard deviation, are more effective when the data are evenly or normally distributed. In contrast, natural breaks are typically used when the data distribution is skewed.

Brewer (2006) stated that “there is no one correct way to class a dataset”. Monmonier (2018) pointed out in his book *How to Lie with Maps* that different sets of categories (or class limits) can lead to completely different interpretations. He further mentioned that mapping software can “encourage laziness” and inadvertently assist the “first-timers” who simply use the default classification methods to define class limits and the number of classes when creating a choropleth map without exploring the other options available in the software.

Numerous data classification methods, or methods for determining class limits, for choropleth maps are described in the literature, for example Slocum et al. (2014) highlighted six frequently used methods of data classification. These include equal intervals, quantiles, mean-standard deviation, maximum breaks, natural breaks, and the optimal method. Cromley (2019) observed that the most frequently used optimal classification is the Jenks optimal classification. Tyner (2014) mentioned arithmetic progression and geometric progression as additional classification methods.

De Smith et al. (2018) compiled a comprehensive list of frequently used data classification methods for choropleth maps, which are typically used to analyse univariate data that include a single variable, such as population density or mortality rate. Besides those mentioned above, De Smith et al. (2018) included unique values, exponential intervals, percentiles, and box plots. Other data classification methods that are worth mentioning include equal feature areas (Lloyd & Steinke, 1977), harmonic series (Kraak & Ormeling, 2011), and nested means (Dent et al., 2009; Kraak & Ormeling, 2011).

For a more customised approach, class breaks could be manually adjusted to accentuate a particular phenomenon relevant to the analysis. While other potential data classification methods are mentioned in the literature, they are not readily available in GIS applications.

Schiewe (2023) introduced a change preservation metric that evaluates the effectiveness of a data classification method. The metric could be used as a data classification method that "explicitly takes the preservation of changes into account" (Schiewe, 2023; 2024). Całka (2018) and Jiang (2012) described a head-tail break data classification, which is useful for data with a heavy-tailed distribution that are not normally distributed. Traun and Loidl (2012) proposed an autocorrelation-based regioclassification method, which groups polygons if their attributes are similar and spatially adjacent.

While there are various GIS software applications available for spatial analysis, mapping, and visualisation (both open source and licenced), ArcGIS Pro[2] and QGIS[3] are considered the most popular and frequently used applications globally, including in South Africa (GIS Geography, 2022). Bolstad (2012) mentioned that ArcGIS is by far the most popular GIS software. Furthermore, GIS Geography's (2022) ranking of the 30 best GIS software applications placed ArcGIS Pro in the top spot, followed by QGIS. Hence, for the purpose of this research, the effectiveness of data classification methods for choropleth maps available in either ArcGIS Pro or QGIS will be examined. The following section provides a detailed description of each data classification method.

2.3.3.1 Data Classification Methods Available in ArcGIS Pro or QGIS for Calculating Class Limits

Table *1* shows the nine data classification methods that are available in either ArcGIS Pro or QGIS. The geometric interval data classification method is only available in ArcGIS Pro, while QGIS offers specific methods, including logarithmic scale and pretty breaks. Defined (or fixed) interval, equal interval, manual interval, natural breaks (Jenks), quantiles, and standard deviation are available in both ArcGIS Pro and QGIS.

Table 1: Data classification methods available in ArcGIS Pro and QGIS

Data Classification Method	GIS Software	
	ArcGIS Pro	QGIS
Defined (or Fixed) Interval	☑	☑
Equal Interval	☑	☑
Geometric Interval	☑	
Logarithmic Scale		☑
Manual Interval	☑	☑
Natural Breaks (Jenks)	☑	☑
Pretty Breaks		☑
Quantiles	☑	☑
Standard Deviation	☑	☑

[2] https://www.esri.com/en-us/home
[3] https://www.qgis.org/en/site/

To illustrate the diverse visualisations that can be achieved using these data classification methods, a series of choropleth maps depicting population density by municipality in South Africa was created (see *Figure 1* to *Figure 7*). For this purpose, the Census 2011 municipal boundaries along with population density estimates were used. There are 234 municipalities across nine provinces in the country. The minimum population density is 0.33 people per square kilometre, while the maximum is 2 961 people per square kilometre, with a mean density of 96.5 (Statistics South Africa, 2012c).

2.3.3.2 Choropleth Maps Depicting Population Density by Municipality

With the exception of the manual and defined interval methods, which require the user to set custom or manual limits for each class, the choropleth maps that follow highlight population density based on seven data classification methods: equal interval, geometric interval, logarithmic scale, natural breaks (Jenks), pretty breaks, quantiles, and standard deviation. Where possible, five classes were chosen.

Defined (or Fixed) Interval and Manual Interval

Both these data classification methods require manual input. For a defined (or fixed) interval, the user specifies a fixed size for each class break. If the data ranges between a minimum value of 0 and a maximum of 100 and the user specifies a fixed size of 10, the GIS application will generate ten classes: 0–10;10–20; 20–30; 30–40; 40–50; 50–60; 60–70; 70–80; 80–90; and 90–100. The manual intervals data classification method requires the user to define both the custom class limits and the number of classes.

Equal Interval

The equal interval classification method creates equal-sized class breaks based on a specified number of classes, using the minimum and maximum data values. *Figure 1* shows five classes, each with a class limit of approximately 540. The number of features within classes is skewed, with 227 of the 234 features (municipalities) allocated to the first class (0.3–538). This method works best if the data are uniformly distributed, meaning the data distribution has no peak and is consistent (Kraak et al., 2021).

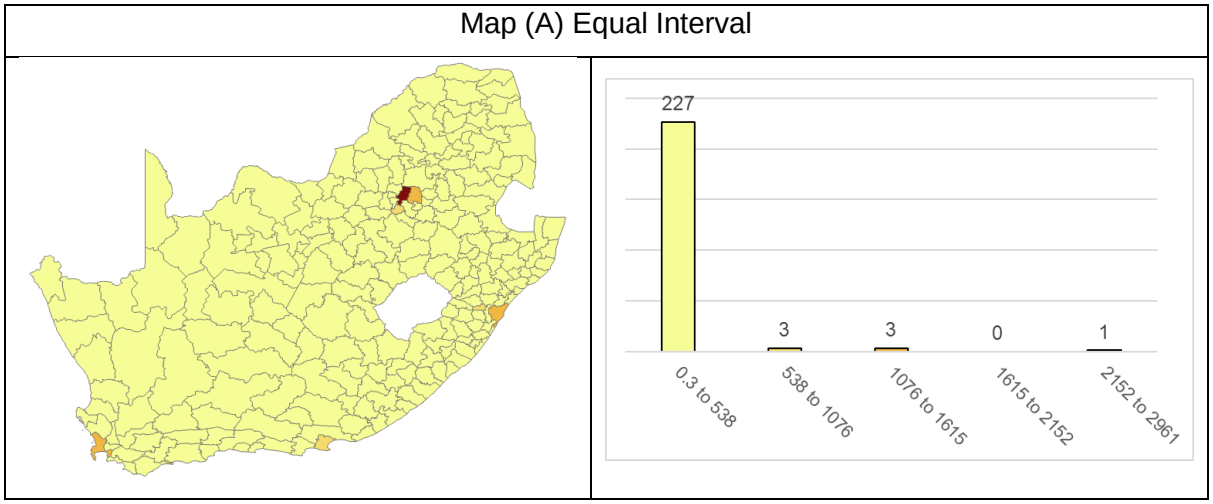


Figure 1: Equal interval data classification

Geometric Interval

The geometric interval data classification method is only available in ArcGIS Pro. In theory, the method attempts to group an equivalent number of features in each class while maintaining consistent class intervals. This is done by minimising the sum of squares for the data in each class. The population density map in *Figure 2* shows a significantly lower feature count for the fourth and fifth class breaks. Additionally, the class intervals are inconsistent. The first class ranges from 0.3 to 11, while the last two classes range from over 195 to 728 and from over 728 to 2 961, respectively. This method is usually preferred if the data are skewed (Evans, 1977).

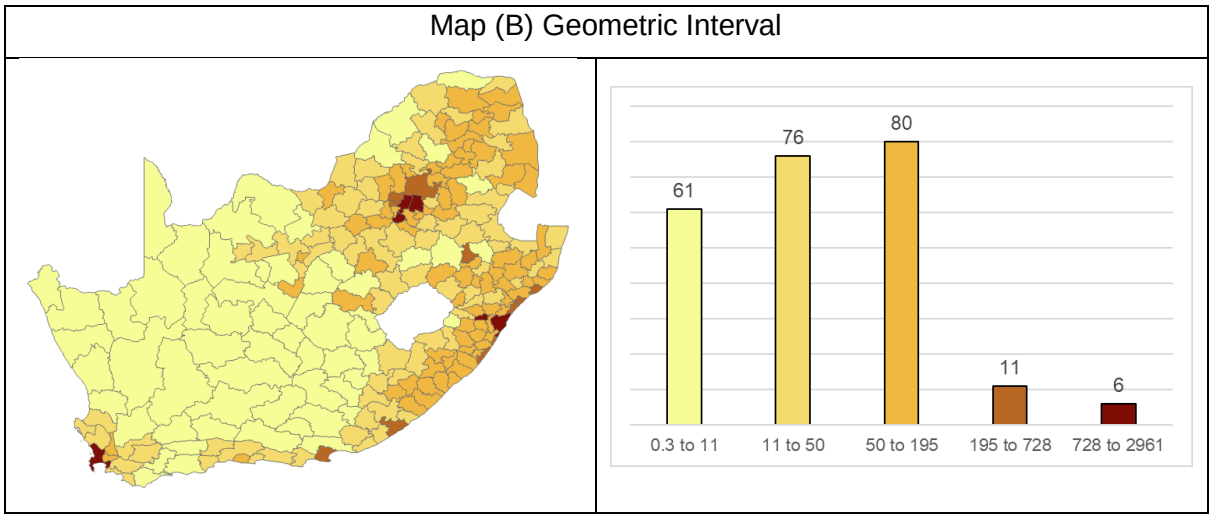


Figure 2: Geometric interval data classification

Logarithmic Scale

The logarithmic scale, which is only available only in QGIS, creates an exponential increase between each class break. This method is useful when data span a wide range of values. The population density map in *Figure 3* illustrates these increments based on five class breaks: 0.33–10^0 (0.33–1); 10^0–10^1 (1–10); 10^1–10^2 (10–100); 10^2–10^3 (100–1 000); and 10^3–10^4 (1 000–10 000). Since the legend is difficult to interpret, the map designer should consider changing the legend text manually.

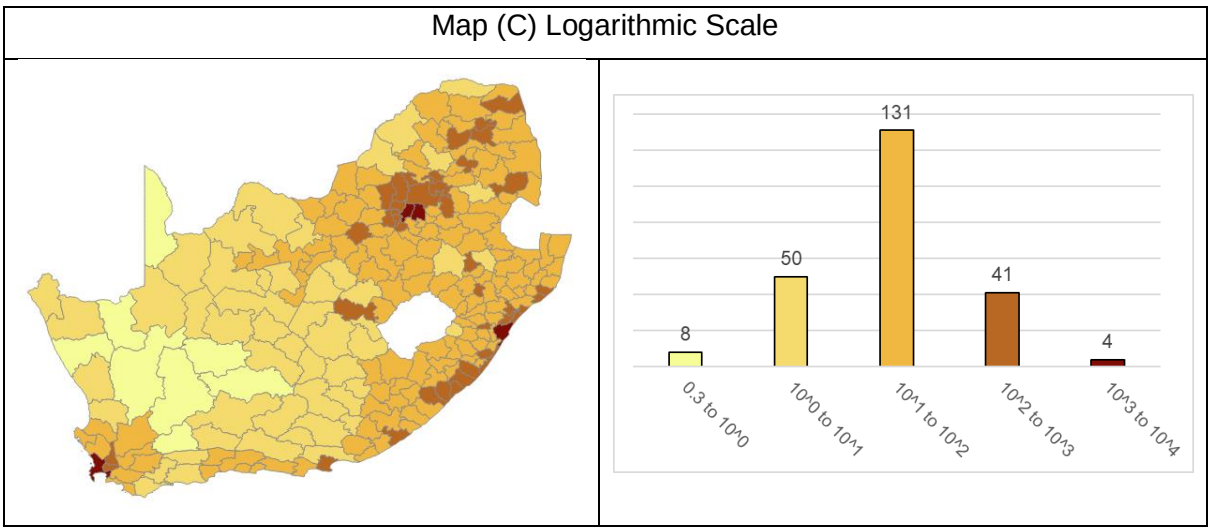


Figure 3: Logarithmic scale data classification

Natural Breaks (Jenks)

Jenks implemented a data classification algorithm in 1977, which was proposed by Fisher, to identify data belonging to the same class (as cited in Dent et al., 2009). The method, known as the optimal method, is now more commonly referred to as natural breaks (Jenks) in modern GIS software applications. The objective is to minimise the measure of data classification errors. Data with similar values are grouped into classes, with class breaks defined where there are significant differences between the data values. The number of features in each class is usually unevenly distributed, and the intervals between class breaks are not consistent. One advantage of this method is that it is effective for data that are not normally distributed (Całka, 2018). The population density map in *Figure 4* shows that most municipalities are allocated to the first two classes, with 149 and 69, respectively, while only a few polygons are allocated to the other classes.

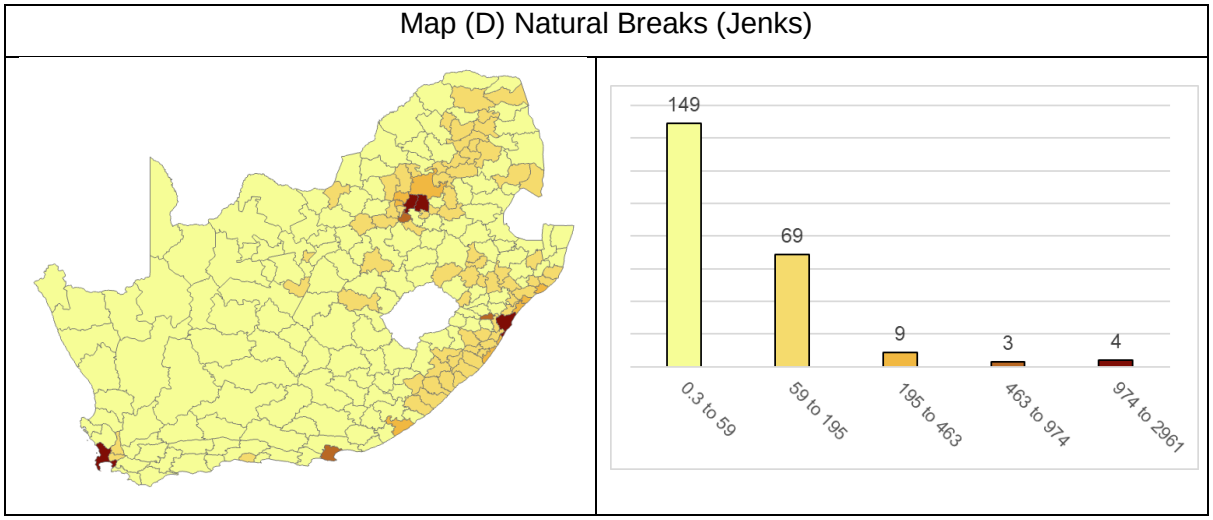


Figure 4: Natural breaks (Jenks) data classification

Pretty Breaks

The pretty breaks data classification method, which is only available in QGIS, is similar to the equal interval method, in which equal-sized class breaks are created based on the minimum and maximum data values. The pretty breaks data method, however, does attempt to round (or simplify) the upper- and lower-class breaks, making them easier to read. As an example, the first and second class break intervals in *Figure 5* were 0.3 to 538, and 538 to 1 076, respectively, compared to the pretty breaks data of 0.3 to 500, and greater than 500 to 1 000, respectively. Again, this method works best if the data are uniformly distributed, meaning the data distribution has no peak and is consistent (Kraak et al., 2021).

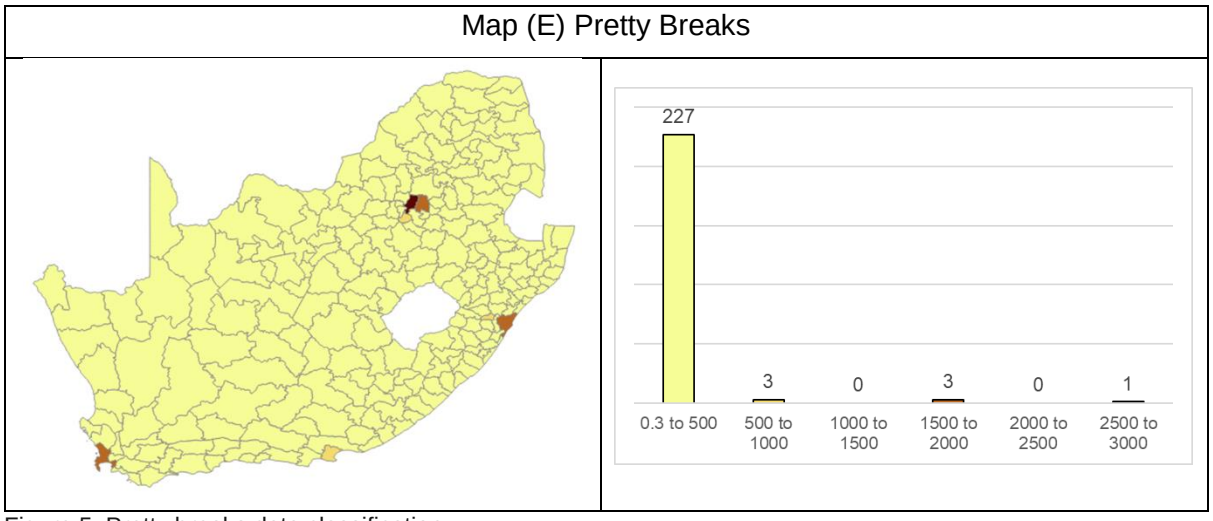


Figure 5: Pretty breaks data classification

Quantiles

Quantiles attempt to group an equal number of features in each class (De Smith et al., 2018). As a result, class intervals are usually not consistent. This method is effective if the data are not normally distributed. The population density map in *Figure 6* indicates that four of the five

classes comprise 47 municipalities each, with 46 municipalities in the last class. One advantage of this method is that no class interval will be empty, meaning features are distributed across all classes (Vasilca, 2019), compared to equal interval and pretty breaks, where some classes may be empty.

Map (F) Quantiles

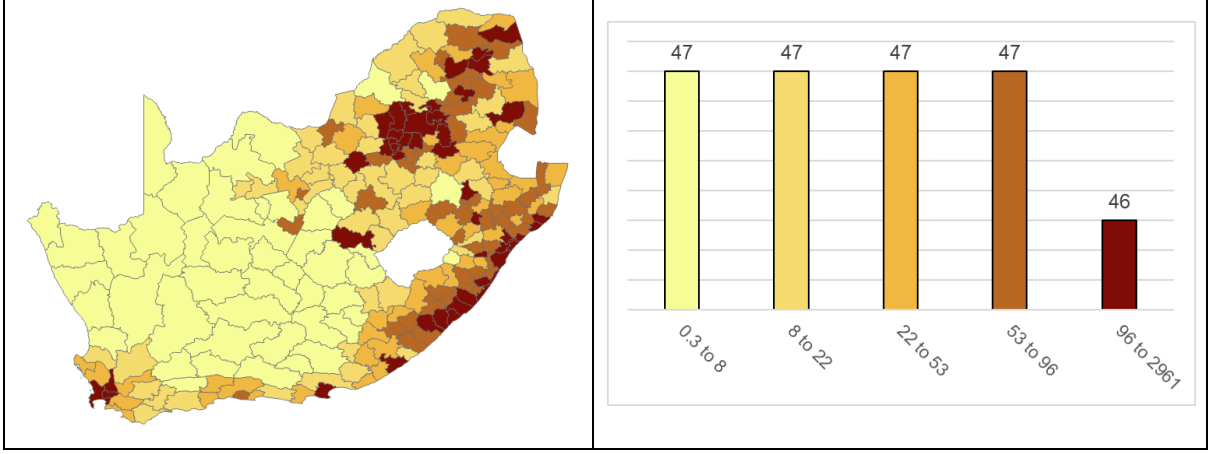


Figure 6: Quantiles data classification

Standard Deviation

Standard deviation measures the difference between each feature's data value and the overall mean. ArcGIS Pro creates class breaks with equal ranges. The population density map in *Figure 7* shows an interval size of one standard deviation, resulting in only three classes. The legend displays the difference between the means but not the actual class breaks (Całka, 2018). The number of classes is generated automatically based on the interval size. This method is usually recommended when the data are normally distributed without outliers (Slocum et al., 2014).

Map (G) Standard Deviation

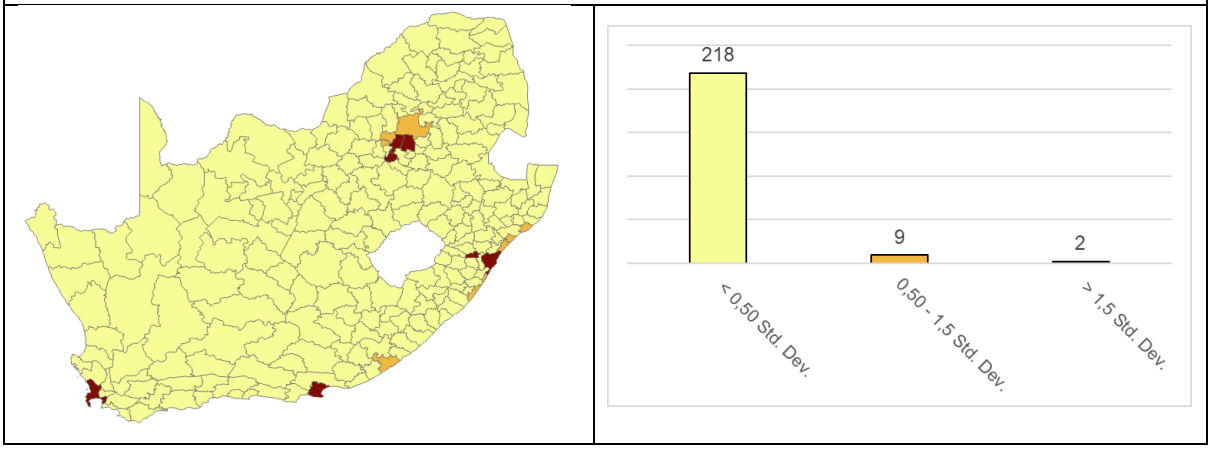


Figure 7: Standard deviation data classification

2.3.4 Problems with Simple Choropleth Maps

One disadvantage of choropleth maps when visualising statistical data is the tendency to "overemphasize large, yet often sparsely populated, administrative areas because of their strong visual weight" (Besançon et al., 2020). This is also true for population data depicting urban and rural areas, as rural areas encompass a larger geographic space than urban areas (Harris et al., 2017). Furthermore, since statistical data are usually captured on various administrative boundaries, such as enumerator areas or wards, which are demarcated artificially, "it is not possible to show variation within enumeration areas" (Tyner, 2014). Slocum et al. (2014) stated that choropleth maps are most effective and accurate when the size and shape of the polygons are fairly similar.

One of the major challenges associated with choropleth maps is selecting a suitable data classification method that displays and communicates the data distribution in a clear and effective way (Slocum et al., 2014). Using the wrong data classification method can distort spatial patterns, resulting in a misleading representation of population densities. This may highlight incorrect or inappropriate locations for the positioning of service centres, ultimately leading to poor service delivery and frustrated citizens. Another challenge with choropleth maps in South Africa is selecting a classification method that effectively displays unequal and dispersed population densities (demand). It should emphasise not only the city centres and their surroundings but also secondary or tertiary populated areas such as townships and informal settlements that are segregated from the city centre.

Kraak et al. (2021) pointed out that classification could potentially increase uncertainty, as the patterns derived from a classification are influenced by the positioning of class breaks. Additionally, Jenks (1963) commented that "a cartographer makes a series of judgments without really understanding what effect these judgments will have upon the reader's interpretation of the distribution". Choosing the wrong or inappropriate data classification method for visualising data with choropleth maps could potentially distort spatial patterns, causing misleading representations and a possible oversimplification of information (Brewer, 2006; Evans, 1977 Monmonier, 2018). Schiewe (2023) further noted that standard or traditional data classification methods are not ideal for visualising multi-temporal data, where temporal changes are either lost or incorrectly highlighted.

2.4 Measurement Techniques

As discussed in Section 2.3.3, a key consideration in choosing a data classification method for a specific data set is the data distribution. "Different frequency distributions suggest different class interval systems" (Evans, 1977). Some methods, such as standard deviation,

are more effective when the data are evenly or normally distributed, while natural breaks are often preferred for skewed data distributions.

Besides data distribution, the literature highlights two effective techniques for measuring or assessing data classification methods for choropleth maps. These include, firstly, a user study in the form of a survey or questionnaire and, secondly, mathematical equations that measure the error between class breaks. This is useful for understanding the extent of data generalisation within class breaks.

In user studies, questionnaires are frequently used to assess respondents' interpretation of choropleth maps that depict different data classification methods. The user study typically includes a series of map-specific questions. The correct and incorrect answers are analysed to determine which method(s) the target audience understands better, specifically the respondents participating in the user study.

The literature discusses various mathematical equations for determining the error between class breaks, with references dating back to the 1970s. These equations calculate an accuracy score or index that indicates the level of data generalisation within each class break. Jenks and Caspall (1971) highlighted the fact that since a choropleth map is a "generalization of reality based on an aerial (or location) distribution, it must include some degree of error".

The next section provides a detailed description of each technique.

2.4.1 User Studies

"User studies can objectively establish which method is most appropriate for a given situation" (Kosara et al., 2003). User studies are conducted using a survey or questionnaire. Survey questionnaires have been used for many years (Clifford et al., 2016), including in the field of geography, and have proven to be an important tool for evaluating people's perceptions or interpretations of a subject. These questionnaires are frequently referred to as explicit reports due to their explicitness, "People know they are providing information to a researcher when they are surveyed" (Montello & Sutton, 2012). Designing a questionnaire requires careful planning, and the questions should be structured to address a specific topic or research problem effectively. Clifford et al. (2016) further described the following three basic principles for designing questions:

- "Keeping the questions simple,
- Define and describe relevant terminology clearly and effectively, and lastly
- Try to use plain and simple wording; easily understood by the target audience".

As internet connectivity is becoming more available and affordable, online questionnaires are increasingly being considered as a great cost-effective approach compared to traditional methods such as face-to-face or telephonic interviews (Regmi et al., 2016). To design online questionnaires, Regmi et al. (2016) defined six key components. Firstly, a user-friendly design makes it easy for respondents to navigate the interface and respond to the various questions. Secondly, selecting the appropriate respondents to participate in the survey is a crucial aspect to consider. Not just regarding their required level of expertise on a subject but also their age, technical literacy, and experience with online applications. Thirdly, multiple responses from respondents should be avoided as much as possible by requesting that potential participants register for the survey. The forth component is data management. Data management is important. Secure data storage, along with the ability to export survey results, enables researchers to analyse data using various platforms such as MS Excel or a database management system. Next, ethical issues, which include informed consent, privacy and confidentiality, and the right to withdraw are noted in the questionnaire. The final component involves piloting. Once the questionnaire design is complete and available on an online platform, it is advisable to conduct a pilot study with a small group of targeted individuals. The pilot is required to test the functionality of the online questionnaire, identify potential issues, and evaluate the structure and format of the captured data.

The following section highlights various user studies described in the literature. The review begins with examples of how questionnaires were used to assess map reading skills. The Santa Barbara Sense of Direction Scale developed by Hegarty et al. (2002) has proven to be a useful tool for predicting spatial abilities in different environments. The scale consists of 27 self-evaluation questions, with responses ranging from 1 (strongly agree) to 7 (strongly disagree). During the discussion, Hegarty et al. (2002) noted that while evidence suggests that people often “overestimate their abilities”, their study’s observations reveal that people are “somewhat truthful and accurate in estimating their environmental spatial abilities”.

Lee and Bednarz (2012) developed a spatial thinking ability test. The test was designed to cover the following spatial thinking components: orientation and direction, map and graphic information, choosing locations on a map, slope profiles, spatial distribution of map features, visualising three-dimensional images, “overlaying and dissolving maps”, and understanding of point, line and polygon features. In their conclusion, Lee and Bednarz (2012) highlighted the fact that spatial thinking is a “collection of various skills”, and in order to test students’ knowledge of spatial thinking, they should “demonstrate what they have learned in different ways”.

In another example, Tomaszewski et al. (2015) presented a modified version of the spatial thinking ability test. They conducted a logistic regression analysis to assess the significance of the variables. The results of the study indicate that students from urban schools performed better than those from rural schools. Tomaszewski et al. (2015) also found statistically significant differences between the gender groups.

A map reading experiment regarding the interpretation of COVID-19 data compared choropleth map visualisations with graduated symbols, which were considered the most frequently used techniques for visualising COVID-19 data at the time. Results from an online questionnaire revealed the opposite, showing that choropleth maps were much easier to read and understand than graduated symbols (Sukraini et al., 2022).

Albert et al. (2016) tested the general map reading skills of 488 higher education students. The questions focused on orientation, distance, topographic elements, geographic names, map symbols, and hypsography. The results of their study revealed variations in map reading competency based on gender, nationality, and age group.

Rautenbach et al. (2014) conducted a map literacy test using both two-dimensional maps and three-dimensional models. The test was conducted using various methods, including a focus group and a questionnaire, and it examined aspects such as map orientation, direction, and distance. Twenty-one students participated in the experiments. The preliminary results from the study showed that participants performed equally well when exposed to two-dimensional maps and three-dimensional models. Subsequently, Rautenbach et al. (2017) developed and evaluated a task taxonomy for spatial planning through a map literacy experiment involving 49 map-literate participants using topographic maps. The task taxonomy included tasks ranging across six levels of increasing difficulty. Qualtrics was used to design the online questionnaire. Participation was voluntary, and there was no time limit for completing the questionnaire. The results were evaluated based on an accuracy score, which reflected the correct and incorrect responses of participants. The results indicate that the accuracy score of participants correlated with their self-perceived difficulty levels for each question. Rautenbach et al. (2017) also found that gender “had no effect on confidence and task completion time”. Lastly, the map literacy experiment demonstrated that there was no significant correlation between participants’ self-rated experience levels and their accuracy in performance.

Lloyd and Bunch (2008) evaluated map reading efficiency based on gender, memory and geographic information. Their experiment included a map reading task, in which respondents were asked to locate specific states on a map of the United States based on their names. The

slowest mean reaction time and the lowest mean accuracy were recorded for females (Lloyd & Bunch, 2008).

Questionnaires are also useful for assessing how respondents interpret maps depicting statistical data. Schiewe (2019) evaluated the visual perception of spatial patterns with choropleth maps based on three effects: “dark-is-more bias, area-size bias, and data-classification effect”’. The study involved designing an online questionnaire, which was used to examine responses from 260 participants. The results revealed that higher values are indeed associated with darker colours. The analysis also confirmed area-size bias, as 30–40% of participants overlooked smaller areas on the map.

Afifah (2019) tested data classification methods and colour symbol schemes to visualise population density in the Special Region of Yogyakarta. Although a proportion assessment test was used to evaluate the effectiveness of data classification methods, a questionnaire was also designed featuring choropleth maps to evaluate various colour symbol schemes. Based on the time and duration required to assess maps using eye-tracking technology, effective colour symbol schemes were identified. The results from his proportion test suggest that the arithmetic interval data classification method was the most effective. The eye-tracking analysis revealed that a diverging colour scheme was the most effective.

The use of questionnaires to evaluate the effectiveness of data classification methods for choropleth maps is not a new concept. Through questionnaires, respondents are typically asked to answer a series of map-specific questions based on choropleth maps that depict various data classification methods. Percentage accuracy scores, which represent the percentage of correct answers provided by respondents, are used to compare data classification methods and determine which methods are most suitable for a specific data set and use case.

Brewer and Pickle (2002) assessed various data classification methods for classifying epidemiological data using choropleth maps. The study comprises nine series of mortality maps for the United States, which are based on seven different classification methods. These include hybrid equal intervals, quantiles, box plots, standard deviation, natural breaks (Jenks), minimum boundary error, and shared area. A total of 56 respondents, all students at Pennsylvania State University, participated in the study. The map interpretation questions were designed in such a way that certain question types were more difficult than others. Brewer and Pickle (2002) conducted a logistic regression analysis to evaluate the accuracy of responses derived from the seven data classification methods. Findings from the study indicate that quantiles and the minimum boundary error classification methods produced the

most accurate results, which were best interpreted by participants, followed by natural breaks (Jenks) and a hybrid version of equal intervals.

2.4.2 Error Calculation Between Class Breaks

Calculating the error between class breaks to evaluate data classification methods for choropleth maps is well-documented in the scientific literature. Jenks and Caspall (1971) highlighted the fact that since a choropleth map is a generalisation of reality based on an aerial or locational distribution, it must include some degree of error. Jenks and Caspall (1971) further noted that factors such as the visual attractiveness of patterns and the creation of class breaks defined at “critical” values were considered effective by writers such as Jones (1930) and Schultz (1961). In general, Jenks and Caspall (1971) commented that writers or researchers recognise the need for an objective approach to define and evaluate accurate data classes.

For example, Figure *8* illustrates the level of generalisation, distortion, or error that occurs when data are grouped into different classes. *Figure 8(A)* shows an unclassed three-dimensional prism map. Each polygon is displayed at a certain unique height (as opposed to a unique colour for each polygon), hence no generalisation. Figure *8(B)* shows a two-dimensional choropleth map with generalisation based on a data classification method with five class intervals. Lastly, Figure *8(C)* illustrates the generalisation or distortion of Figure *8(B)*, this time using a three-dimensional prism map that shows areas (polygons) belonging to the same class. Thus, polygons in the same class are displayed at a uniform height, which makes it difficult for the user to discern variations within that class. This phenomenon is known as generalisation, distortion, or error.

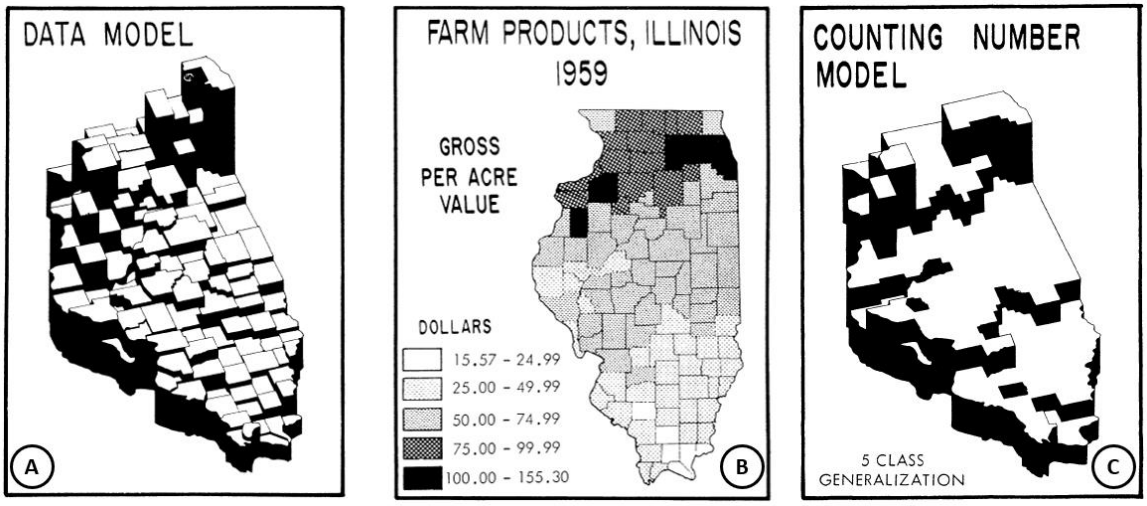


Figure 8: Generalisation based on data classification

Source: Jenks and Caspall (1971)

Robinson (1984) and Slocum et al. (2014) described two statistical criteria for calculating the error for class limits, namely goodness of variance fit and goodness of absolute deviation fit.

Goodness of Variance Fit

A measurement technique frequently described in the literature is the goodness of variance fit (GVF) measure (Armstrong et al., 2003; Chandra & Mistri, 2011; Declerq, 1995; Golian et al., 2010; Robinson, 1984; Slocum et al., 2014; Smith, 1986). The GVF score ranges from 0 to 1, with 1 representing the highest accuracy. This means there is no generalisation or distortion, ensuring that each polygon belongs to a distinct class. The GVF measures the "sum of squared deviations about the class mean" based on the following formula:

$$GVF = 1 - \frac{SDAM - SDCM}{SDAM}$$

Where $SDAM$ is the squared deviation from the array mean and $SDCM$ is the total sum of square deviation from the array mean (Robinson, 1984).

An existing R package called Jenks71 and classInt is available for automatically calculating GVF, the overview accuracy index, and tabular accuracy index scores simultaneously for any given data classification method, using a specified number of class intervals (the overview accuracy index and tabular accuracy index are described in more detail below). For example, see the Jenks71 R script (Figure *9*) depicting five class intervals based on the quantile data classification method.

```
install.packages("spData")
data(jenks71, package="spData")
install.packages("classInt")
library(classInt)


q5 <- classIntervals(jenks71$jenks71, n=5, style="quantile")
print(jenks.tests(q5, jenks71$area))
```

Figure 9: Jenks71 R package

Goodness of Absolute Deviation Fit

The goodness of absolute deviation fit (GADF) is also used to assess the accuracy of data classification methods. GADF is calculated with the following formula:

$$GADF = 1 - \frac{ADCM}{ADAM}$$

Where $ADCM$ is the sum of absolute deviations about class median (Slocum et al., 2014) (Vasilca, 2019) and $ADAM$ is the sum of absolute deviations about the median for the entire data set.

Jenks and Caspall (1971) developed three additional algorithms to assess the error of a classification method: the overview accuracy index, tabular accuracy index, and boundary accuracy index. These measurement techniques were designed to evaluate data classification methods based on these questions: (a) "Which map creates the most accurate overview?; (b) Which map provides the most accurate intensity values?; (c) Which map has boundaries that occur along major breaks in the statistical surface?" (Jenks & Caspall, 1971). The accuracy index ranges from 0 to 1, with 1 representing the highest level of accuracy.

Lastly, another measure for calculating classification error is based on the sum of absolute deviations from the class medians (ADCM), which is illustrated in Table *2*. This calculation was used to develop an optimal data classification method. The classification errors for the optimal method (on the right) and the quantiles method (on the left) are calculated separately. For example, Class 2 in the quantiles classification includes 14, 31, and 32. The median is 31, so the error calculation is as follows: $(14-31=17)+(31-31=0)+(32-31=1)=(17+0+1)=18$. Note the significantly smaller error for the optimal classification method when the data are grouped differently. The calculated error for the quantiles is significantly higher at 87, compared to the optimal method, which has an error of only 7.

Raw Data: 11, 12, 13, 14, 31, 32, 33, 99, 100					
Quantiles Classification			Optimal Classification		
Class	Value	Error	Class	Value	Error
1	11, 12, 13	2	1	11, 12, 13, 14	4
2	14, 31, 32	18	2	31, 32, 33	2
3	33, 99, 100	67	3	99, 100	1
	ADCM =	87		ADCM =	7

Table 2: Sum of absolute deviations from the class medians

Source: Slocum et al. (2014)

The following section highlights various examples from the literature where data classification methods were evaluated based on the error calculation between class breaks.

Declerq (1995) compared the GVF to two other measurement techniques: the goodness of deviation around the median fit and the GADF for data classification accuracy. Their intention was to determine the optimal number of class intervals using various data classification methods. These include: the Jenks optimal method, equal intervals, a suboptimal method "minimising image fragmentation", and another suboptimal method "minimising class break

complexity". During the study, a GVF accuracy of 0.95 or higher was considered "satisfactory". Results from this study suggest that seven or eight classes are required for accurate choropleth maps.

Chandra and Mistri (2011) used the GVF measurement technique to test for the most suitable classification method when the data are normally distributed. For this, they used the population density data of the Bankura District. Five data classification methods were tested: equal range (or equal interval), nested mean, parameters of normal distribution (or standard deviation), quantiles, and areal equal steps. The equal range classification achieved the highest accuracy of 82.96%. This was followed by standard deviation, quantiles, and nested means. Areal equal steps was ranked the lowest.

Additionally, Vasilca (2019) used the GADF to assess the effectiveness of the data classification methods available in ArcGIS Pro. For this purpose, thematic maps were created to illustrate the correlations between emergency calls and the populations of various regions in Romania. The GADF measure yielded the highest value for natural breaks at 0.82, followed by geometric interval, which scored 0.76. The lowest accuracy score was quantiles at 0.49.

Smith (1986) compared five traditional data classification methods – quartile, equal interval, standard deviation, natural breaks, and an optimisation method – based on the GVF measurement. It wat determined that only the optimisation method, nowadays commonly referred to as natural breaks (Jenks), produced accurate results. For this study, a sample of 117 data sets was derived from the 1977 County and City Data Book, which includes data such as birth rate, population density, and income.

2.5 Population Distribution in South Africa

South Africa has an uneven population distribution, which leads to unequal access to service facilities. Weir-Smith and Dlamini (2024) further mentioned that "unequal spatial concentration is at the heart of economic imbalance in South Africa" and that post-apartheid policies did not resolve economic imbalances caused by the apartheid era. A report by the World Bank (2018) on poverty and inequality describes South Africa as one of the most unequal countries in the world, stating that "inequality has increased since the end of apartheid in 1994".

People in South Africa are highly segregated, and this segregation is inherently geographical (Brown & Chung, 2006). The Group Areas Act 41 of 1950 (apartheid) "prohibited the multiracial use or occupation of urban land" (Strauss, 2019). Segregated zones were established in urban areas, allowing only certain race groups to live and work there. This resulted in the establishment of densely populated townships and informal settlements outside city centres

and suburban living spaces. Mostly poor people reside in these areas, and access to basic services and public service centres is limited and insufficient. Around 67% of the country's population live in urban spaces, with an estimated urban growth rate of 1.97% (UN-Habitat, n.d.), affecting service delivery and resulting in an ever-increasing unemployment rate and rising crime levels. Moeti et al. (2023) analysed quality of life survey data from the Gauteng City-Region Observatory and found that residents of informal settlements have significantly poorer access to healthcare facilities compared to those living in formal housing structures.

2.5.1 Population Data

Statistics South Africa[4] is the custodian of demographic data in South Africa. The last national census survey was conducted in 2022. Unfortunately, demographic data has not yet been released at a more granular level than local or metropolitan municipalities. Hence, for this research, demographic (or population) data from the previous national census conducted in 2011 were used. The following section first describes national and provincial population dynamics based on Census 2022, followed by a more granular description of population distribution based on the Census 2011 data (Statistics South Africa, 2012c).

The country covers a land area of approximately 1.2 million square kilometres. The total population is estimated to have reached 62 million in 2022, an increase from 51.7 million in 2011. Of the nine provinces, most people reside in Gauteng (15 million) and KwaZulu-Natal (12.4 million). The dominant population groups are Africans (81.4%) and coloured people (8.2%). The white population is estimated to be 7.3%. The total number of households is estimated at 17.8 million. Access to basic services, such as water, sanitation, and electricity, was also documented. It was estimated that around 82% of households has access to piped water, which includes access inside a dwelling or in the yard. Most households (70%) have flush toilets, followed by access to pit toilets without ventilation (12.5%). It was also reported that 94.7% of households use electricity as their main source of energy (Statistics South Africa, 2023a).

For Census 2011, the country was demarcated into approximately 103 000 enumerator areas "based on specifications of administrative boundaries, size, and population density" (Statistics South Africa, 2012a), which were used to capture all the demographic data. These enumerator areas were subsequently classified into ten enumerator area types "according to a set criteria profiling land use and human settlement within the area" (Statistics South Africa, 2012b). Of the 51.8 million people in South Africa, more than half (55.8%) live in formal residential areas,

[4] https://www.statssa.gov.za/

followed by traditional residential areas and informal residential areas, with 31.3% and 5.8%, respectively. See Table *3*.

Table 3: Population distribution by enumerator area type

Enumerator Area Type	Population	Percentage
Formal residential	28 885 090	55.8%
Informal residential	2 991 477	5.8%
Traditional residential	16 213 521	31.3%
Farms	2 078 722	4.0%
Parks and recreation	35 999	0.1%
Collective living quarters	609 907	1.2%
Industrial	157 978	0.3%
Small holdings	451 811	0.9%
Vacant	59 588	0.1%
Commercial	286 478	0.6%
Total	**51 770 571**	**100.0%**

Source: Compiled from Statistics South Africa; Interactive data in SuperCROSS (2012)

Census data are accessible to the public through an online portal[5] and a software application called Census 2011 Community Profiles in SuperCROSS. SuperCROSS allows users to download a range of demographic variables, including age, gender, population group, income, and employment data, for various geographic units or scales. These include a small area layer, sub-place, ward, main place, municipality, district, and province.

The small area layer is the smallest geographic unit for which demographic data are available. The data set is an aggregation of enumerator areas consisting of approximately 84 000 polygons. Sub-places represent “a suburb, section or zone of a township, smallholdings, village, sub-village, ward or informal settlement” (Statistics South Africa, 2012b). Main places include cities, towns, townships, tribal authorities, and administrative areas.

Based on the 2011 municipal demarcation, South Africa has 234 municipalities, including 226 local municipalities and eight metropolitan municipalities. Some municipalities are characterised by a diverse and distinctive geographic spread of urban, farm, and traditional spaces. The population within municipalities is unevenly distributed. In addition to densely populated central business districts, most municipalities also feature scattered pockets of populated townships or informal settlements located in peri-urban areas or outside the city centre. Henderson (2006) noted that a graphic visualisation of the data is “essential to assess the data distribution”. The histogram (Figure *10*), which contains 20 bins calculated in ArcGIS

[5] https://superweb.statssa.gov.za/webapi/jsf/login.xhtml

Pro and illustrates population density (POPKM2), highlights the fact that at the municipal level, the population is not normally distributed but rather skewed at 6.7 (kurtosis = 54.6).

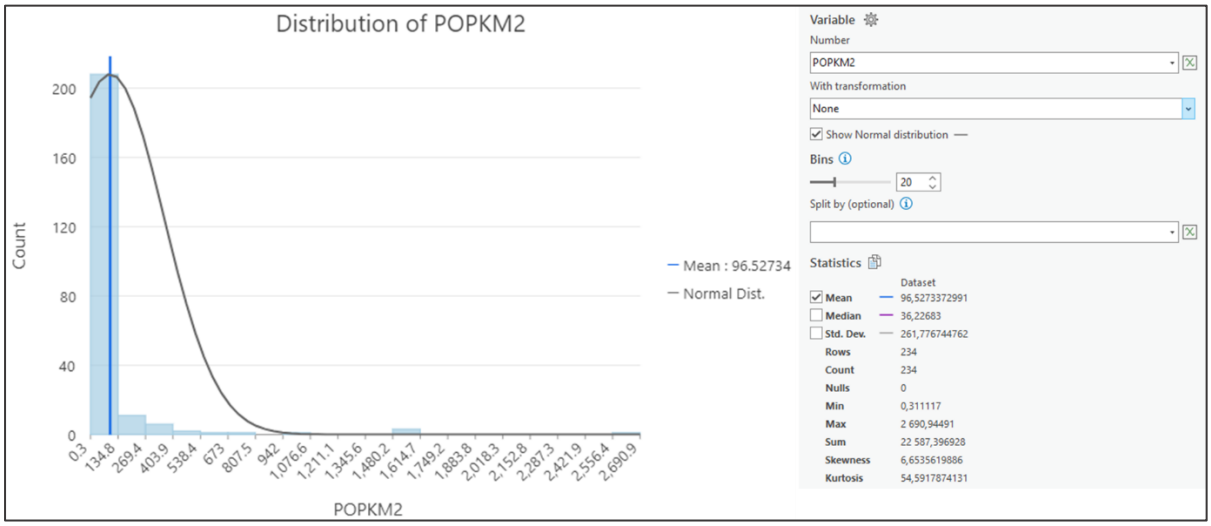


Figure 10: Population distribution by municipality

Due to its history, the geographic distribution of the South African population presents unique challenges for visualising spatial patterns on a map for effective decision-making. To effectively connect the population with public services, decision makers must understand where the population demand is in relation to public service facilities (Snyman & Coetzee, 2024). To do this, supply (service centres or facilities) and demand (population) maps are needed to visualise potential gaps and shortfalls. Choropleth maps are commonly used and effective for visualising population distribution (see Section 2.5.2 for a discussion on the use of choropleth maps in South Africa).

The challenge with data classification methods for choropleth maps in South Africa lies in selecting a classification method that effectively displays the unequal and dispersed population densities (demand). It should emphasise not only the city centres and their surroundings but also secondary or tertiary populated areas, such as townships and informal settlements, that are segregated from the city centre.

2.5.2 Choropleth Maps in South Africa

Choropleth maps are frequently used by the Electoral Commission of South Africa (IEC) to visualise voting and voter registration patterns. The IEC's (2019) Atlas of Results[6] enables users to analyse political party support with pregenerated choropleth maps (Figure *11*) for different geographic units such as voting districts, wards or municipalities. The Atlas is an online portal that showcases election results from 1999 to 2019. Voter registration choropleth

[6] https://atlas.elections.org.za/npeatlas/#

maps at the voting district level (highlighting high- and low-density patterns) enable decision makers to optimise voting district boundaries and voting station locations in preparation for election day (IEC, 2019).

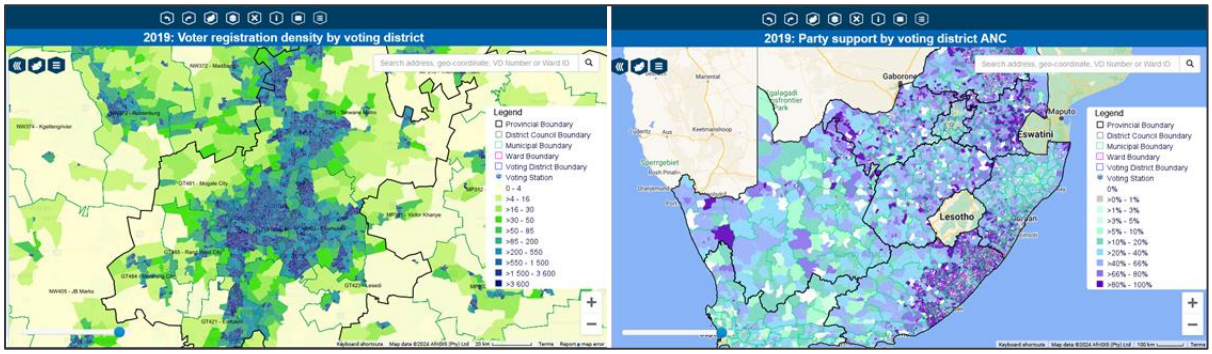


Figure 11: Choropleth maps showing voter registration and party support

Source: IEC (2019)

Statistics South Africa (2023b) uses choropleth maps to display population distribution and growth patterns across the country.[7] Other examples include (a) a study by Public Health that used choropleth maps to analyse cervical cancer screenings (Makura et al., 2016); and (b) an analysis of population distribution in relation to economic activity across the Gauteng city region (Mosiane & Murray, 2021). In a study conducted by the University of Pretoria, choropleth maps were used to show both isochrones – indicating travel time to catheterisation laboratories (cath lab) facilities – and population density by ward. The aim of the study was to identify densely populated areas across South Africa where access to cath lab facilities exceeds the specified travel-time thresholds (Coetzee et al., 2021).

A Scopus search for articles and abstracts that included the key phrases ‘choropleth map’ and ‘South Africa’ yielded four more relevant publications. Bekker (2023) analysed the occurrence of protest incidents using both dot distribution and choropleth maps and suggested that protests per capita are best represented with choropleth maps. Friesen et al. (2018) described a proof of concept web application that displays community-oriented primary care data using both choropleth maps and proportional circle maps. Friesen et al. (2018) concluded that for choropleth maps, “an automated determination of the optimal number of classes” based on a specific variable would be desirable. In health science, Khumalo et al. (2022) utilised the capabilities of choropleth maps to analyse the distribution of community health workers in relation to HIV prevalence in the KwaZulu-Natal province. A final example of the use of choropleth maps is a study by Motlana et al. (2021) that visualised the spatial distribution of cancer cases across three public hospitals in KwaZulu-Natal from 2015 to 2017.

[7] https://census.statssa.gov.za/#/

2.6 Geographic Accessibility and Population Demand

For this research, visualising population demand for geographic accessibility analysis was selected as a use case to assess the effectiveness of data classification methods for choropleth maps. This section provides a brief overview of the concept of geographic accessibility.

“Accessibility is the most widely used metric in measuring the value of a location in public service delivery” (Church & Murray, 2009). In connecting the population to public services, decision makers must understand where the population demand is in relation to public service facilities (Snyman & Coetzee, 2024). To achieve this, it is essential to create maps that visualise both supply (service centres or facilities) and demand (population) in order to identify gaps and shortfalls. “Improving service delivery to all people in South Africa is a key priority of government“ (Green, 2012). Population demand refers to the “number of people who may need the services” (Ma et al., 2018).

Effectively communicating or presenting the results of a geographic accessibility analysis is as important as the results themselves. Although visualising population demand is not the only metric used to analyse geographic accessibility, it is one of the prominent outputs of such a study, which is usually presented to a target audience. Generally, choropleth maps are effective for visualising statistical data such as population densities (Tyner, 2014). Other critical factors to consider are a road network and the location of service centres or facilities (DPSA, 2021). If service centres are situated too far away from the population demand, reaching these centres are “costly and time-consuming” (Church & Murray, 2009) which could result in poor service delivery and frustrated citizens. Rodrique et al. (2009) defined spatial or geographic access as “the measure of the capacity of the locations to be reached by, or to reach, different locations” based on both time and distance (Ashiagbor et al., 2020) between people and service centres.

Snyman and Coetzee (2024) measured geographic access to service facilities in the rural areas of the Eastern Cape, South Africa, where roads and footpaths are often not mapped. For the study, they compared travel distances based on a straight line, a road network, and an augmented travel network that includes a triangular irregular network to serve as a proxy for unmapped roads and footpaths. The results from their study suggest that an augmented travel network is a suitable alternative for measuring geographic access to service centres, especially in data-poor areas where rural roads and footpaths are not mapped. Choropleth maps were used throughout the study to visualise travel distances to service centres.

Earlier work by the Department of Public Service and Administration (DPSA, 2012) includes the development of a practical step-by-step guideline for improving geographic access to government service points. While the guideline was comprehensive regarding developing access standards, collecting and using quality geospatial data, conducting accessibility studies, and developing an implementation strategy, little attention was given to the geographic visualisation of results.

Questions relating to geographic accessibility differentiate between densely and sparsely populated areas, while also identifying locations that are either over- or underserved for the optimal provision of service centres. Questions were structured based on the following facility location models as identified in the DPSA's (2012) geographic access guideline:

- Expansion model
- Reduction model
- Relocation model

The expansion model encompasses two approaches: greenfield and brownfield. For the greenfield approach, optimal locations for service centres are determined regardless of the current footprint of service centres. The brownfield approach considers the current footprint when determining optimal locations (DPSA, 2012). If service centres are not optimally located (close to the people) due to possible settlement growth or movement patterns, they could either be relocated (relocation model) or closed down (reduction model).

Examples of geographic accessibility studies that mapped population demand using choropleth maps include published work by Ab Hamid et al. (2023), Chen et al. (2023), and DPSA (2013).

The literature review explored the use of choropleth maps and emphasised the importance of choosing an appropriate data classification method for a specific dataset. Also, previous research identified two effective techniques for measuring or assessing data classification methods for choropleth maps. These include, firstly, a user study in the form of a survey or questionnaire and, secondly, mathematical equations that measure the error between class breaks. To assess different data classification methods based on these two techniques, maps are needed.

The next chapter (Chapter 3) focuses on the research design with a technical description of the processes that were followed to create choropleth maps for evaluation. These include the selection of study areas, geographic units, number of classes, colour schemes, and data classification methods.

3. RESEARCH DESIGN

Chapter 3 describes the choropleth map design process. The first section provides an overview of the research approach and ethical considerations, followed by a technical description (Sections 3.3 to 3.6) of how the data and choropleth maps were prepared. Figure *12* illustrates the five steps of the map design process.

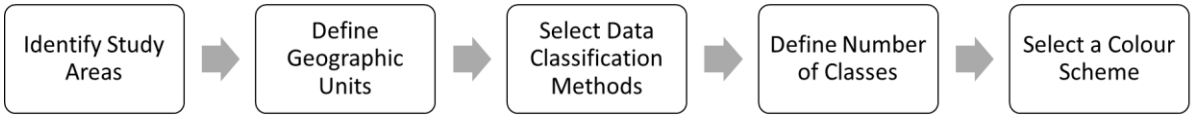


Figure 12: Choropleth map design process flow

3.1 Approach

Based on the recommendations of the empirical literature review, this research adopted two quantitative research methods (Research Objective 2) to assess the effectiveness of data classification methods for choropleth maps. These include elements of both explicit reports (experimental design) and statistical data analysis (nonexperimental design). The explicit reports were based on a user study (presented in Chapter 4), which consisted of an open-ended online questionnaire that assessed participants' interpretation of choropleth maps depicting different data classification methods for geographic accessibility analysis (Research Objective 3). For the user study, convenience sampling was used which included voluntary participation of students from the University of Pretoria.

Statistical data analysis involved measurement techniques to assess and evaluate the effectiveness of data classification methods for visualising population demand in South Africa. For this, the error between class breaks was measured using a recommended mathematical equation (Research Objective 4). The statistical data analysis process and results are presented in Chapter 5.

For this research, priority was given to the user study. The error between class breaks measurement was used to compare findings from the user study against a well-known mathematical equation. The combination of these two methods provided a more comprehensive view of the use of data classification methods for choropleth maps.

3.2 Ethical Considerations

The research aim and objectives were approved by the Ethics Committee, Faculty of Natural and Agricultural Sciences, University of Pretoria (NAS021/2023). Please see Appendix A for

the ethics approval document. Participants who accepted the invitation to participate in the user study were informed about the research's purpose and objectives and they were required to give their consent before completing the questionnaire.

The following geographic data sets were used with permission from the data custodians:

- Census 2011 population data, aggregated by small area layer and sub-place from Statistics South Africa. https://www.statssa.gov.za/?page_id=425
- Eskom Holdings SOC Ltd Spot Building Count; a geo-referenced dwelling or building frame point data set. https://www.eskom.co.za/paia-popia/

3.3 Study Areas

The aim was to identify suitable study areas that are representative of the unique spatial distribution of people in South Africa, as described in Section 2.5. Choropleth maps showing various study areas at different geographic units, or geographic scales as defined in Section 3.4, were designed to eliminate a possible learning effect. Specifically, to avoid or limit the likelihood that participants' responses to a question based on a choropleth map would be influenced or affected by their answers to previous questions in the questionnaire.

The population in South Africa is not evenly distributed. In addition to the densely populated central business districts, which are characterised by high-rise residential buildings, and suburban residential zones in metropolitan or local municipalities, there are also scattered pockets of densely populated townships and informal settlements located in peri-urban areas or on the outskirts of city centres. The challenge with data classification methods for choropleth maps lies in effectively displaying the varying densities. Choropleth maps should highlight not only primary locations such as city centres and their surrounding suburban areas but also populated secondary and tertiary locations such as townships and informal settlements.

Statistics South Africa is the custodian of demographic data in South Africa. Although a national census survey was conducted in 2022, demographic data has not yet been released on a more granular level than local and metropolitan municipalities. Hence, for this research, demographic (or population) data from Census 2011 were used (Statistics South Africa, 2012c), as the data are available at more granular levels, referred to as small area layers and sub-places. Small area layer polygons are the smallest geographical units with demographic data, whereas sub-places are aggregated polygons derived from small area layers that represent suburbs or villages (Statistics South Africa, 2012b).

Based on the Census 2011 results, the country has a population of 51.7 million covering a total land area of 1.2 million square kilometres. South Africa has 234 municipalities, which

include 226 local municipalities and eight metropolitan municipalities (Statistics South Africa, 2012c). Furthermore, areas are categorised into ten enumerator area types: formal residential, informal residential, traditional residential, farms, parks and recreation, collective living quarters, industrial, small holdings, vacant, and commercial, highlighting the diverse landscape of the country (SuperCROSS, 2012).

To effectively evaluate and assess the suitability of data classification methods for choropleth maps that depict population demand in a diverse geographic setting, representative municipalities were selected as study areas where populations are distributed across all ten enumerator area types. Table *4* shows the top ten municipalities based on total population distributed across all enumerator area types.

Table 4: Top ten municipalities based on total population distributed across all enumerator area types

Local and Metropolitan Municipality	Total Population	Rank
eThekwini Metropolitan Municipality	3 442 360	1
City of Tshwane Metropolitan Municipality	2 921 488	2
Buffalo City Metropolitan Municipality	755 094	3
Mangaung Metropolitan Municipality	747 432	4
Polokwane Local Municipality	629 000	5
The Msunduzi Local Municipality	618 536	6
Thulamela Local Municipality	618 462	7
Mbombela Local Municipality	588 794	8
Rustenburg Local Municipality	549 574	9
Bushbuckridge Local Municipality	541 249	10

Source: Compiled from Statistics South Africa; Interactive data in SuperCROSS (2012)

The aim was to identify four suitable study areas that are representative of the unique spatial distribution of people in South Africa (see Figure *13*), namely the City of Tshwane Metropolitan Municipality, Buffalo City Metropolitan Municipality, Mangaung Metropolitan Municipality, and Polokwane Local Municipality.

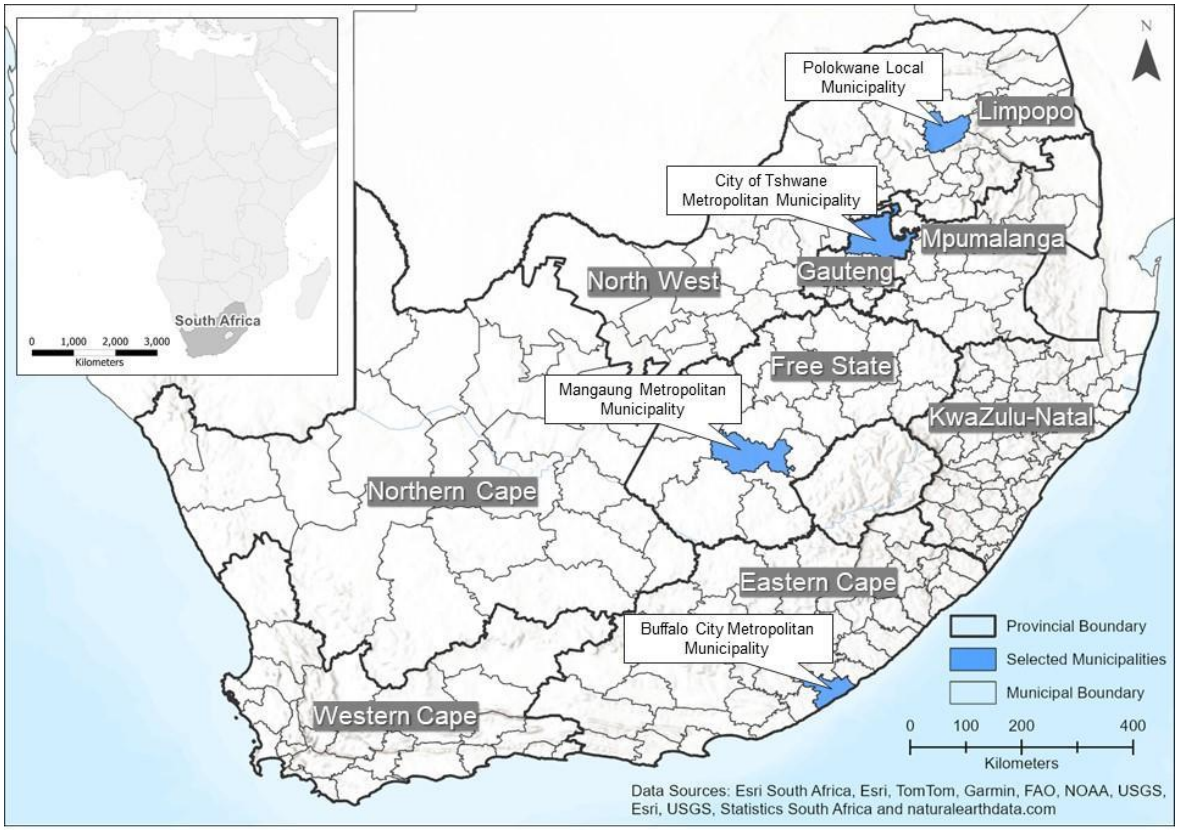


Figure 13: Municipalities in South Africa showing the four selected study areas

Source: Map created in ArcGIS Pro using built-in data from Esri (2024). These include Esri South Africa, TomTom, Garmin, FAO, NOAA and USGS. Other sources include Statistics South Africa (2011) and naturalearthdata.com (2023)

Table *5* depicts the percentage distribution of the population per enumerator area type for each of the four selected study areas. Most people reside in formal residential areas, except for the Polokwane Local Municipality, where 54.8% of the population lives in traditional residential areas, followed by 37.4% in formal residential areas.

Table 5: Percentage population distribution per enumerator area type for each of the four selected study areas

Enumerator Area Type	Buffalo City Metropolitan Municipality	City of Tshwane Metropolitan Municipality	Mangaung Metropolitan Municipality	Polokwane Local Municipality
Formal residential	67.4%	78.1%	83.4%	37.4%
Informal residential	12.6%	10.3%	8.4%	2.3%
Traditional residential	16.4%	5.2%	1.6%	54.8%
Farms	1.2%	0.6%	1.6%	0.7%
Parks and recreation	0.0%	0.0%	0.0%	0.1%
Collective living quarters	1.3%	2.1%	2.3%	1.2%
Industrial	0.3%	0.3%	0.7%	0.1%
Small holdings	0.2%	2.3%	1.6%	2.8%
Vacant	0.1%	0.0%	0.1%	0.2%
Commercial	0.6%	1.0%	0.3%	0.5%

Source: Compiled from Statistics South Africa; Interactive data in SuperCROSS (2012)

Figure *14* to Figure *17* show detailed orientation maps of each study area, depicting municipal boundaries as demarcated in 2011. The maps were designed in Maptitude GIS, highlighting primary cities and towns, as well as general topographic features such as road and railway infrastructure, along with natural features including dams, river networks, and nature reserves. Data sources include OpenStreetMap, NAVTEQ, and Statistics South Africa.

City of Tshwane Metropolitan Municipality

The metropolitan area is situated in Gauteng province and covers more than 6 300 square kilometres, making it the largest of the four selected study areas. With a total population estimate of just under 3 million, based on Census 2011, the population density is approximately 460 people per square kilometre (Statistics South Africa, 2012d). Major cities include Pretoria, the country's capital, along with various densely populated localities such as Mamelodi, Atteridgeville, Centurion, Akasia, Ga-Rankuwa and Mabopane, which are in close proximity to the capital. Secondary localities across the metropolitan area include Bronkhorstspruit, Cullinan, Rayton, Ekangala, and Temba.

"Tshwane stretches almost 121 km from east to west and 108 km from north to south, making it the third-largest city in the world after New York and Tokyo/Yokohama" (City of Tshwane, n.d.).

"The municipality's main economic sectors are community services and government, followed by finance and manufacturing" (Statistics South Africa, 2012d).

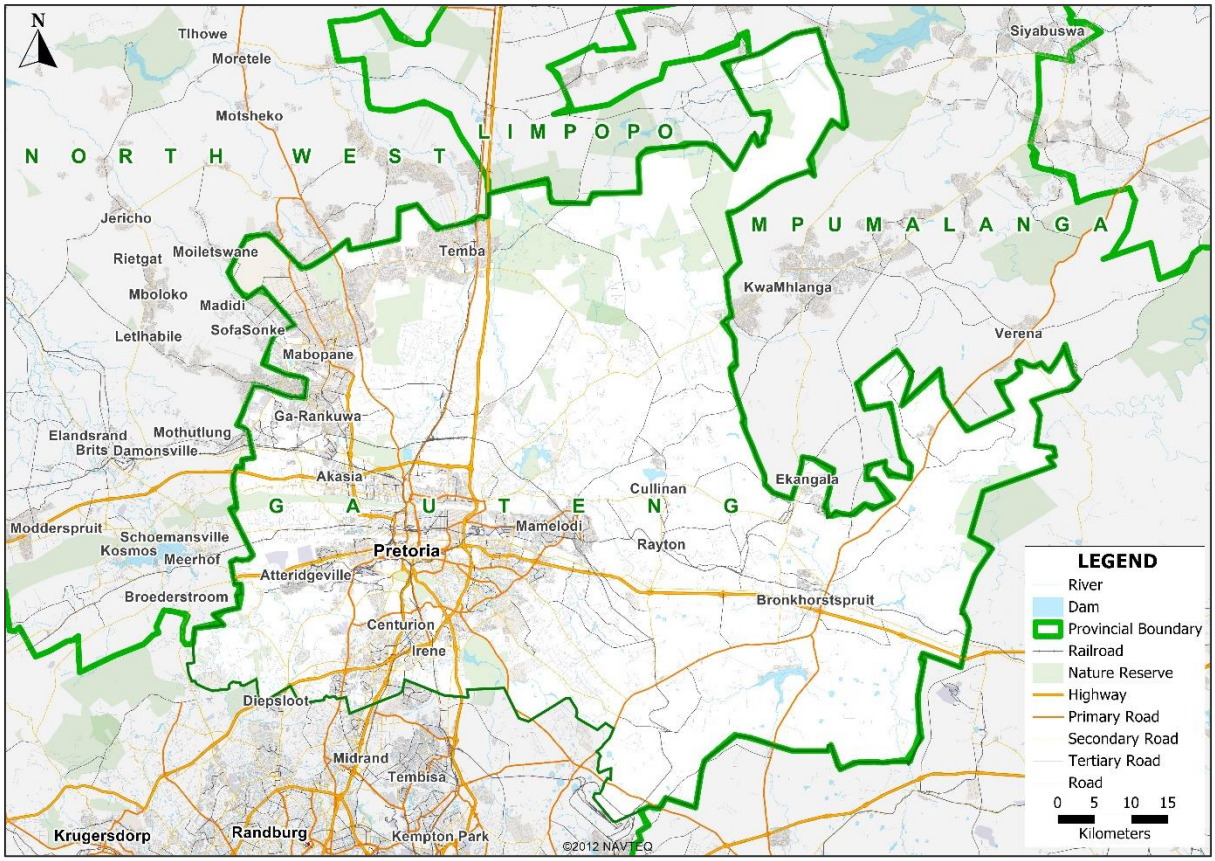


Figure 14: City of Tshwane Metropolitan Municipality

Source: Map created using data from NAVTEQ (2012), OpenStreetMap (2024), and Statistics South Africa (2011)

Buffalo City Metropolitan Municipality

The Buffalo City Metropolitan Municipality is situated in the Eastern Cape province and has a coastline of 68 kilometres along the Indian Ocean. The metropolitan is approximately 2 500 square kilometres with a population density of just under 300 people per square kilometre. East London is the largest city in the metro, situated on the coast and “boasting air, road, rail, and sea logistics” (Buffalo City Metropolitan Municipality, n.d.). Densely populated localities across Buffalo City include Mdantsane, Bhisho, and Qonce.

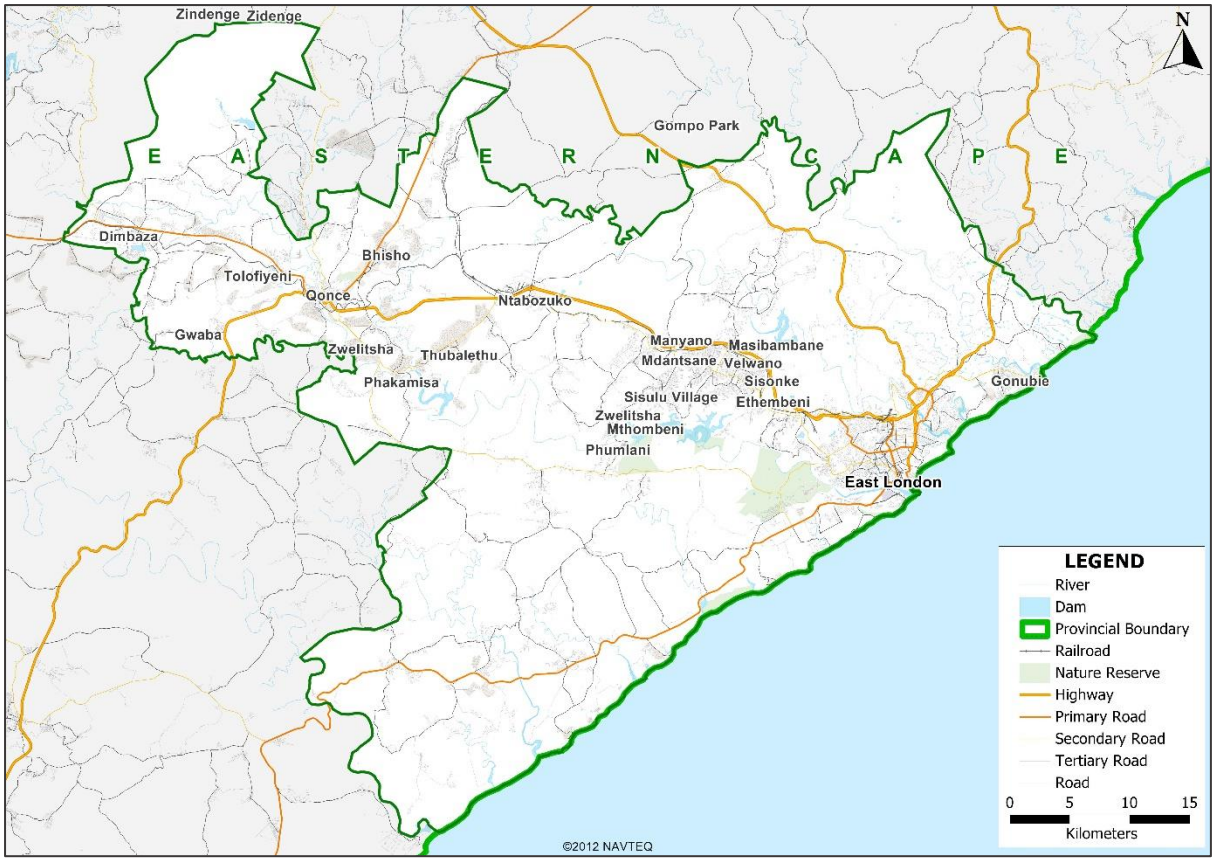


Figure 15: Buffalo City Metropolitan Municipality

Source: Map created using data from NAVTEQ (2012), OpenStreetMap (2024), and Statistics South Africa (2011)

Mangaung Metropolitan Municipality

The Mangaung Metropolitan Municipality is located in the Free State province. Mangaung translates to the place of the cheetah. “The economy is strongly driven by the government sector”, followed by the finance sector (Statistics South Africa, 2012e). The largest city is Bloemfontein, which is also the largest city in the province, with densely populated secondary localities including Botshabelo, Mangaung, and Thaba Nchu. The physical size of the metropolitan area is just below 6 300 square kilometres with around 750 000 people (120 people per square kilometre).

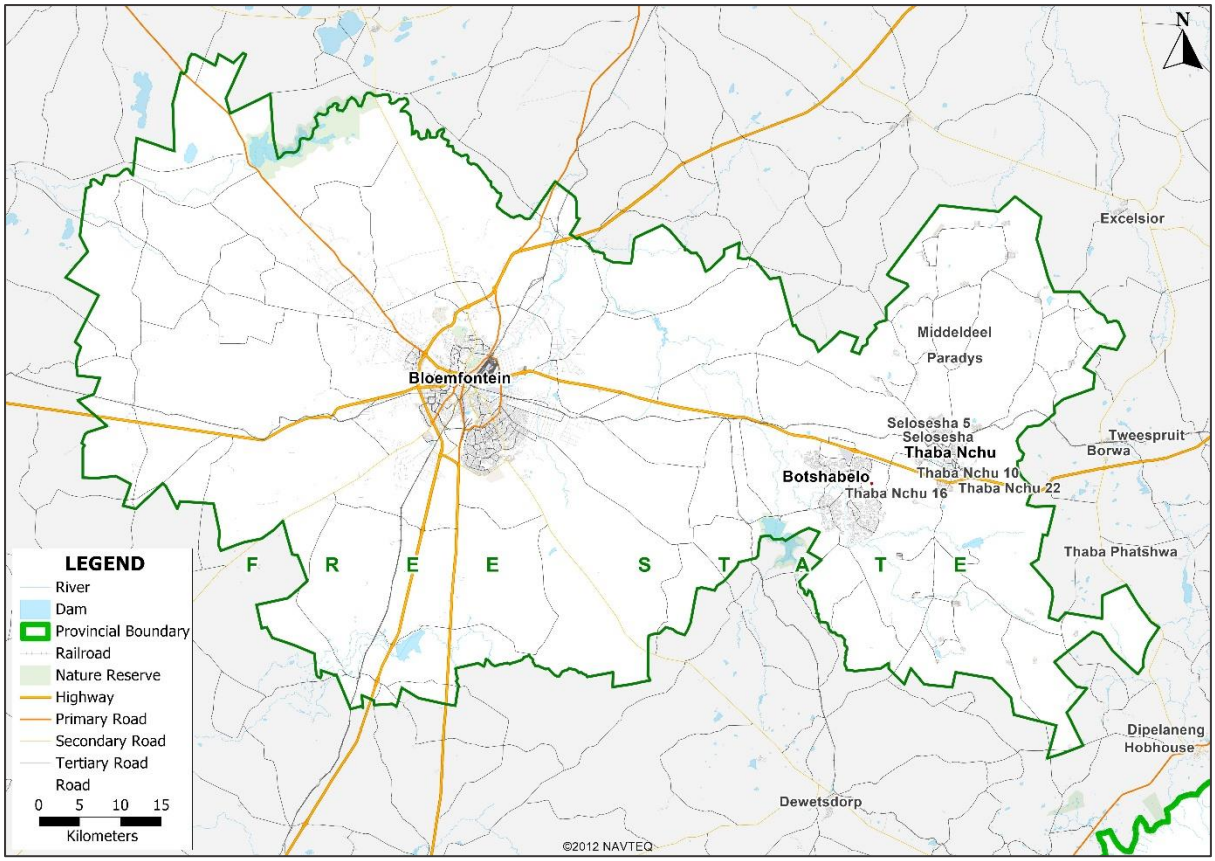


Figure 16: Mangaung Metropolitan Municipality

Source: Map created using data from NAVTEQ (2012), OpenStreetMap (2024), and Statistics South Africa (2011)

Polokwane Local Municipality

The Polokwane Local Municipality is situated in the Limpopo province, bordering Botswana, Zimbabwe, and Mozambique. The municipality spans just over 3 800 square kilometres and has a population of close to 630 000 (Statistics South Africa, 2012c). Polokwane is the largest city in the municipality and serves as the capital of the province. Most people in the municipality reside in traditional residential dwellings (55%), followed by formal residential dwellings (37%) and small holdings (3%) (Statistics South Africa, 2012c).

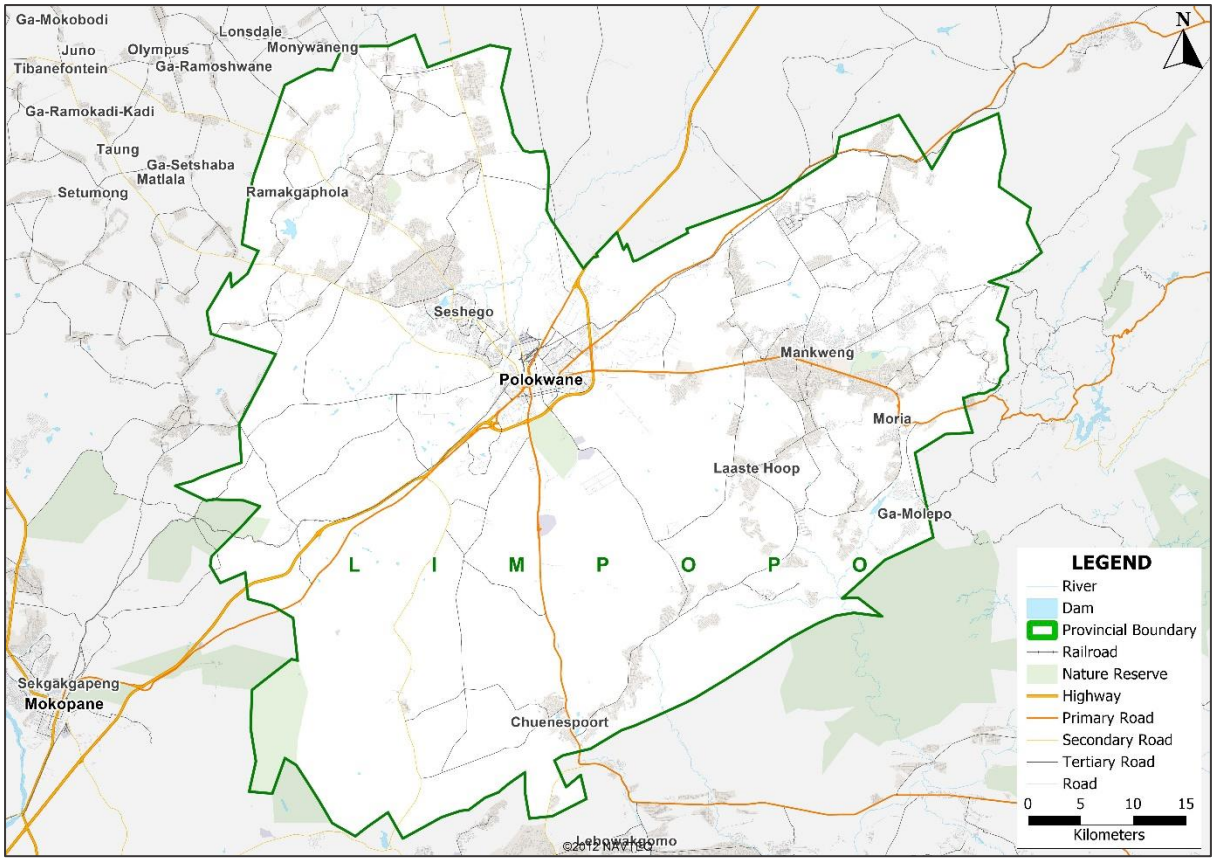


Figure 17: Polokwane Local Municipality

Source: Map created using data from NAVTEQ (2012), OpenStreetMap (2024), and Statistics South Africa (2011)

3.4 Geographic Units

A geographic unit, also referred to as data resolution, “is related to scale, indicates the granularity of the data that are used in mapping” (Slocum et al., 2014). Since geographic units, such as census blocks, wards, and hexagons, are important factors to consider when selecting geospatial data visualisation techniques, various geographic units that depict population distribution were identified for this research.

The aim was to evaluate and test the impact of different geographic units on the effectiveness of data classification methods for choropleth maps. Firstly, geographic units depicting population demographics were selected from the Census 2011 Community Profiles in SuperCROSS database, an official and freely available data source. These include a small area layer and sub-places. Small area layer polygons are the smallest geographical units with demographic data, whereas sub-places are aggregated polygons derived from small area layers that represent suburbs or villages. Secondly, since both small area layer and sub-place polygons vary in size, hexagons were created to represent equal-sized polygons, the researcher also wanted to determine whether equal- or varied-sized polygons could influence participants’ interpretation of choropleth maps. Rather than computing spatial overlays, which

present their own challenges, it was decided to superimpose the Census 2011 population data onto the hexagons. A point data set called Spot Building Count representing dwelling locations was aggregated per hexagon to indicate densities, or rather household densities, per hexagon.

The Spot Building Count data are maintained by Eskom,[8] the main electricity supplier in South Africa. Points are captured from Spot 5 imagery (European Space Agency, n.d.) and verified through various sources, including schools, 1:50 000 topographic data, and dwelling points captured by Statistics South Africa (n.d.). Hexagons were created with an area of two square kilometres, because the sub-place median polygon size of all four municipalities combined is approximately 1.5 square kilometres, which was rounded up to 2 square kilometres. Figure *18* and Figure *19* show the four municipalities depicting equal- and varied-sized polygons: hexagons, small area layer, and sub-places.

[8] https://www.eskom.co.za/

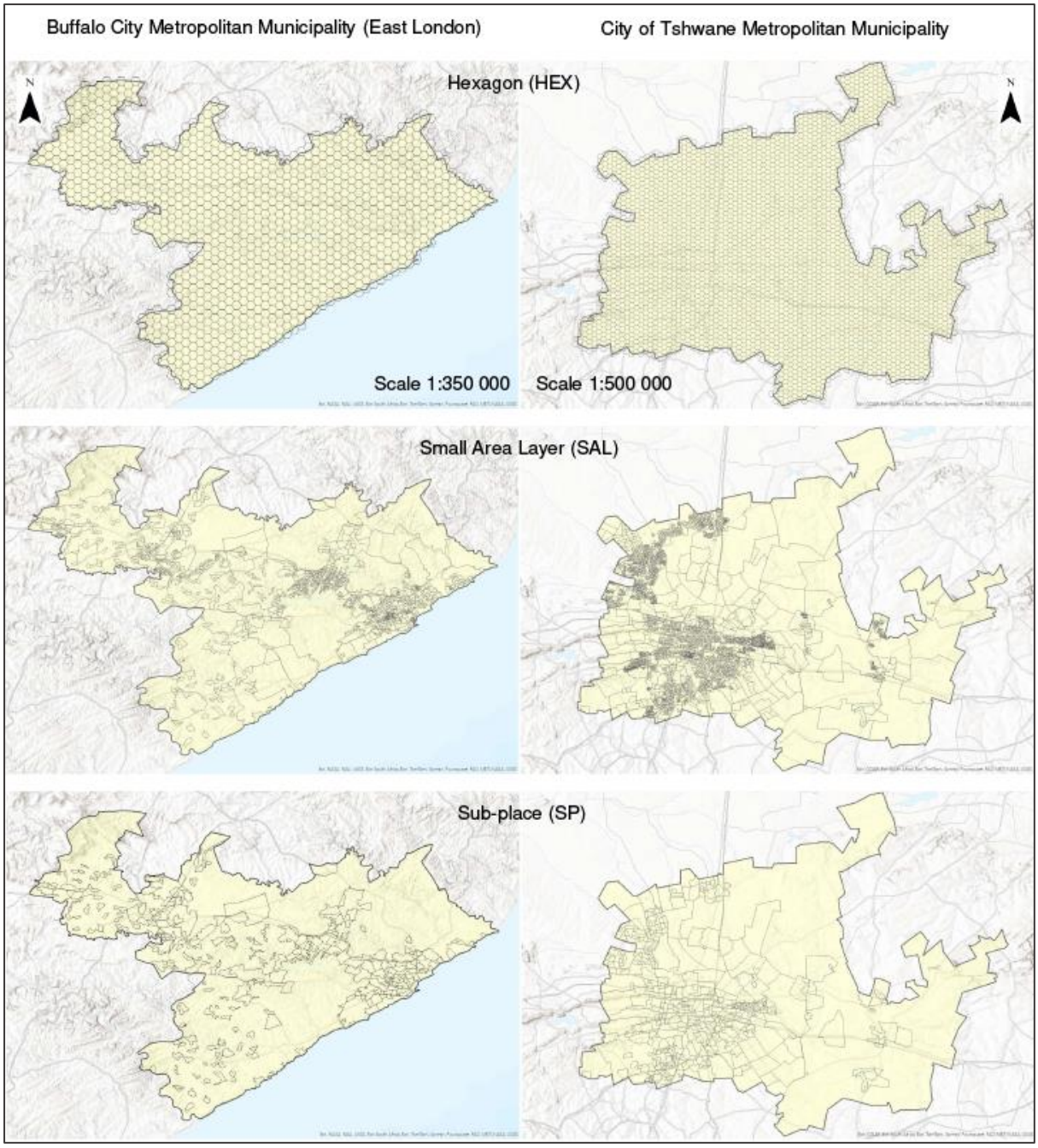


Figure 18: Geographic units for the Buffalo City Metropolitan Municipality and City of Tshwane Metropolitan Municipality

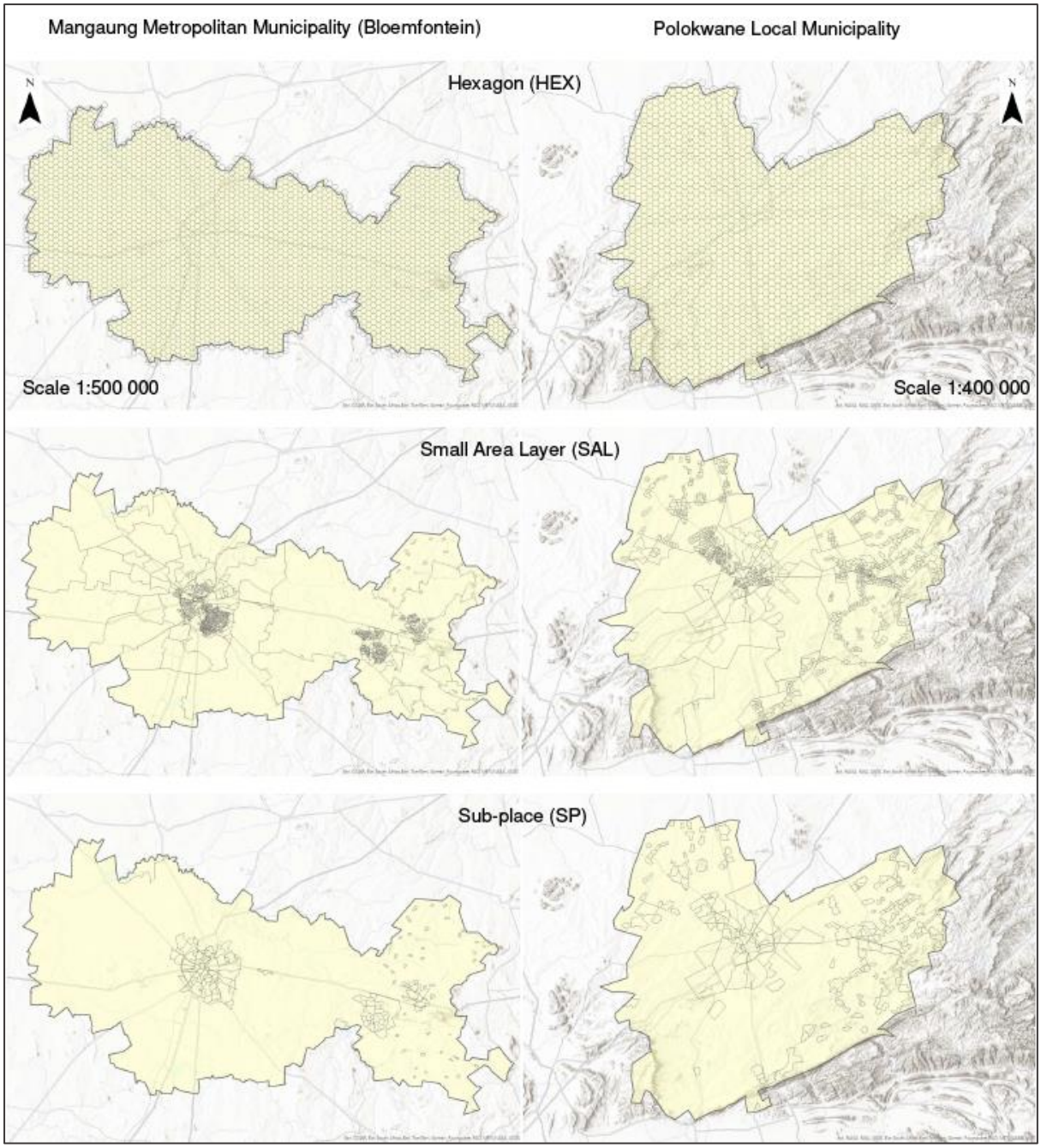


Figure 19: Geographic units for the Mangaung Metropolitan Municipality and Polokwane Local Municipality

3.5 Data Classification

Data classification methods are used for choropleth maps to categorise data by determining upper- and lower-class limits based on a specified number of classes. Ideally, choropleth maps depicting population distribution in South Africa should highlight not only the primary high-density areas, such as city centres and their surroundings, but also populated secondary and tertiary locations, such as townships and informal settlements.

Although various open source and licenced GIS software applications are available and used for spatial analysis, mapping, and visualisation in South Africa, ArcGIS Pro[9] and QGIS[10] (Figure *20*) are considered to be the most popular and frequently used applications globally and locally (GIS Geography, 2022; Khan & Mohiuddin, 2018). Hence, the data classification methods available in both these applications were considered for this research. Bolstad (2012) mentioned that ArcGIS is the most popular GIS software. According to GIS Geography's (2022) ranking of the 30 best GIS software applications, ArcGIS Pro was rated the best, followed by QGIS 3. While other GIS applications are also used in South Africa, they are not compared to ArcGIS Pro and QGIS frequently. These applications include GeoDa, MapInfo, Maptitude, TransCAD, Global Mapper, and SAGA.

Figure 20: ArcGIS Pro and QGIS

There are nine data classification methods available in both ArcGIS Pro and QGIS. These include: defined (or fixed) interval, equal interval, geometric interval, logarithmic scale, manual interval, natural breaks (Jenks), pretty breaks, quantiles, and standard deviation. The geometric interval data classification method is available exclusively in ArcGIS Pro. Methods specific to QGIS include logarithmic scale and pretty breaks. Defined (or fixed) interval, equal interval, manual interval, natural breaks (Jenks), quantiles and standard deviation are available in both ArcGIS Pro and QGIS. Refer to Section 2.3 for examples and a detailed description of these data classification methods.

Except for the manual and defined interval methods, which require users to set custom/manual limits for each class, this study initially identified seven data classification methods, namely equal interval, geometric interval, logarithmic scale, natural breaks (Jenks), pretty breaks, quantiles, and standard deviation.

One of the key aspects recommended in the literature for selecting a data classification method for a specific data set is conducting a data distribution test, also known as a test for normality in the data. Hence, for these seven methods, histograms and descriptive statistics

[9] https://www.esri.com/en-us/home
[10] https://www.qgis.org/en/site/

were generated, showing both population and household distribution per study area and geographic unit (hexagon, small area layer, and sub-place).

The histograms in Figure *21* show the data distribution for each study area and geographic unit. Based on a visual inspection, it is evident that the data are not normally distributed. Descriptive statistics, as shown in *Table 6* to *Table 9*, confirm this. The skewness and kurtosis measure the "degree of normality of distributions, or the lack thereof" (Ho & Yu, 2015). A skewness value between −0.5 and 0.5 suggests a normal distribution of frequencies (Hatem et al., 2022).

The results suggest that the household distribution by hexagon and population density by small area layer and sub-place are not normally distributed in any of the four study areas; instead, they are highly skewed. The degree of skewness of the four study areas per geographic unit ranges from 2.49 to 8.02. This indicates that the majority of polygons exhibit a low population distribution (or low household count for hexagons and low population density for small area layers and sub-places), while only a small number of outliers show extremely high density.

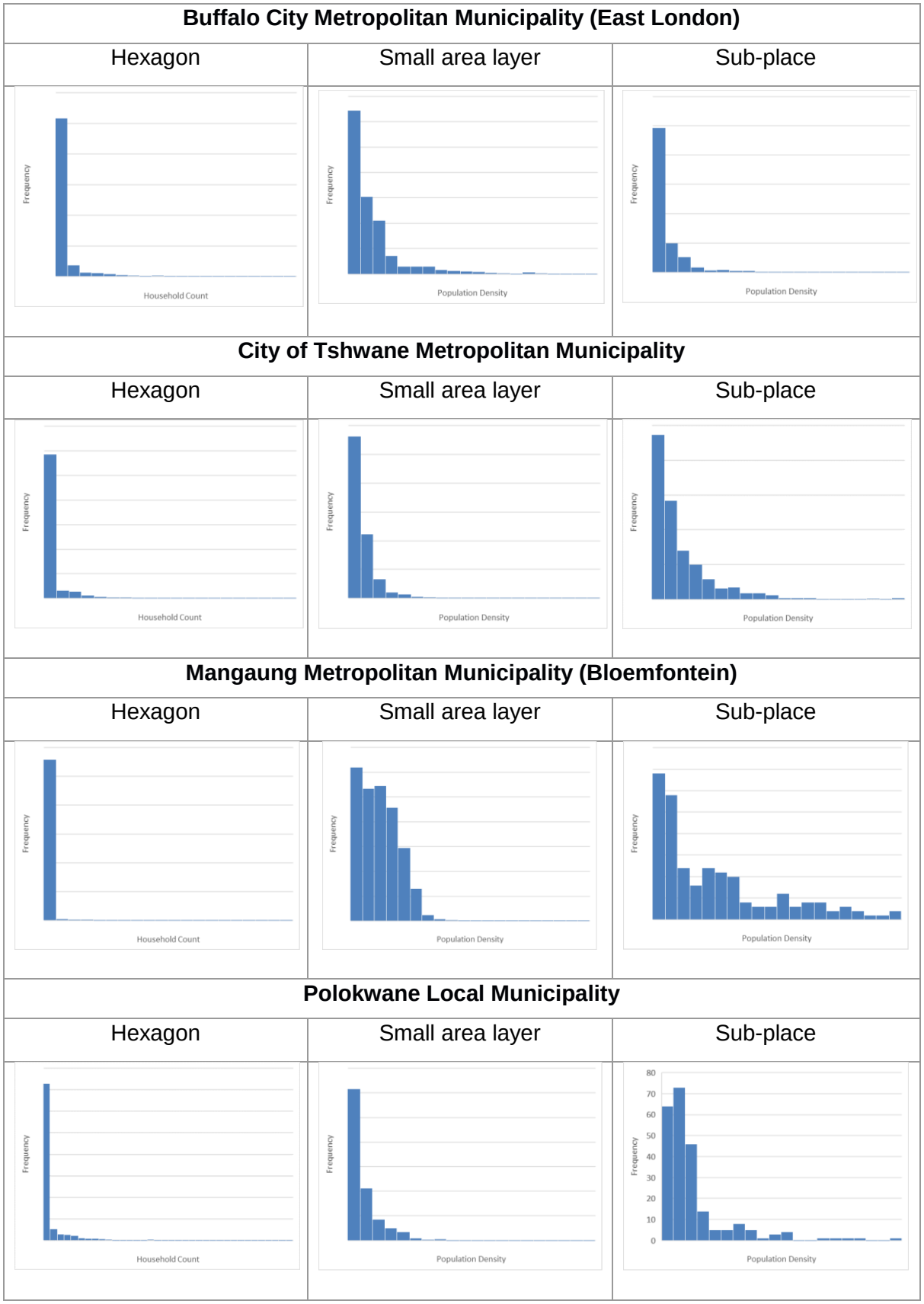


Figure 21: Histograms showing data distribution for each study area per geographic unit

Table 6: Descriptive statistics for Buffalo City Metropolitan Municipality

Buffalo City Metropolitan Municipality (East London) – Descriptives

	TYPE		Statistic	Std. Error
Count	Hexagon – Household count	Mean	162.69	11.518
		Median	20.00	
		Std. Deviation	400.813	
		Minimum	0	
		Maximum	5 192	
		N	1 211	
		Skewness	4.844	0.070
		Kurtosis	33.565	0.140
	Small area layer – Population density	Mean	5 270.11	180.390
		Median	3 116.77	
		Std. Deviation	6 715.746	
		Minimum	1	
		Maximum	55 690	
		N	1 386	
		Skewness	2.835	0.066
		Kurtosis	10.311	0.131
	Sub-place – Population density	Mean	1 946.80	185.049
		Median	931.28	
		Std. Deviation	3 447.080	
		Minimum	0	
		Maximum	34 544	
		N	347	
		Skewness	5.279	0.131
		Kurtosis	37.045	0.261

Table 7: Descriptive statistics for City of Tshwane Metropolitan Municipality

City of Tshwane Metropolitan Municipality – Descriptives

	TYPE		Statistic	Std. Error
Count	Hexagon – Household count	Mean	193.98	8.626
		Median	19.50	
		Std. Deviation	504.134	
		Minimum	0	
		Maximum	6 851	
		N	3 416	
		Skewness	4.765	0.042
		Kurtosis	32.086	0.084
	Small area layer – Population density	Mean	7 046.88	129.917
		Median	4 484.17	
		Std. Deviation	8 738.313	
		Minimum	1	
		Maximum	127 262	
		N	594	
		Skewness	4.509	0.036
		Kurtosis	33.953	0.073
	Sub-place – Population density	Mean	3 408.64	169.267
		Median	2 195.19	
		Std. Deviation	4 125.388	
		Minimum	0	
		Maximum	31 648	
		N	31 648	
		Skewness	2.491	0.100
		Kurtosis	9.457	0.200

Table 8: Descriptive statistics for Mangaung Metropolitan Municipality

Mangaung Metropolitan Municipality (Bloemfontein) – Descriptives

	TYPE		Statistic	Std. Error
Count	Hexagon – Household count	Mean	63.71	6.027
		Median	1.00	
		Std. Deviation	325.762	
		Minimum	0	
		Maximum	4 586	
		N	2 921	
		Skewness	8.020	0.045
		Kurtosis	74.624	0.091
	Small area layer – Population density	Mean	5 270.11	180.390
		Median	3 116.77	
		Std. Deviation	6 715.746	
		Minimum	1	
		Maximum	55 690	
		N	1 386	
		Skewness	2.835	0.066
		Kurtosis	10.311	0.131
	Sub-place – Population density	Mean	1 946.80	185.049
		Median	931.28	
		Std. Deviation	3 447.080	
		Minimum	0	
		Maximum	34 544	
		N	347	
		Skewness	5.279	0.131
		Kurtosis	37.045	0.261

Table 9: Descriptive statistics for Polokwane Local Municipality

Polokwane Local Municipality – Descriptives

	TYPE		Statistic	Std. Error
Count	Hexagon – Household count	Mean	86.99	5.525
		Median	4.00	
		Std. Deviation	237.826	
		Minimum	0	
		Maximum	2 447	
		N	1 853	
		Skewness	4.982	0.057
		Kurtosis	31.772	0.114
	Small area layer – Population density	Mean	2 746.80	114.345
		Median	1 585.38	
		Std. Deviation	3 664.407	
		Minimum	0	
		Maximum	41 990	
		N	1 027	
		Skewness	4.967	0.076
		Kurtosis	39.967	0.152
	Sub-place – Population density	Mean	1 305.02	100.149
		Median	899.76	
		Std. Deviation	1 528.705	
		Minimum	0	
		Maximum	10 266	
		N	233	
		Skewness	2.742	0.159
		Kurtosis	9.411	0.318

Based on the skewness of the population distribution for each study area and geographic unit (hexagon, sub-place and small area layer), the standard deviation, equal interval and pretty breaks data classification methods were excluded from further analysis, as these methods are best suited for data that are normally distributed (Slocum et al., 2014; Tyner, 2014; Vasilca, 2019). Furthermore, equal intervals and pretty breaks are most effective when the data are uniformly distributed, meaning the data distribution has no peak and is consistent (Kraak et al., 2021). Hence, these methods would not effectively represent data distribution in South Africa, nor would the data variability or nuances stand out.

The remaining data classification methods that were tested in the user study along with their accuracy score measurements (or errors between class breaks), include:

- Geometric interval,
- Logarithmic scale,
- Natural breaks (Jenks), and
- Quantiles.

3.6 Map Design

A total of 48 choropleth maps were created to depict the four study areas (local and metropolitan municipalities), three geographic units (hexagons, small area layers, and sub-places), and four data classification methods (see Figure *23* to *Figure 34*). Based on the recommendations found in the literature, five class intervals were used for all the maps where possible. The default number of classes in both ArcGIS Pro and QGIS was set to five.

Most GIS software applications include a variety of built-in colour and pattern schemes to display differences between class intervals when designing a choropleth map. Nevertheless, the researcher opted for the well-known generic colour scheme design web application, ColorBrewer,[11] (Figure *22*) to select differential colours for each class interval. To ensure consistency, a single colour scheme was used for all maps. ColorBrewer is a free web application, created by Brewer, that allows users to specify the number of data classes and select a predefined colour scheme. All maps were standardised by using a single sequential colour scheme ranging from light yellow to dark blue. Light grey was chosen for the boundary's outline colour.

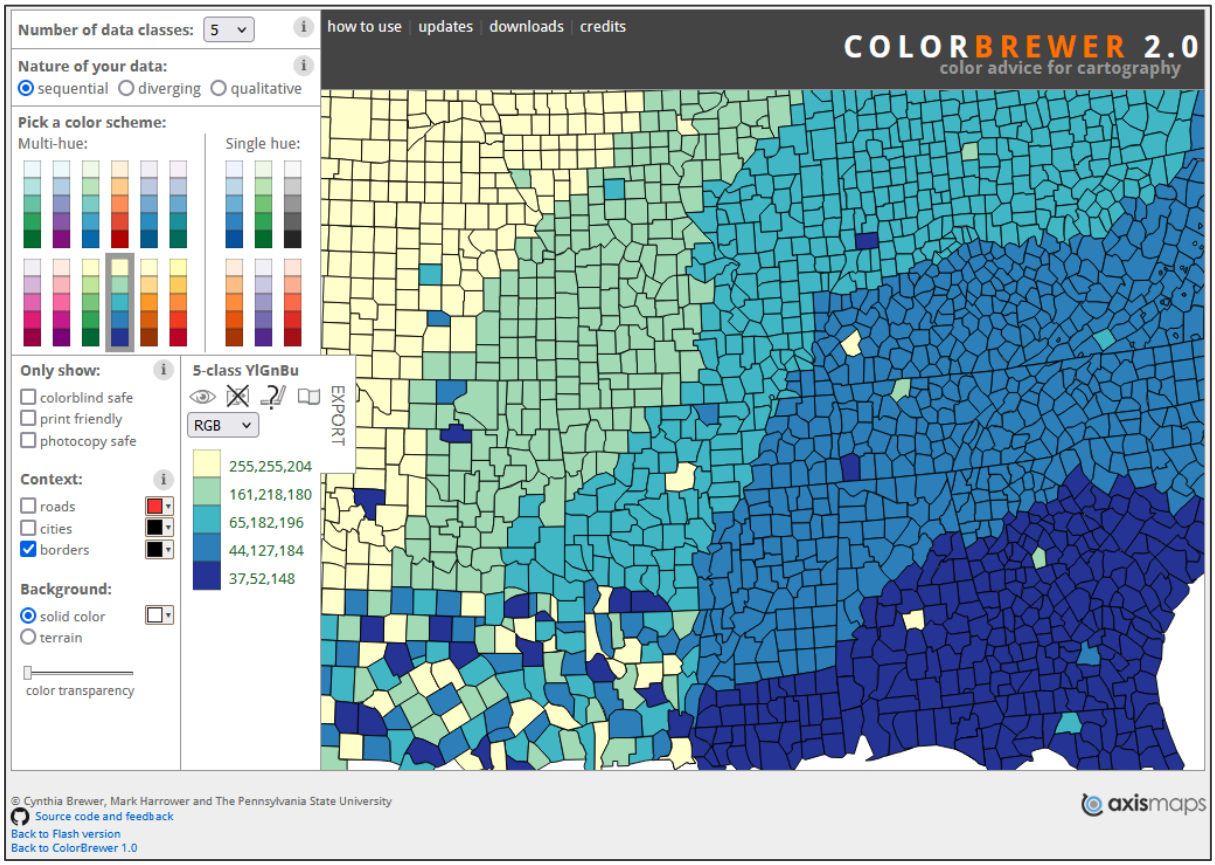


Figure 22: *ColorBrewer 2*.0

[11] https://colorbrewer2.org/

Although ColorBrewer requires a certain degree of manual work for the map designer – specifically copying RGB or hex values for each colour and manually updating the five auto-generated colour schemes in ArcGIS Pro and QGIS – a similar colour scheme is built in and available in both GIS software applications: Yellow-Green-Blue in ArcGIS Pro and YIGnBu in QGIS.

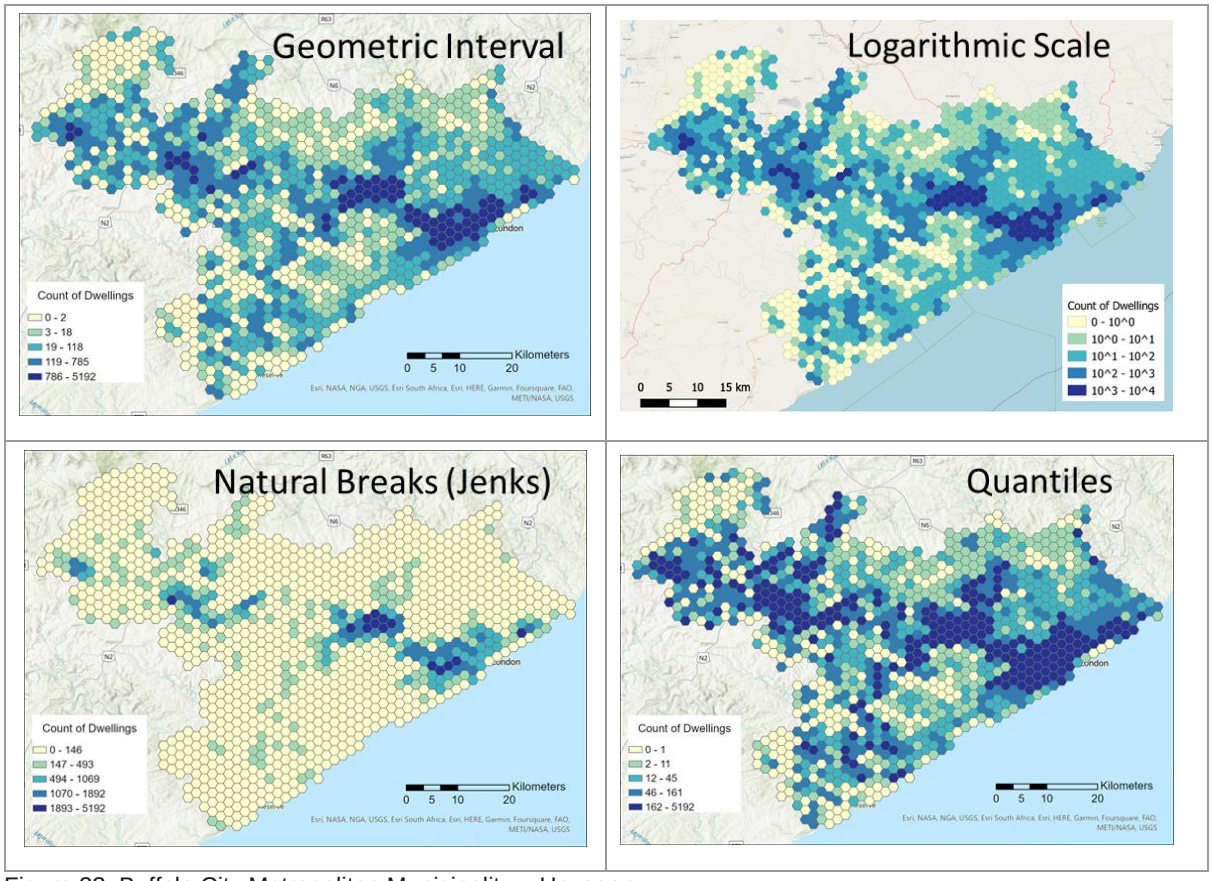


Figure 23: Buffalo City Metropolitan Municipality – Hexagon

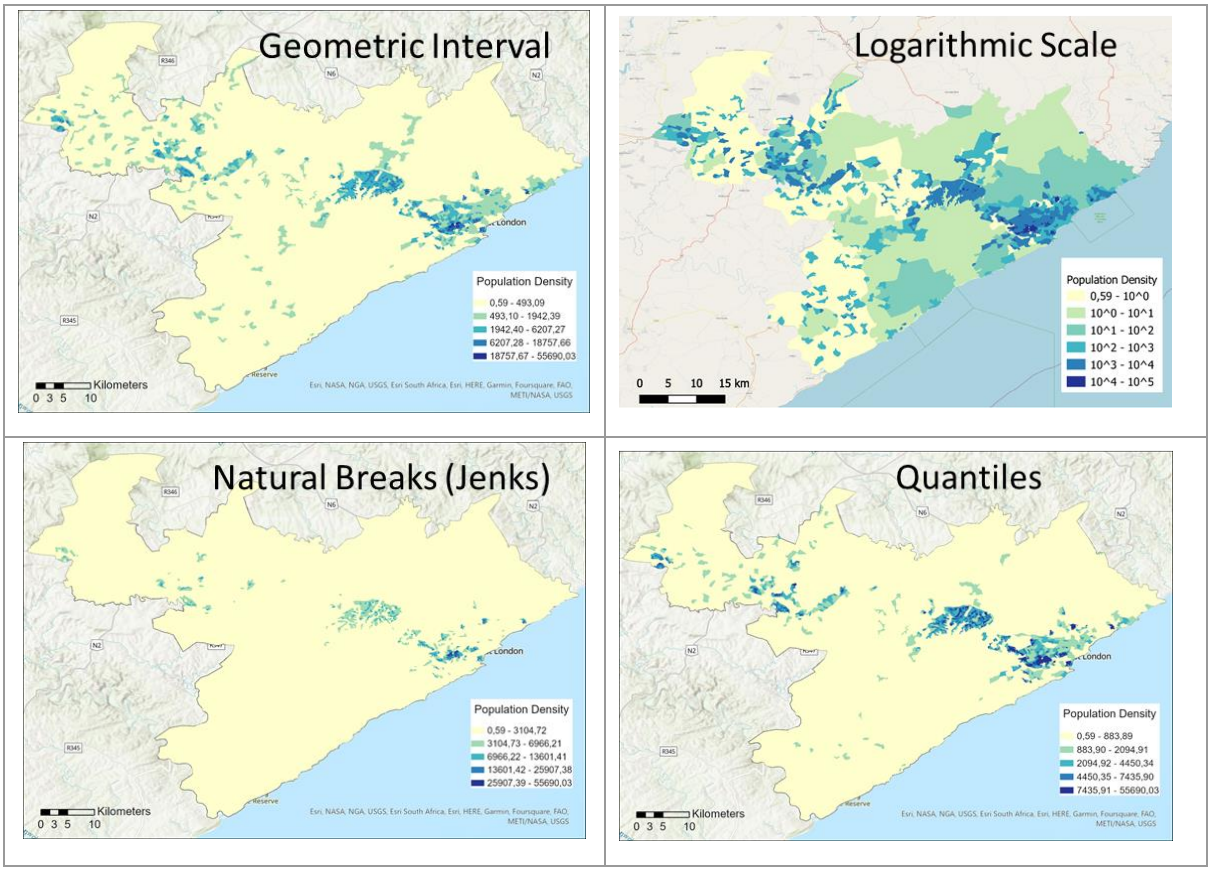


Figure 24: Buffalo City Metropolitan Municipality – Small area layer

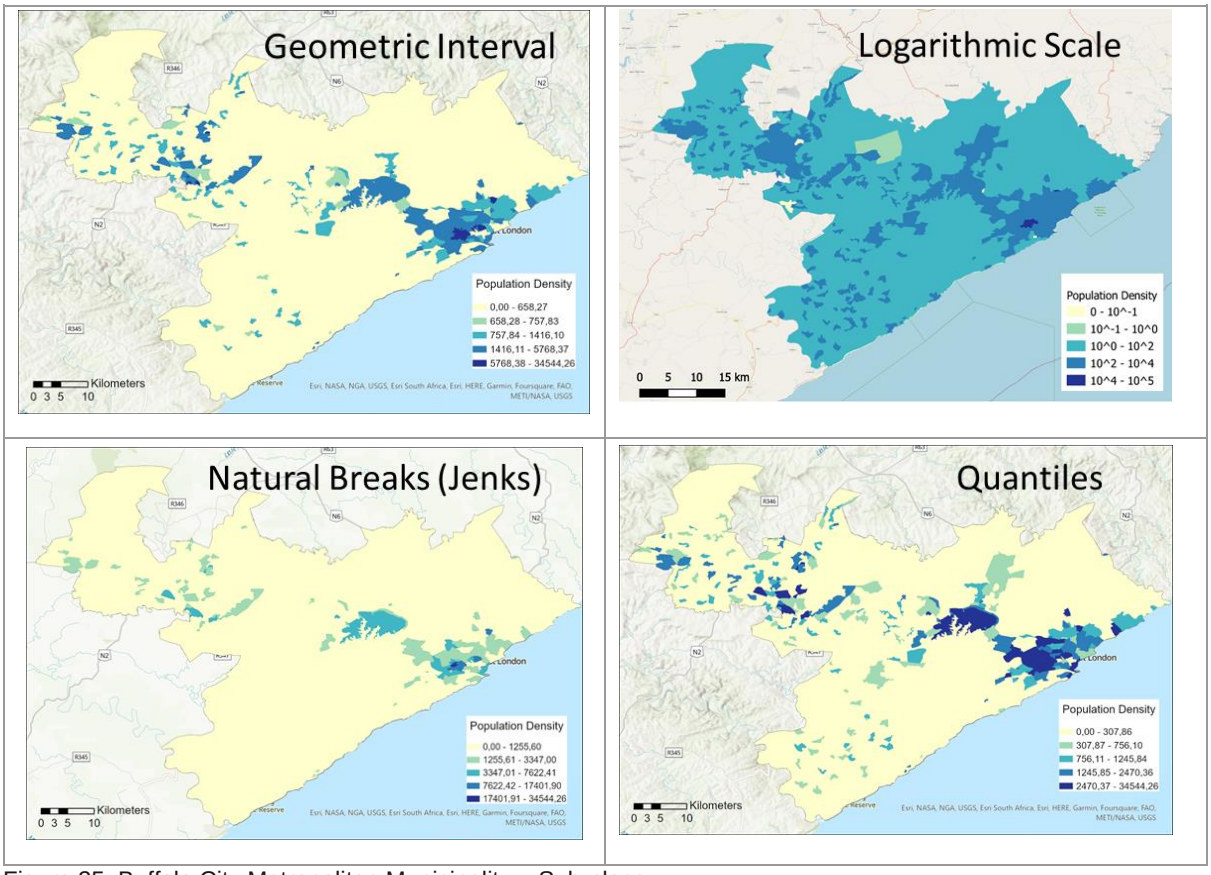


Figure 25: Buffalo City Metropolitan Municipality – Sub-place

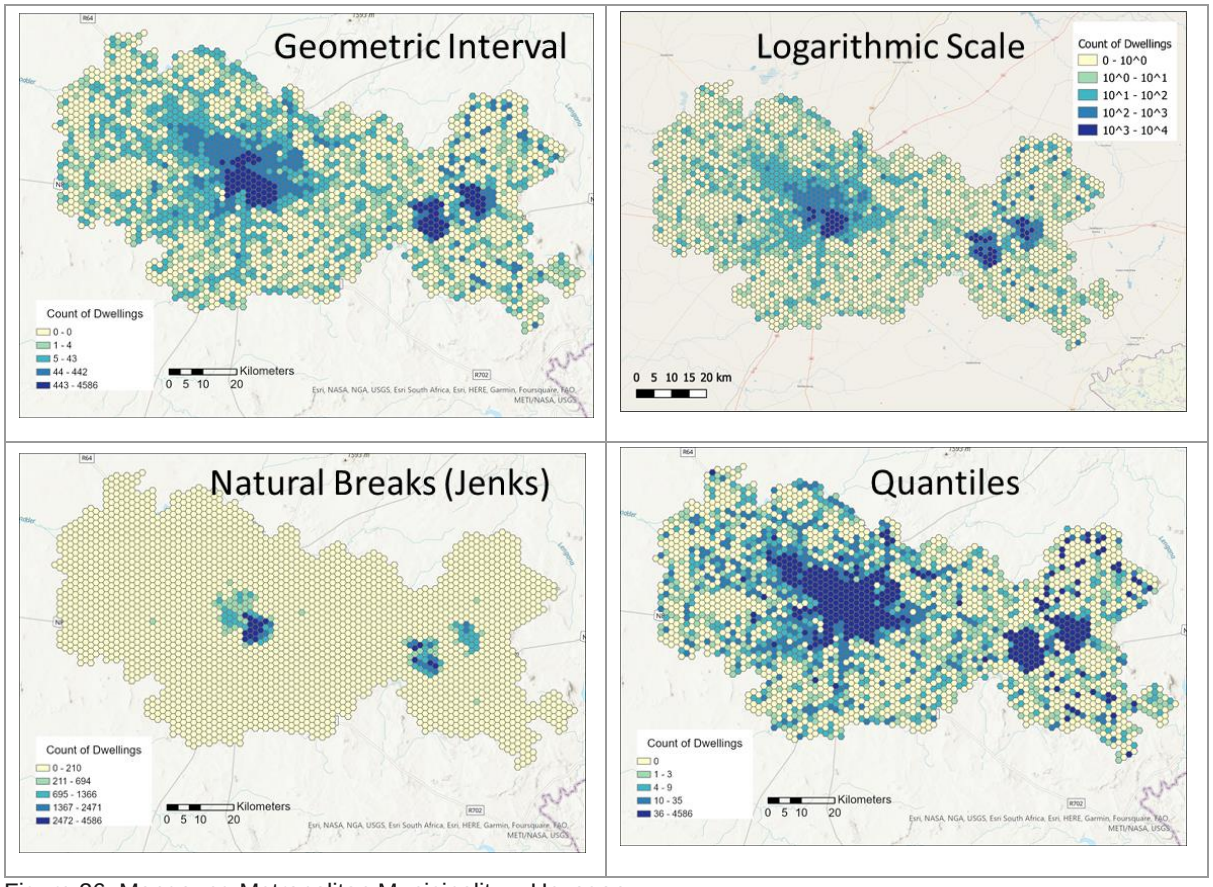


Figure 26: Mangaung Metropolitan Municipality – Hexagon

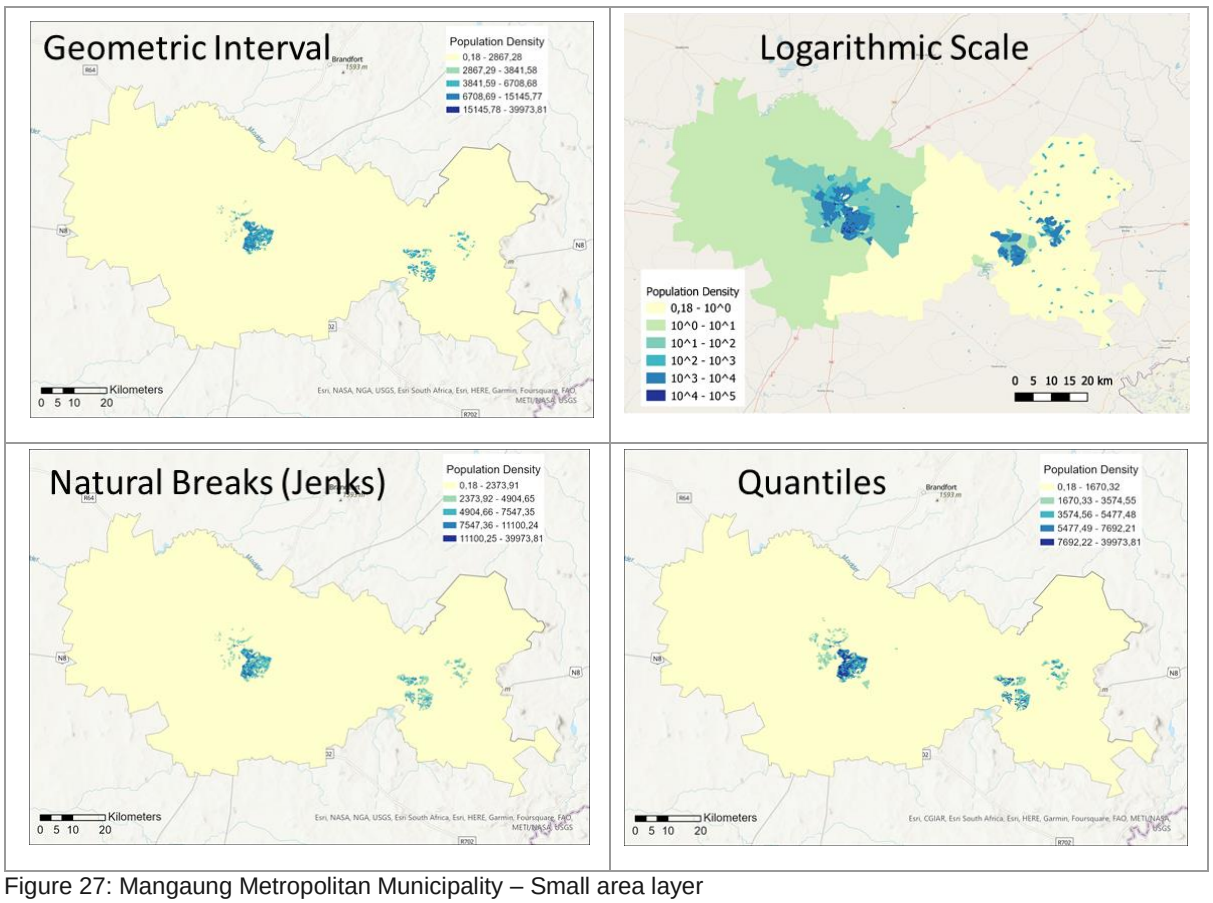


Figure 27: Mangaung Metropolitan Municipality – Small area layer

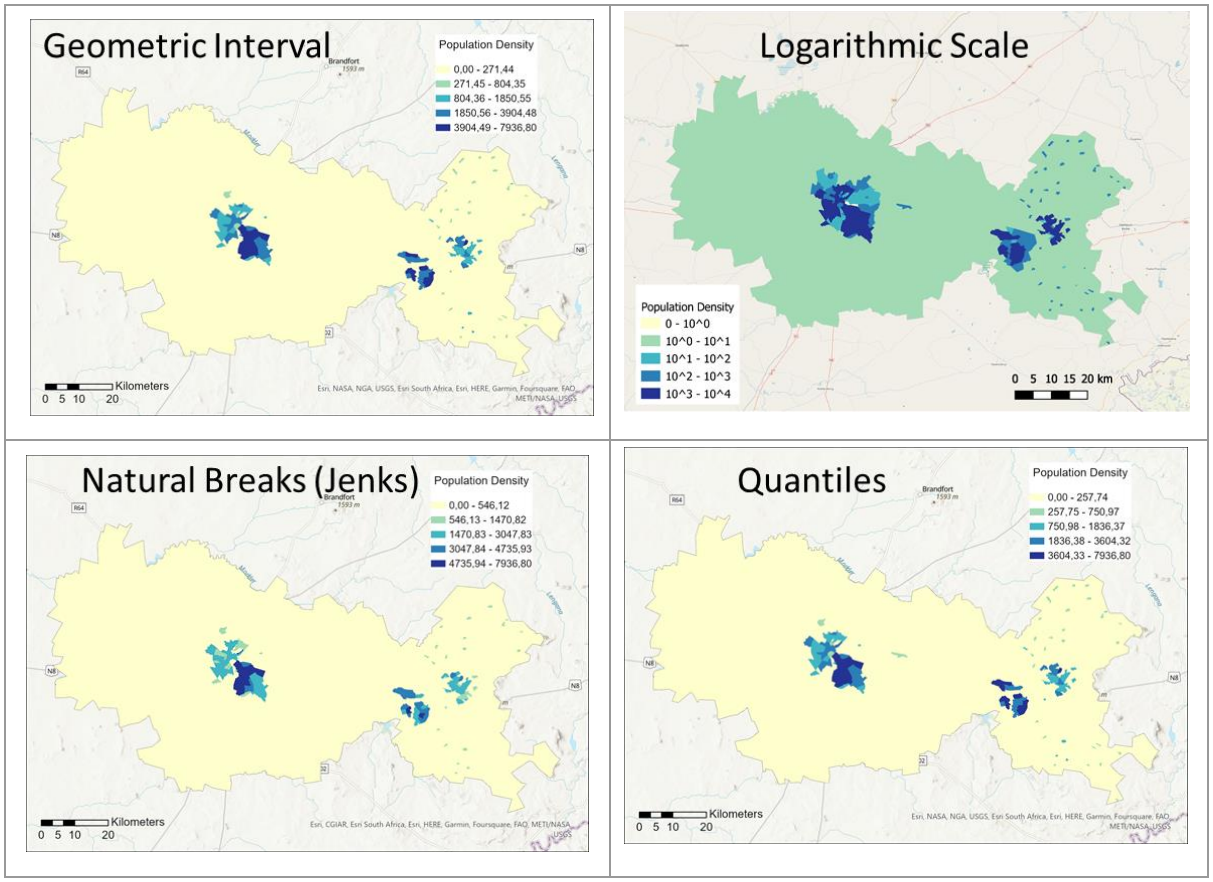


Figure 28: Mangaung Metropolitan Municipality – Sub-place

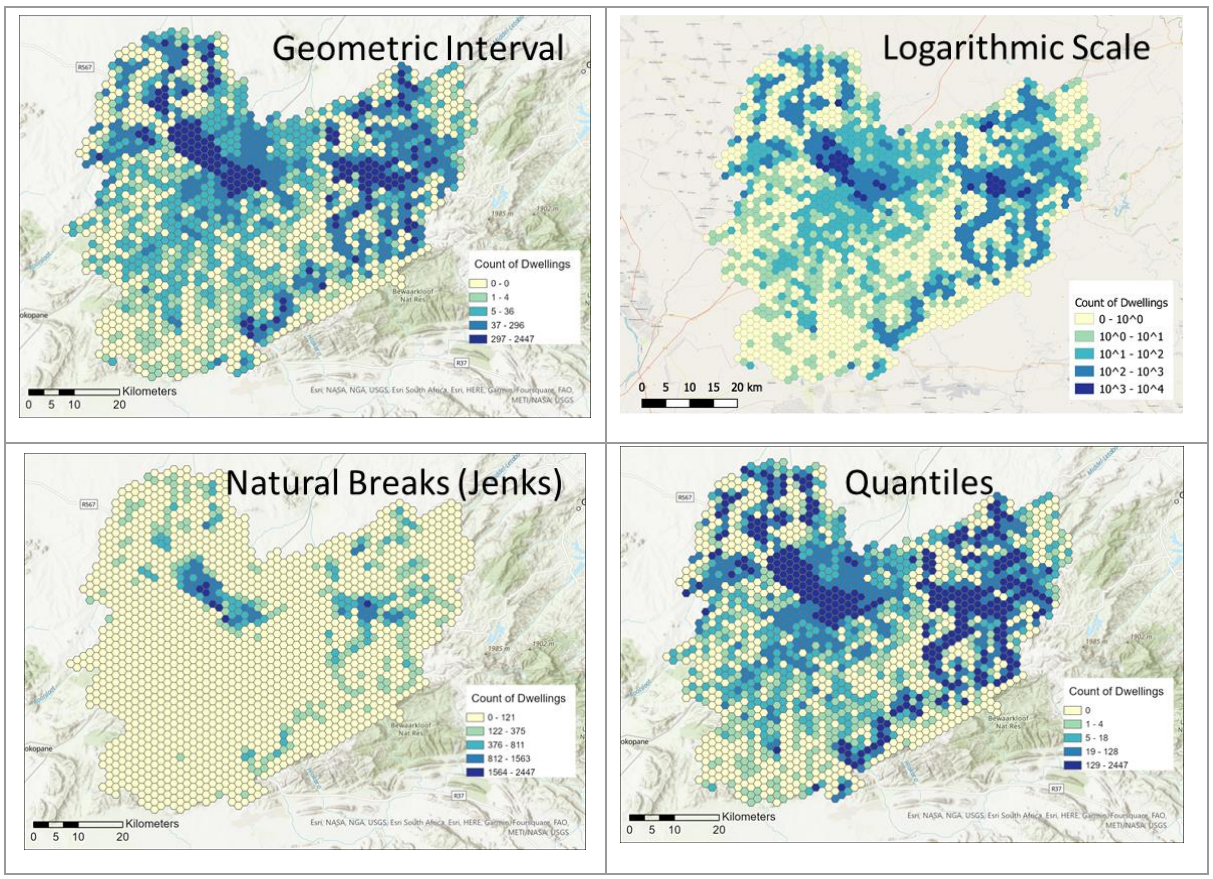


Figure 29: Polokwane Local Municipality – Hexagon

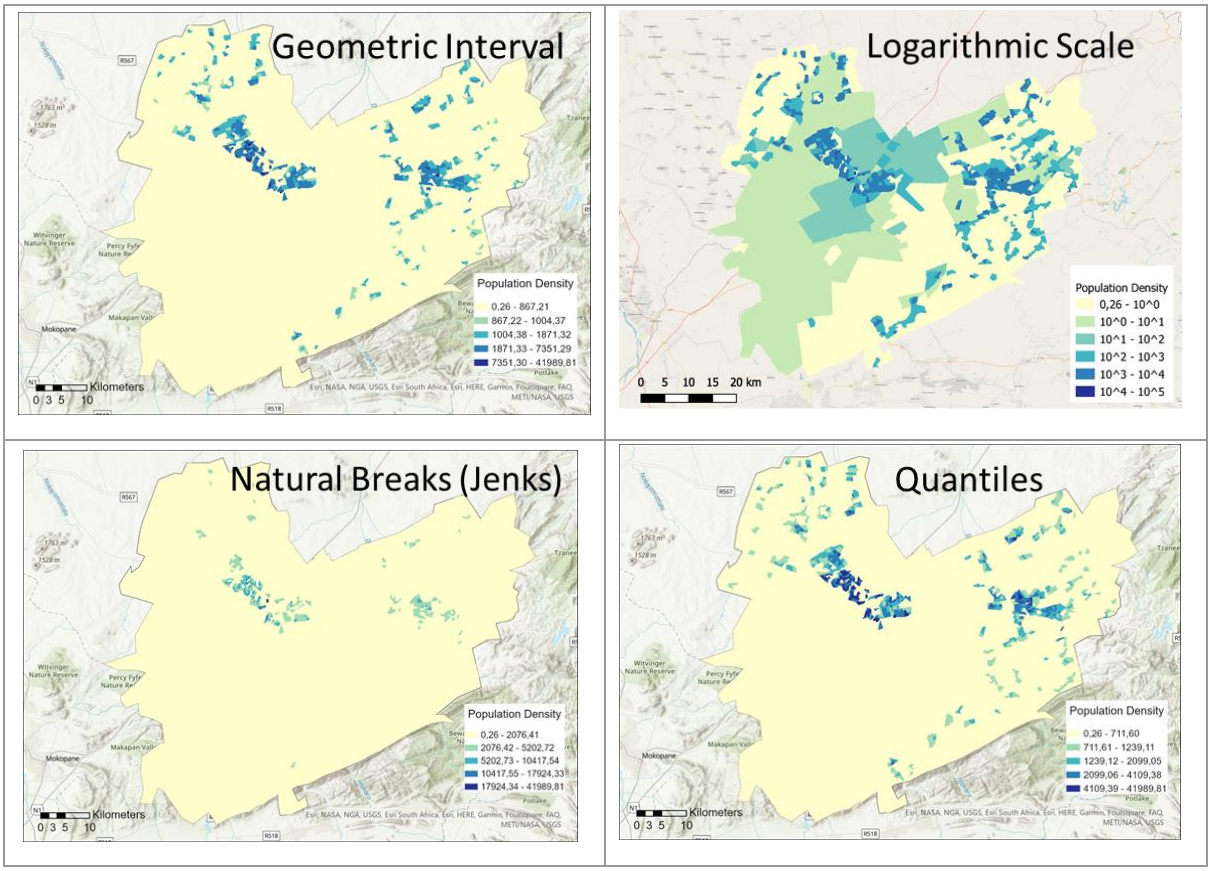


Figure 30: Polokwane Local Municipality – Small area layer

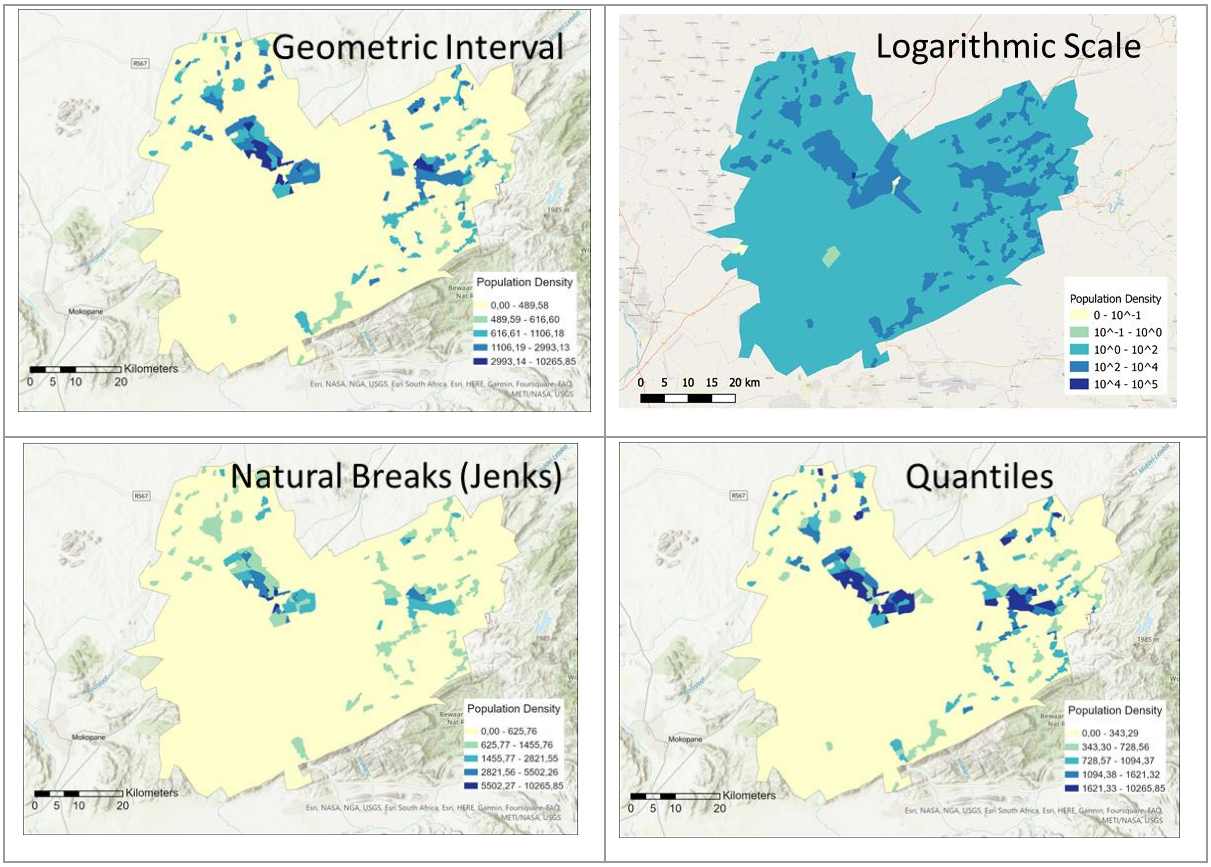


Figure 31: Polokwane Local Municipality – Sub-place

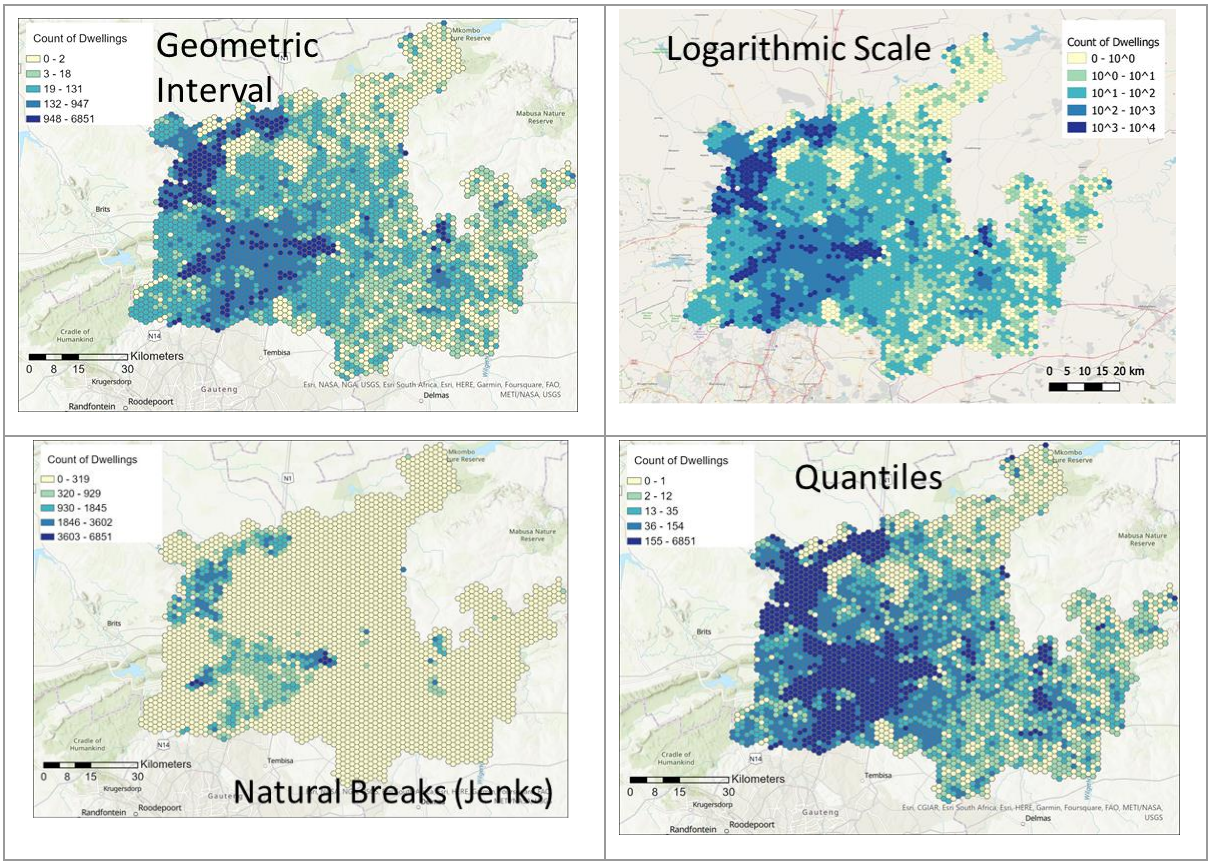


Figure 32: City of Tshwane Metropolitan Municipality – Hexagon

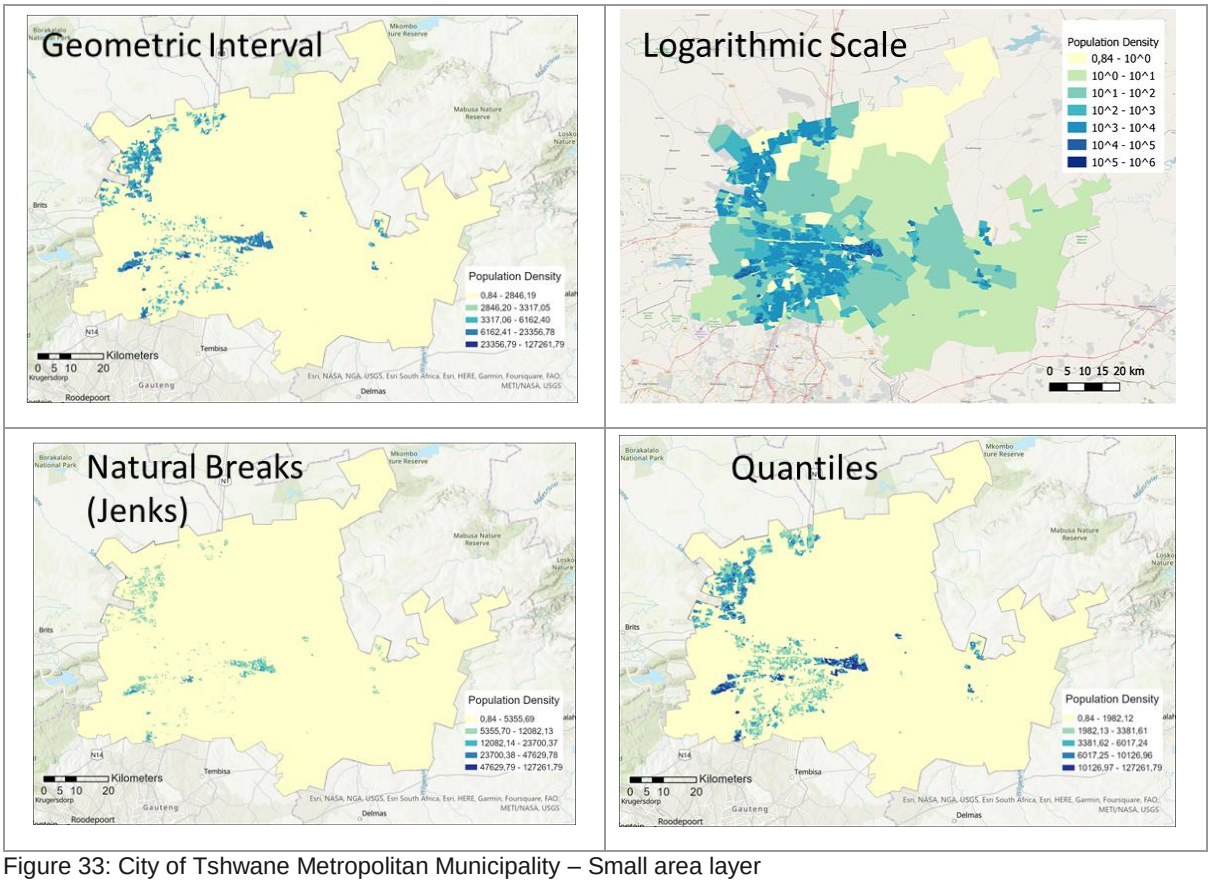


Figure 33: City of Tshwane Metropolitan Municipality – Small area layer

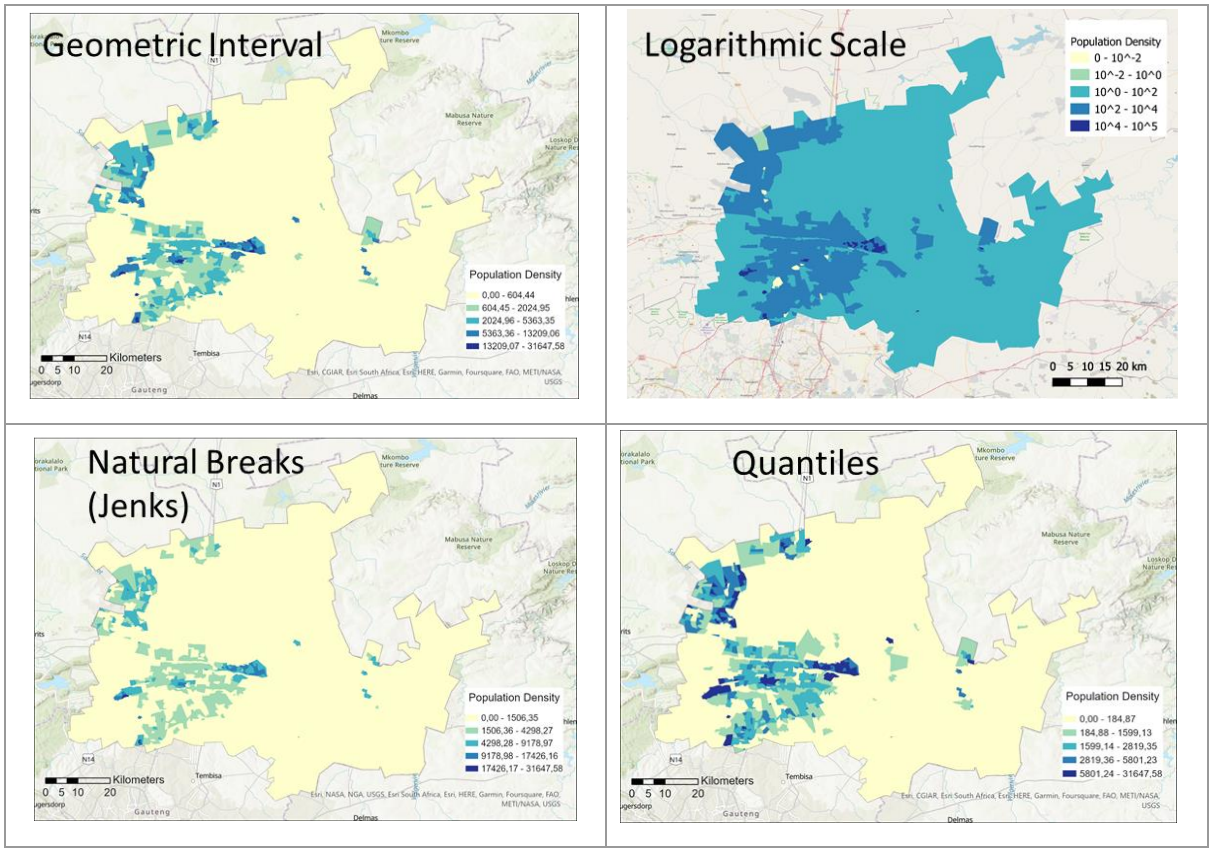


Figure 34: City of Tshwane Metropolitan Municipality – Sub-place

Chapter 3 described the process followed to design choropleth maps so that the maps could be assessed with a user study (described in the next chapter) as well as through a mathematical equation that measures the error between class breaks for each data classification method (Chapter 5). A total of 48 choropleth maps were created depicting the four selected study areas (local and metropolitan municipalities), three geographic units (hexagons, small areas, and sub-places), and four data classification methods: geometric interval, logarithmic scale, natural breaks (Jenks), and quantiles.

4. USER STUDY

4.1 Introduction

Chapter 4 evaluates the suitability of data classification methods for choropleth maps depicting population demand (Research Objective 3: Identify the most suitable data classification method(s) to visualise population demand for decision makers in geographic accessibility studies in South Africa). Four data classification methods, available in both ArcGIS Pro and QGIS, were selected and assessed through a user study.

The chapter is divided into four parts. Section 4.2 describes the user study design. Section 4.3 gives an overview of the respondents who participated. Section 4.4 assesses the results derived from the user study, and Section 4.5 concludes with a discussion of the key findings. See Figure *35*.

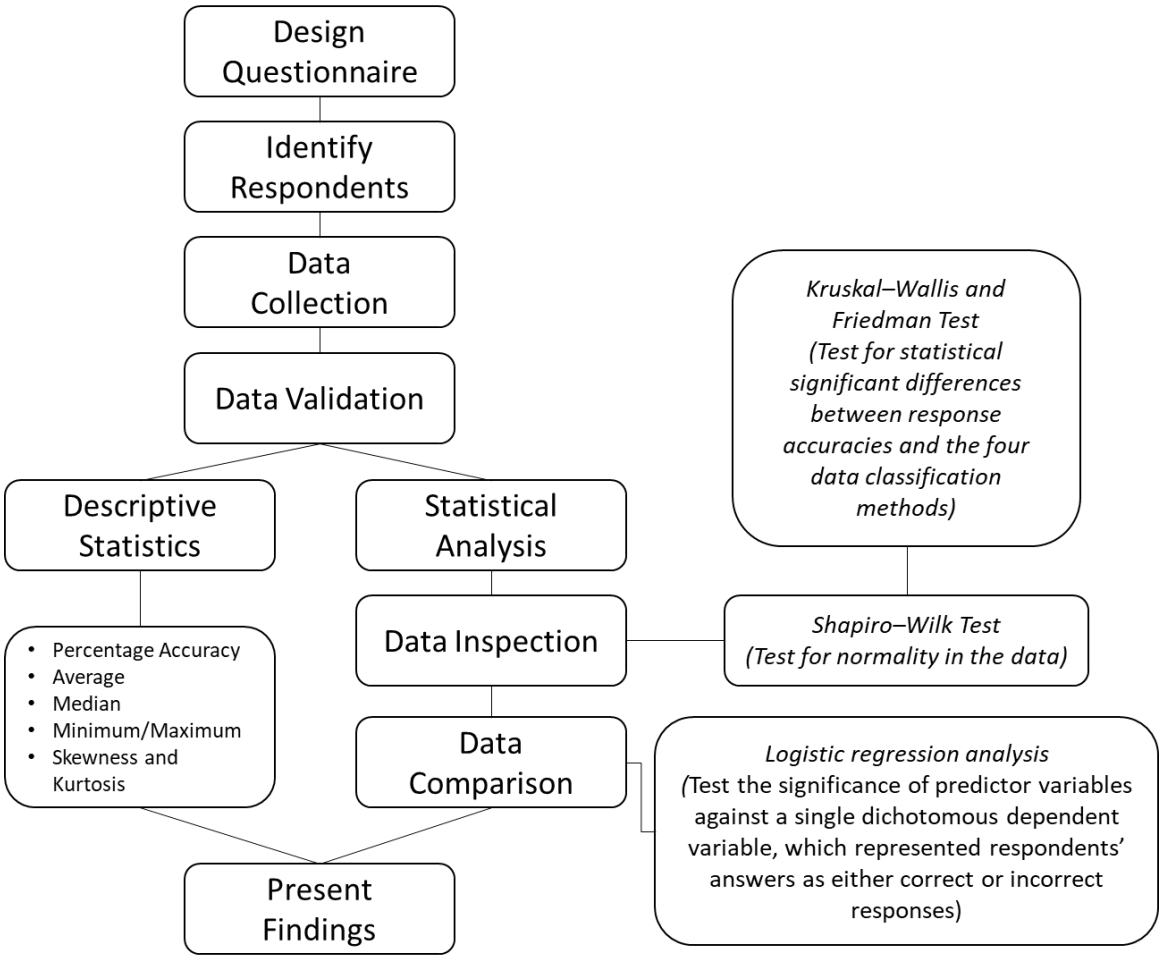


Figure 35: User study flow diagram

The results of the user study were used to elicit recommendations regarding the suitability of each method. These recommendations are useful for non-GIS professionals who need to

analyse population demand in South Africa but are unsure which data classification method will effectively display the country's unique population distribution using choropleth maps.

4.2 Study Design

The user study includes a survey featuring an online questionnaire designed to assess respondents' interpretations of choropleth maps depicting population demand (see Appendix C). The survey was created using Qualtrics,[12] a secure online platform specifically designed for building and customising online questionnaires. Qualtrics enables researchers to create survey projects, allowing them to select and customise from a variety of built-in methods to design questions, such as radio buttons, drop-down menus, and click events. Click events are useful when assessing respondents' interpretation and understanding of images. For example, respondents were asked to identify specific locations by clicking on a map. The questionnaire was accessible online through a web address link. Therefore, an internet connection was required to access and complete the questionnaire.

Qualtrics also captures general statistics such as the starting date and time, as well as the duration to complete the questionnaire, including the time taken for each question. The structure of the questionnaire comprised four parts:

- Introduction and background,
- Demographic, academic, and skills characteristics of respondents,
- Map literacy and colour vision tests, and
- Choropleth map assessment.

4.2.1 Introduction and Background

The starting page of the questionnaire included a brief introduction, providing respondents with a concise description of the research goal, intended outcome, and potential applications of the results. The key concepts and terminology, including geographic accessibility and choropleth maps, were explained, followed by a brief overview of the questionnaire structure. Content regarding research ethics and participants' consent was also added. See the text below.

[12] https://www.qualtrics.com/uk/

Dear participant

This survey evaluates the interpretation of different data classification methods for choropleth maps. Results from the survey will be used to develop a set of good practices for professionals who need to visualise population demand with choropleth maps in order to conduct geographic accessibility studies.

What is geographic accessibility and how is it measured?

Geographic accessibility is measured by calculating the physical distance people travel to specific facilities or service centres (such as clinics, community centres, police stations etc.). "Travelling long distances to reach these centres is costly and time consuming, especially to those who suffer the burden of poverty and deprivation". Measuring geographic accessibility enables policy makers to implement effective strategies for the optimal positioning of service centres (close to the people).

Choropleth (or thematic) maps are frequently used to visualise population demand for geographic accessibility studies. It is however noted that these maps are sometimes incorrectly interpreted or misunderstood which leads to ineffective optimisation strategies. These strategies include recommendations for (1) where to open a new service centre, i.e. the 'expansion model' (2) where to close a current service centre, i.e. the 'reduction model' and (3) where to move a current service centre to a different location, i.e. the 'relocation model'.

What you need to do

You will be shown 2 general map reading questions, as well as 48 map interpretation questions specifically related to geographic accessibility (where to open, close or move service centres). Please read each question carefully and then follow the instructions for providing an answer.

Note that consent cannot be withdrawn once the questionnaire is submitted as there is no way to trace the particular questionnaire that has been filled in. Please answer the questions in the questionnaire as completely and honestly as possible. This should not take more than 30 minutes of your time. This study has received written approval from Research Ethics Committees of the Faculty of Natural and Agricultural Sciences (tel: 012 420 4356). The results of the survey may be published in the media and/or an academic journal without identifying any of the participants individually. We will provide you with a summary of our findings on request. If you have any questions or comments, please do not hesitate to ask the facilitators or contact Lourens Snyman, lourens.snyman@up.ac.za

By clicking 'Next', you agree to the above terms and provide your consent to use these results in this study.

Thank you

4.2.2 Demographic, Academic and Skills Characteristics of Respondents

The first part of the questionnaire was designed to capture the respondents' demographic characteristics and their current level of education. Demographics included age and gender. Details of their current academic enrolment programme were logged, including a summary of modules (subjects) that they are either busy with or have completed. Additionally, various self-evaluation questions were designed, and respondents were asked to rate their level of training or proficiency in different categories related to visualisation, cartography, GIS, and data analysis. These categories included map reading, geography, statistics, cartography, planning, topographic maps, statistical maps, spatial data, web browsers, English as a language, and Google Maps. The rating scale ranged from 0 to 10, with a score of 10 being the highest.

4.2.3 Map Literacy and Colour Vision Test

These questions were added to establish a general baseline of map literacy among of respondents and also to evaluate their ability to differentiate different colour scales, which could influence their ability to visualise choropleth maps. The two general map literacy questions were designed to test spatial orientation skills related to distance, direction, and the general identification of map features (Table *10*).

Table 10: Map literacy questions

Definition	Question
Distance measurement	What is the approximate distance (straight-line) between points A and B on the map?
Sense of direction	If I travel from Pretoria to Kungwini, I will be travelling in a(n) _____ direction.

For the colour vision test, respondents were first asked whether a professional had ever informed them that they had imperfect colour vision. This was followed by a test in which they were required to differentiate between light and dark colour schemes. See *Figure 36*.

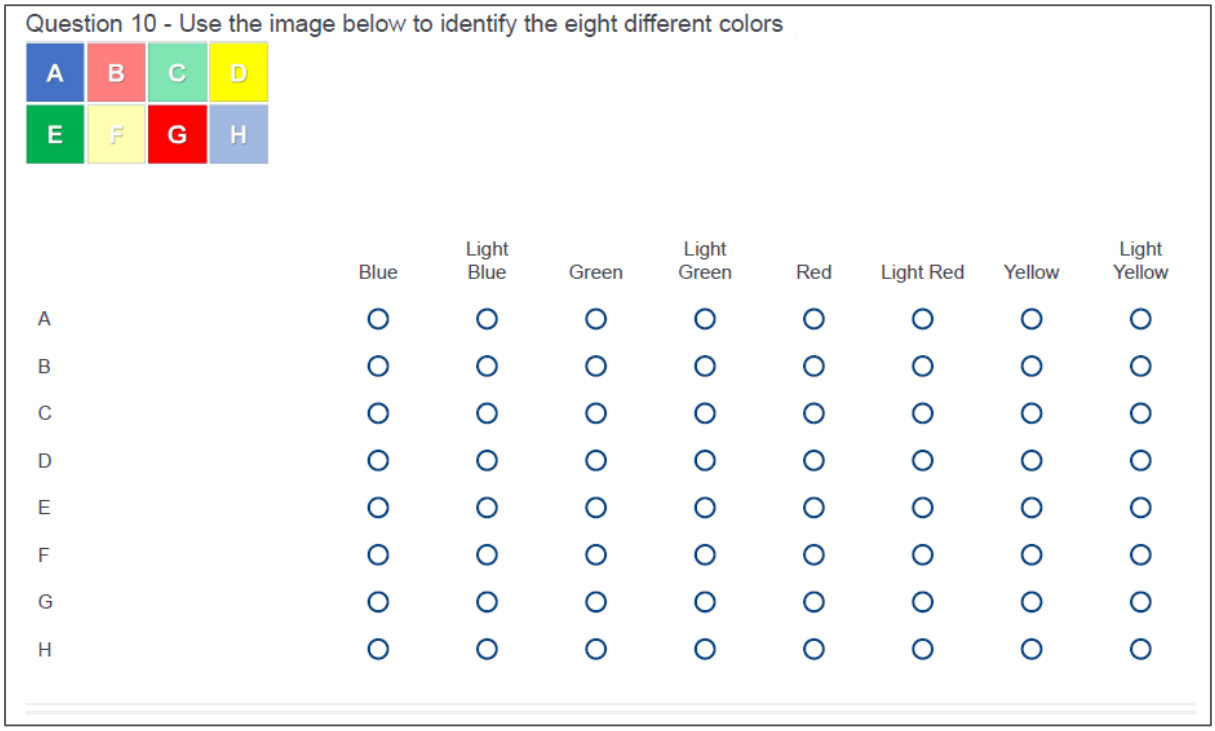


Figure 36: Colour vision test

4.2.4 Choropleth Map Assessment

The last part, which covers the main research objectives, includes map-specific questions (see Table *11*) where respondents' interpretations of choropleth maps depicting different data classification methods were assessed. Questions related to geographic accessibility were designed according to real-world scenarios. Respondents were required to differentiate between densely and sparsely populated areas and to identify locations that are over- or underserved for the optimal provision of service centres. Questions were structured based on the following facility location models identified in the DPSA's (2012) geographic access guideline:

- Expansion model,
- Reduction model, and
- Relocation model.

Respondents were required to click on relevant locations on the map. The locations of existing service centres were randomly plotted; hence, they do not reflect an actual geographic footprint.

Four questions were designed, each with varying levels of difficulty. Question 1 is considered the easiest, while Question 4 is considered the most difficult. Additionally, each question required respondents to click on a specific number of locations on the map, referred to here as 'click events'. The underlying data classification method used for each choropleth map was not disclosed to respondents.

Table 11: Geographic accessibility questions

Definition	Question	No. of Click Events
Q1:	Identify three areas on the map where the dwelling count is very low. Click on the relevant areas.	3
Q2:	Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster (areas with a high dwelling count). The maximum threshold (catchment area of a service centre) is 10 km, i.e. a centre serves only people within 10 km from it. Anybody living further away is not served by that centre.	1
Q3:	Identify two locations on the map to add additional service centres. These centres should be located in high-density clusters (areas with a high dwelling count) with no other facility nearby (further than 10 km away from existing centres).	2
Q4:	Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would-be high-density clusters (areas with a high dwelling count) with no other facility close by (within 10 km).	2

A total of 48 static choropleth maps were used, covering the four study areas (local and metropolitan municipalities), three geographic units, and four data classification methods, as described in Chapter 3.

Respondents were assigned exactly the same questions in the same sequence (within-subject participant assignment). One disadvantage of this method is a potential learning effect, meaning that a response to a question could be influenced or affected by previous questions. Thus, to eliminate sequential dependencies of responses and the potential learning effect, both the geographic unit and locality (study areas) changed throughout the questionnaire (Table *12*). For reference, geographic units include hexagons, small area layers, and sub-places. The four localities (or study areas) are:

- A – City of Tshwane Metropolitan Municipality,
- B – Buffalo City Metropolitan Municipality,
- C – Polokwane Local Municipality, and
- D – Mangaung Metropolitan Municipality.

Table 12: Structure of the questionnaire

Data Classification Method	Hexagon				Small area layer				Sub-place			
Geometric interval	1	2	3	4	5	6	7	8	9	10	11	12
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	A	B	C	D	B	C	D	A	C	D	A	B
Logarithmic scale	13	14	15	16	17	18	19	20	21	22	23	24
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	D	A	B	C	A	B	C	D	B	C	D	A
Natural breaks (Jenks)	25	26	27	28	29	30	31	32	33	34	35	36
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	C	D	A	B	D	A	B	C	A	B	C	D
Quantiles	37	38	39	40	41	42	43	44	45	46	47	48
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	B	C	D	A	C	D	A	B	D	A	B	C

For example, Question 22 (highlighted in orange in Table *12*) includes:

(1) Identify the best area for opening a new service centre, (2) the location (or study area) is Polokwane Local Municipality (C), and (3) the choropleth map shows population density on a sub-place level based on the logarithmic scale data classification method (see Figure 37).

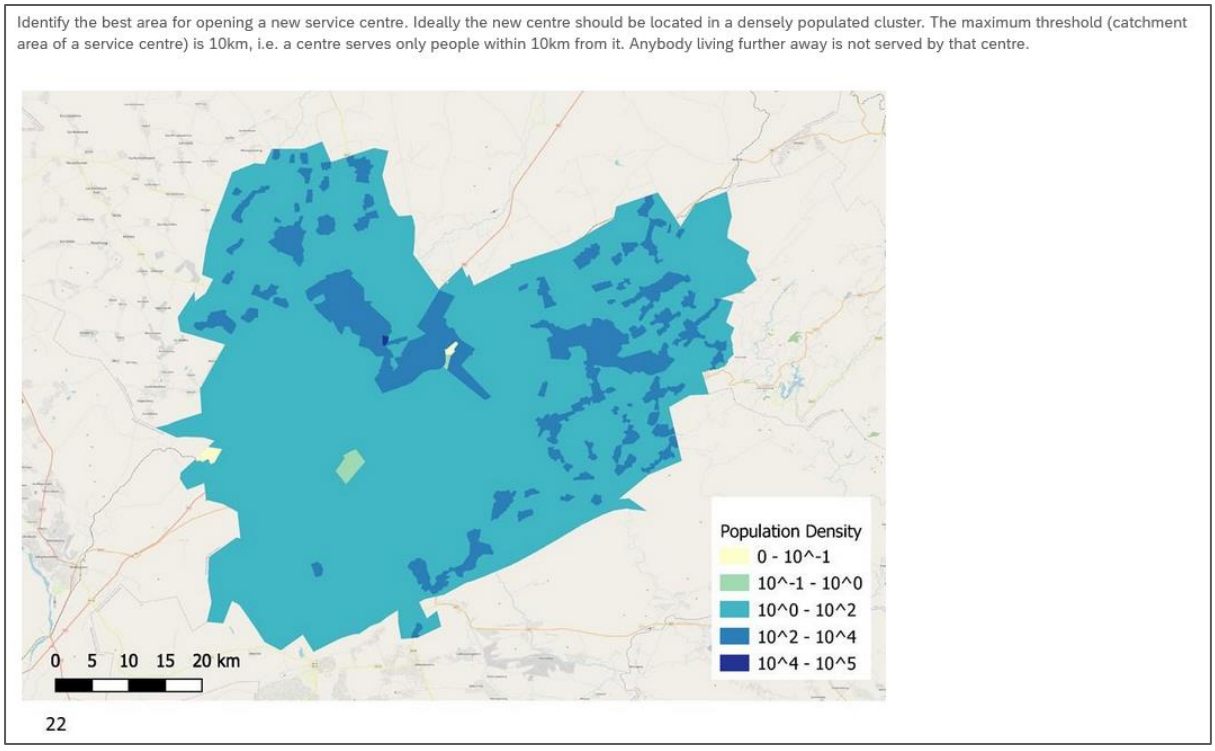


Figure 37: Question 22 – Identify the best area for opening a new service centre

4.3 Respondents

Participation was completely voluntary, and a few lucky draw cash prizes were given to those who completed the entire survey. Participants were students from the University of Pretoria, enrolled in programmes or taking a module at the Department of Geography, Geoinformatics and Meteorology (GGM), with and without prior experience in geography and GIS. Students were chosen for the user study because they, in one way or another, represent the upcoming workforce of the country.

The online questionnaire was available from June 2023 to October 2023, for a duration of four months. Before making the questionnaire available to participants, a pilot was conducted as recommended by Regmi et al. (2016) with three selected respondents to test overall functionality, ease of use, format of the captured data, and, lastly, the duration to complete the questionnaire.

A total of 229 students participated in the user study, excluding those who completed the pilot study. Unfortunately, not all of them completed the questionnaire. Of the total number of participants, only 165 started with the 48 geographic accessibility questions, and of those, 107 completed the entire questionnaire. Only responses from those who completed the entire questionnaire were used for analysis. *Figure 38* shows the number of respondents per geographic accessibility question. The graph clearly shows the drop-out rate as respondents worked through the 48 questions.

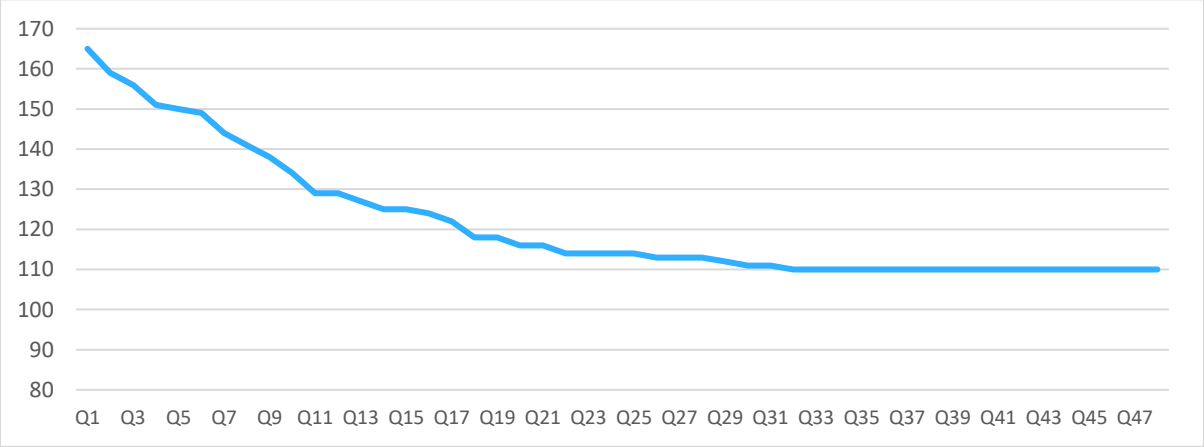


Figure 38: Number of respondents per geographic accessibility question

The age of the respondents ranged from 18 to 27, with the majority falling within the 19–21-year-old age group. Most were female (64%). Males comprised 35%, and one student preferred not to disclose their specific gender.

From the 107 participants, most were registered in the BEd programme (Education) at the time of the survey (40%), followed by BSc Geography and Environmental Science, and BSc

Geoinformatics with 14% and 11%, respectively. Although BEd students are not enrolled in a programme in the Department of GGM, they do have the option to select subjects from GGM. *Figure 39* shows the percentage of participants in each academic programme. The blue bars indicate programmes in which students have already been introduced to GIS. These comprised 34.6% of participants.

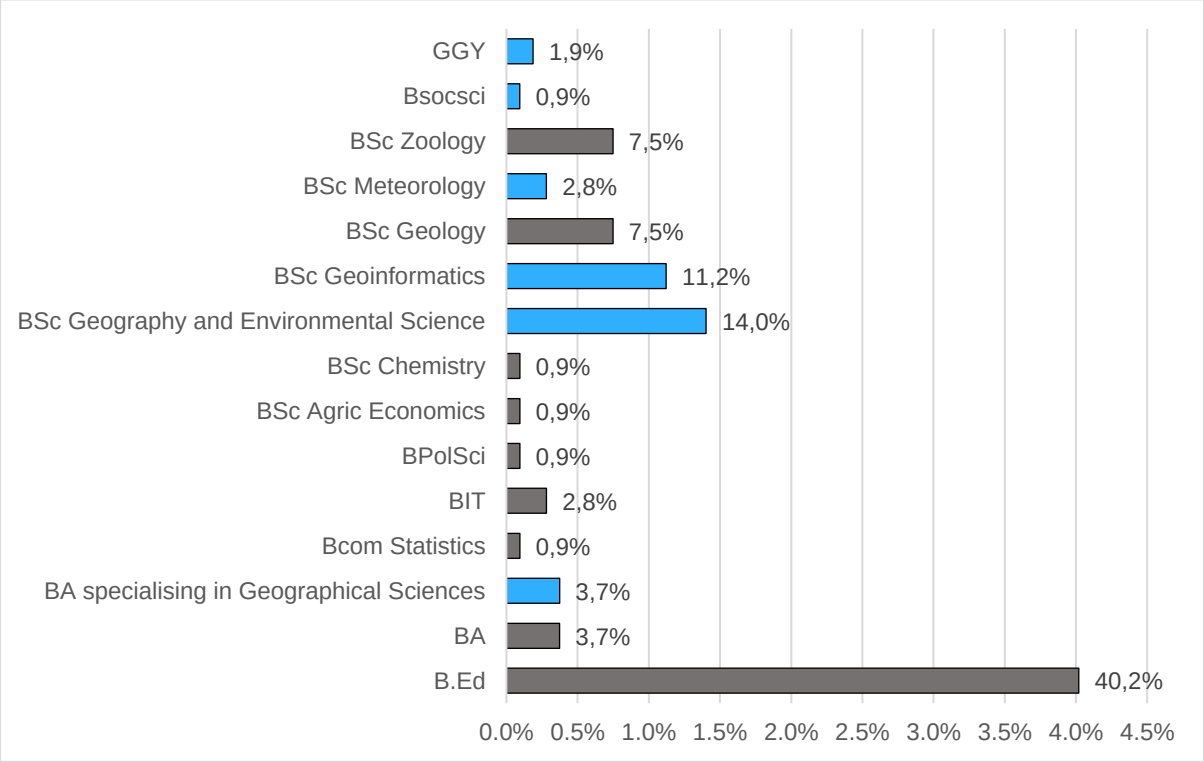


Figure 39: Percentage of students in each academic programme

The majority of respondents (97) took geography as a subject during their high school years. Of those, 42 (or 43.2%) were enrolled in the BEd academic programme.

Students were also required to indicate the modules they have already completed or are currently taking. Again, these figures provide better insights regarding respondents' overall skill levels. Table *13* shows the number of respondents who have completed or are currently taking a module, as well as the associated academic year of enrolment. Most participants were second-year students (78.5%), followed by third-year students (12.1%).

Table 13: Current academic year

Modules (Current or Completed)	Academic Year	No. of Participants
GMC110	1	7
GGY283	2	11
GGY283, GMC110	2	20
GGY283, GMC110, GIS221	2	1
GIS221	2	51
GMC110, GIS221	2	1
GGY283, GIS310, GGY383	3	1
GGY283, GIS310, GMC110	3	8
GGY383	3	2
GIS310	3	1
GIS310, GMC110	3	1
GGY283, GIS310, GGY383, GIS221, GIS708	4	1
GGY283, GIS310, GMC110, GIS708	4	1
GIS310, GMC110, GIS221, GIS708	4	1
Total		**107**

In the first section of the questionnaire, respondents were asked to indicate whether they had ever lived in any of the areas selected for the user study. They could select multiple locations. Since the University of Pretoria is situated in the City of Tshwane Metropolitan Municipality, it was assumed that all participants would have lived there. The aim of this question was simply to determine whether they were familiar with other localities besides Tshwane. As expected, 84% of participants currently reside in Tshwane or have previously done so. Of those, 15.9% indicated that they have also lived in Polokwane. Surprisingly, 14% indicated “None of the Above”. This may suggest that some respondents were unaware that Pretoria is part of the City of Tshwane Metropolitan Municipality, or that they did not read the question carefully. One person stayed in Buffalo City, while another stayed in Mangaung.

4.4 Results

This section evaluates the results obtained from the user study. The first part provides a broad overview of the findings based on respondents’ correct and incorrect answers (accuracy). The second part evaluates responses in reference to the four data classification methods. The objective and main purpose of this research were to determine whether certain data classification method(s) were easier or preferred for interpreting population demand with choropleth maps, specifically in relation to geographic accessibility. Finally, the accuracy of the responses was subsequently compared to various predictor variables defined for the user study to test their statistical significance.

Duration to Complete the Questionnaire

Qualtrics logs the time required to answer each question; as a result, the overall duration to complete the questionnaire was calculated for each respondent. Based on the duration, two extreme outliers were recorded: 1 473 minutes (24 hours) and 1 507 minutes (25 hours), respectively. The average time to complete the questionnaire was 62 minutes, excluding two outliers, while the median time was 41 minutes. Furthermore, the 20th and 80th percentiles were 29 and 60 minutes, respectively, while the 50th percentile equalled a duration of approximately 40 minutes, as shown in Figure *40*.

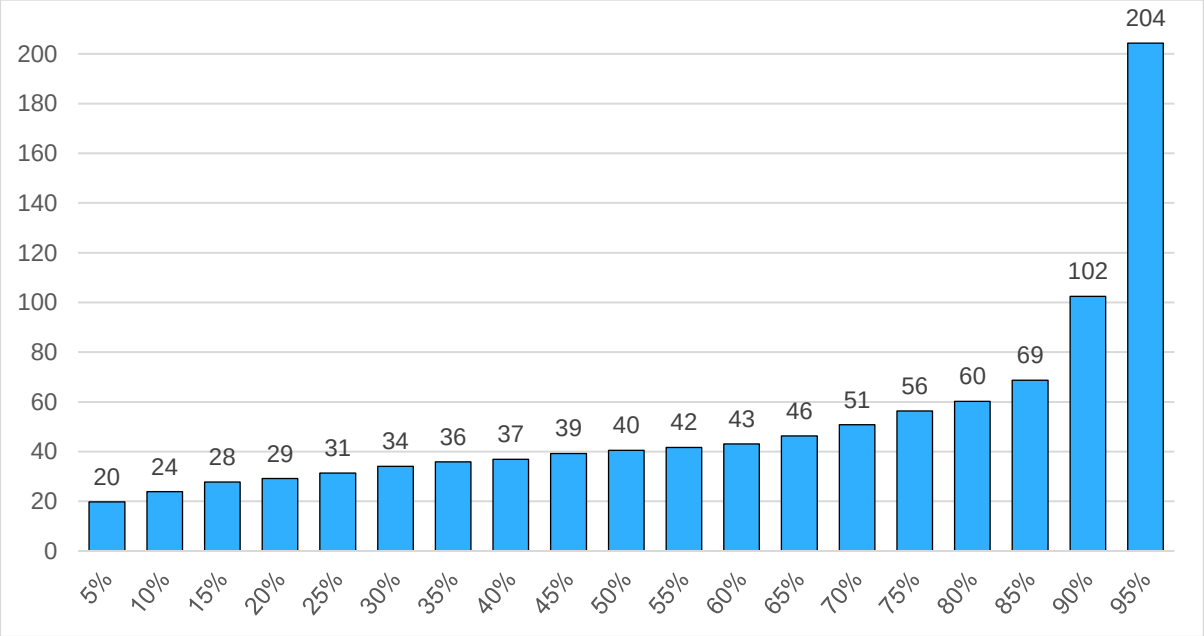


Figure 40: Duration in minutes to complete the entire questionnaire per 5th percentile

The average duration to answer one of the 48 geographic accessibility questions, which is the main component of the questionnaire, was 35 seconds with a median of 18 seconds. Each question comprised three components: (a) click events to identify location(s) on a map; (b) a rating of your confidence level in answering the specific question; and (c) a rating of your perceived difficulty level of the question.

Validating Responses

The click locations of the respondents were assessed, validated, and flagged as either correct or incorrect, with 1 indicating a correct response and 0 indicating an incorrect one. Unfortunately, it was not possible to design the accessibility questions in Qualtrics in a way that allowed responses (click events on the map) to be automatically assessed as correct or incorrect. Although a built-in function is available to add custom regions to a map image

(representing correct areas), it can only draw squares, which lack the granularity needed for evaluation purposes.

Eventually, a manual process in QGIS was used to evaluate the responses. Firstly, for each click event on the map in Qualtrics, the X and Y pixel values were captured, saved, and then imported into QGIS in CSV format. Additionally, the 48 map images used in Qualtrics were imported into QGIS, which allowed for an accurate overlay of the map images and the corresponding responses based on pixel values. For example, *Figure 41* shows the actual click events of all respondents for Question 2:

> *Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10 km, i.e. a centre serves only people within 10 km from it. Anybody living further away is not served by that centre.*

Points, displaying the click events, highlighted in red were manually flagged as incorrect following a visual inspection. The yellow dots indicate the correct locations. Questions were designed based on the DPSA's expansion, reduction, and relocation models (described in Chapter 2). Based on this example, respondents should ideally identify localities where the underlying polygon and neighbouring polygons are shaded in dark blue. Points located in yellow or light blue areas were flagged as incorrect.

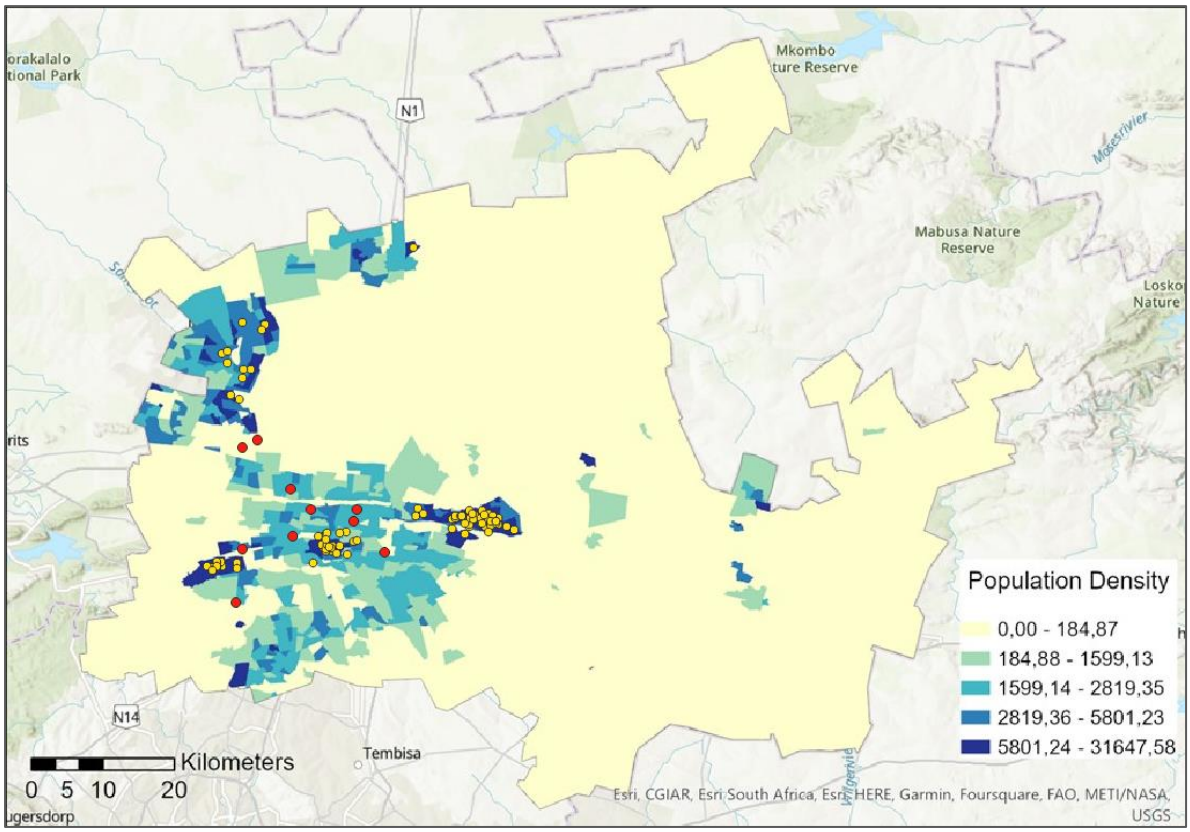


Figure 41: Choropleth map depicting correct and incorrect click events

4.4.1 Evaluating Choropleth Maps to Visualise Population Demand in South Africa

Overall, the respondents performed well in the user study, achieving an average accuracy rate of 89.9% on the 48 geographic accessibility questions. Four respondents scored 100%, while three individuals recorded the lowest accuracy level of 64.6%. The histogram (*Figure 42*) shows the number of respondents (frequency) by percentage accuracy, grouped into five percent intervals ranging from 65% to 100%. A calculated skewness of −1.213 indicates that the data are not normally distributed but are rather negatively skewed, which was expected given the high average percentage of accuracy.

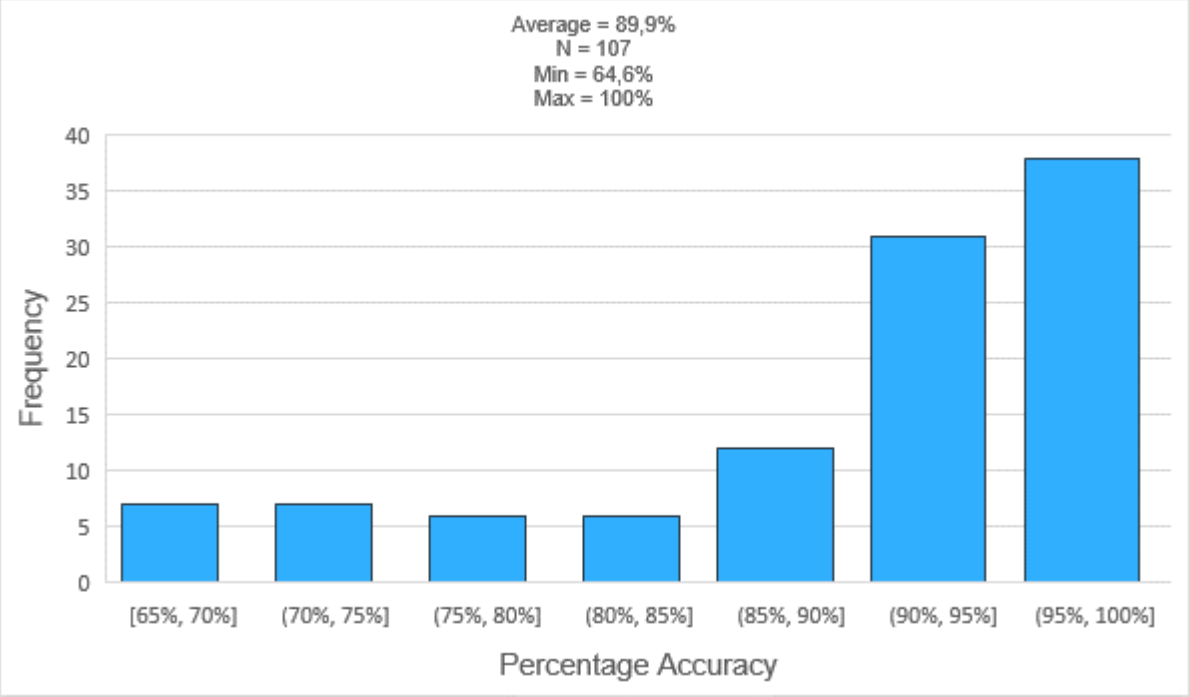


Figure 42: Histogram showing the overall percentage accuracy of respondents

4.4.1.1 Age and Gender

A comparison of percentage accuracy based on gender did not reveal any significant differences. Females performed slightly better, achieving a rate of 90.2%, followed by males at 89.4%. The participant who preferred not to indicate a gender scored 84.4%. Table *14* shows the percentage accuracy per age. On average, participants aged 18 answered most questions correctly at a rate of 96.3%. One participant, aged 27, achieved a 93.8% accuracy, followed by ages 22 and 23 with 91.4% and 90.5%, respectively. The lowest percentage accuracy, which was 88.7%, was calculated for 18 participants, all aged 19.

Table 14: Percentage accuracy by age

Age	No. of Respondents	% Accuracy
18	5	96.3%
19	18	88.7%
20	34	89.8%
21	27	89.9%
22	9	91.4%
23	6	90.5%
24	3	89.6%
25	2	76.0%
26	2	89.6%
27	1	93.8%

4.4.1.2 Education

Respondents with prior knowledge or experience in geography performed slightly better (92.0%) than those without previous experience (88.8%). Additionally, the frequency of map use does influence the percentage accuracy of responses. Students who indicated that they use maps once a week or every day performed well (91.8% and 90.3%) compared to those who use maps only once a month (87.8%) or rarely (86.2%).

Third-year students outperformed both second- and first-year students, achieving 93.9%, compared to 89.4% and 88.8%, respectively. Only three fourth-year students participated in the study. Two of them achieved a percentage accuracy above 98% and one scored 64.6%.

4.4.1.3 Map Literacy

The results from the two map literacy questions indicate that 91.5% of respondents answered the question about direction correctly. In contrast, only 53.2% of the 57 respondents answered the distance-related question correctly. A possible explanation is that respondents estimated the distance by eyeballing the scale bar for reference. Among these 57 participants, the average accuracy for the 48 geographic accessibility questions was 89.8%. Those who answered incorrectly had a very similar accuracy rate of 89.9%. This suggests that there is no conclusive relationship between the map literacy test, which consisted of two questions, and the accuracy of responses in the geographic accessibility questionnaire.

4.4.1.4 Vision Test

Ten respondents (9.3%) indicated that a professional had informed them that they have imperfect colour vision. Their average percentage accuracy was 87%. Of those, only two respondents failed the colour test in the questionnaire, meaning that although they have some imperfect colour vision, they can still differentiate colours. Also, results from the colour test

revealed that ten respondents incorrectly identified only one of the eight colours, while an additional four respondents made two mistakes. Based on these findings, it was decided that the results from these participants should not be excluded from the analysis.

4.4.1.5 Training and Expertise

To understand respondents' technical skills and proficiency, they were asked to rate their level of training or proficiency on a rating scale ranging from 0 to 10, where a score of 10 ranked highest. *Figure 43* shows the average rating for each respondent per category. Proficiency in English received the highest ranking of 8.9, followed by Google Maps with a ranking of 8.6. Cartography and statistics were ranked lowest, with scores of 5.8 and 5.7, respectively. The overall average rating for all categories combined was 7.1.

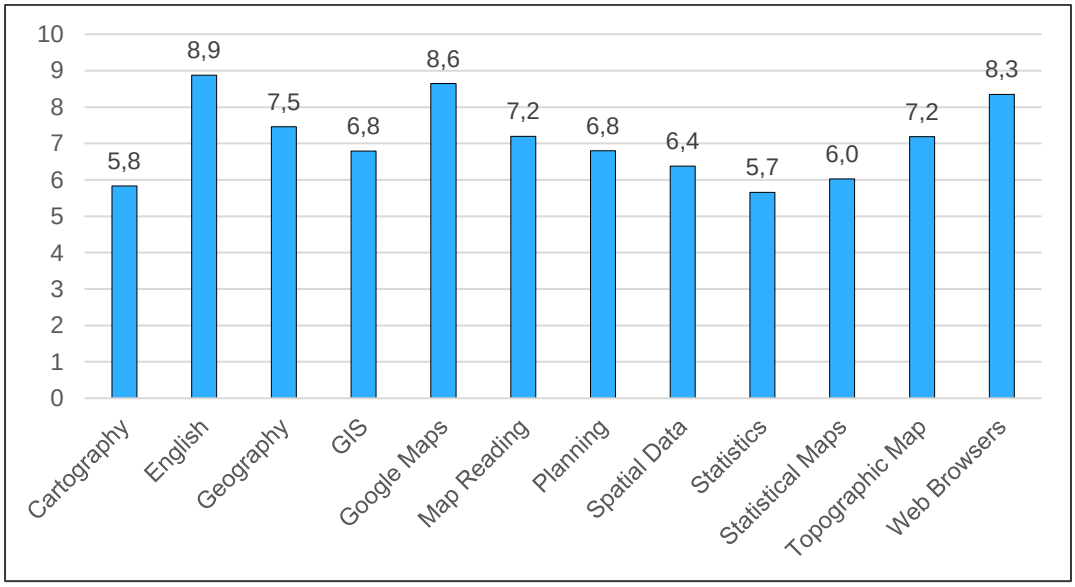
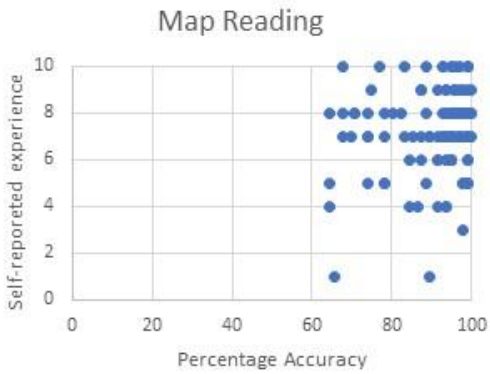
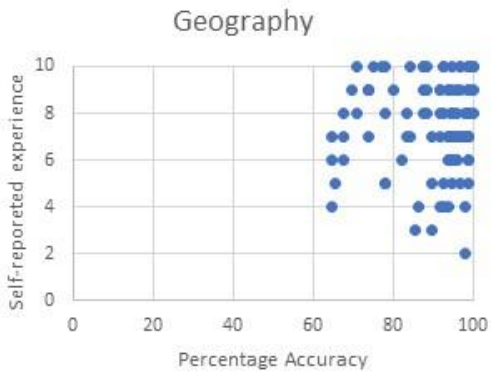


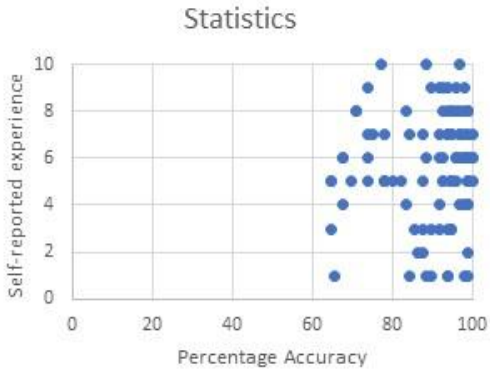
Figure 43: Average level of training (or expertise) per category

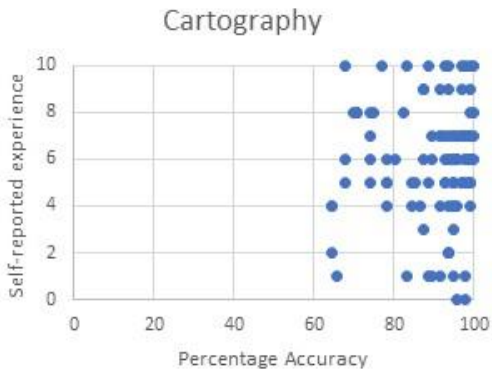
Initially, the researcher compared respondents' self-perceived levels of efficiency with their percentage accuracy scores to assess the strength of the linear relationship. The goal was to determine whether specific skills would increase the probability of respondents correctly interpreting choropleth maps.

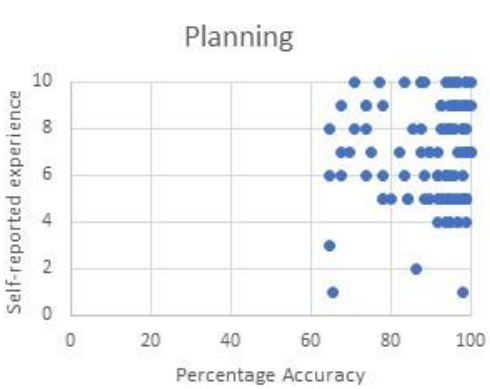
The scatterplots presented in *Figure 44* compare the percentage accuracy per respondent with their self-perceived level of expertise for each category separately. In general, neither the Pearson coefficient nor the scatterplots revealed a linear relationship between response accuracy and self-perceived experience levels.

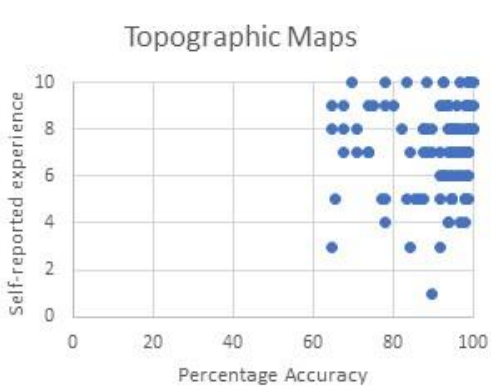


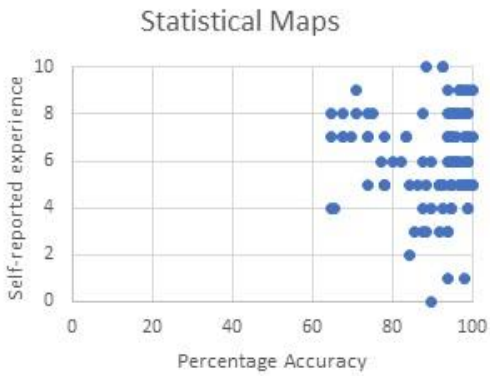


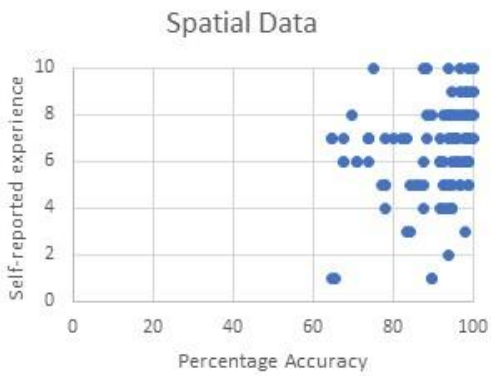












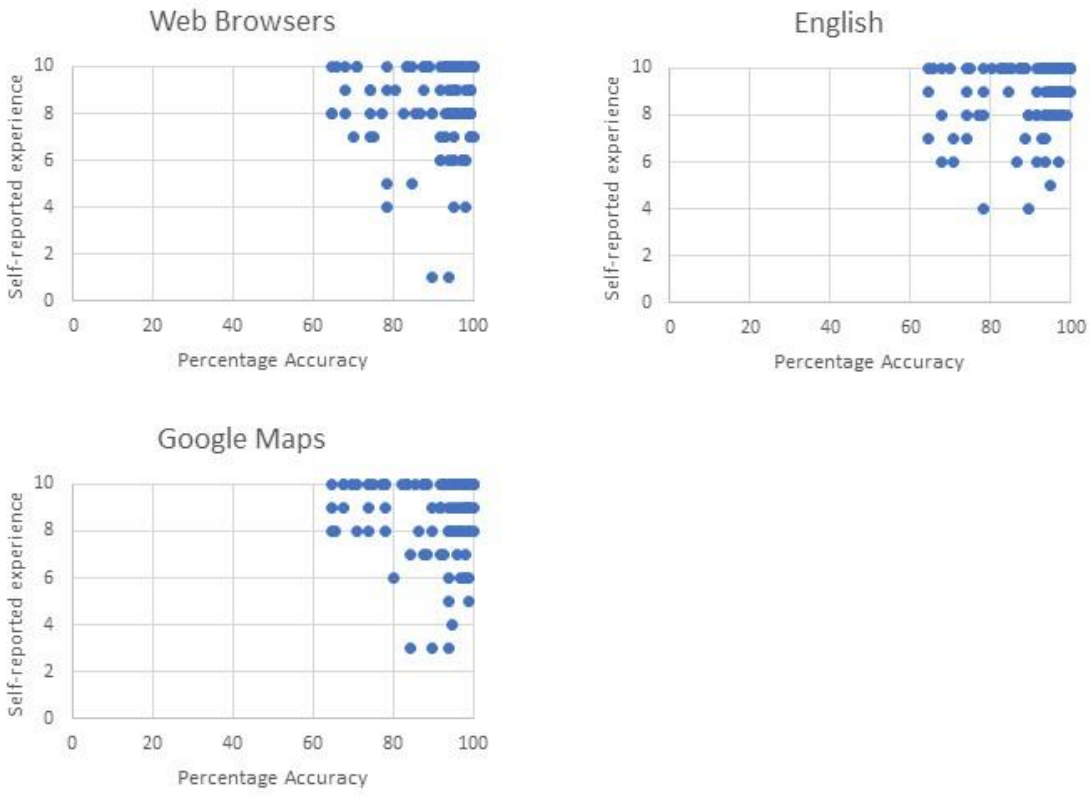


Figure 44: Linear relationship between respondents' percentage accuracy and self-perceived experience levels

Subsequently, a second test was conducted to determine potential relationships in the data by comparing the average accuracy score of respondents to each recorded level of efficiency (1–10) per category. As an example, Table *15* shows the results for self-perceived efficiency with spatial data only. Of the 107 respondents, seven rated their experience with spatial data as a 10. The average percentage accuracy was 91.5%, ranging from a minimum of 75% to a maximum of 100%.

Table 15: Accuracy score compared to respondents' self-perceived level of efficiency working with spatial data

Efficiency Level	Average % Accuracy	Min. % Accuracy	Max. % Accuracy	No. of Respondents
1	77.3	64.6	89.6	4
2	93.8	93.8	93.8	1
3	88.5	83.3	97.9	3
4	90.5	78.1	94.8	7
5	90.1	77.1	99.0	17
6	88.5	67.7	99.0	19
7	87.8	64.6	100.0	27
8	94.2	69.8	100.0	17
9	97.7	94.8	100.0	5
10	91.5	75.0	100.0	7
Total				**107**

Again, a Pearson product moment correlation coefficient (PPMCC) test was conducted to compare the average percentage accuracy of participants against each efficiency level, for all the different skills categories (Table *16*). The PPMCC strength of the relationships was grouped into four classes: strong, moderate, weak, and none, or very weak (adopted from Xiao et al. (2016).

Table 16: Linear relationship strength between accuracy scores and self-perceived level of efficiency

Category	Pearson Product Moment Correlation Coefficient	Strength
Map reading	0.435	Moderate
Geography	−0.242	Weak
Statistics	0.426	Moderate
Cartography	−0.139	Weak
GIS	−0.425	Moderate
Planning	0.443	Moderate
Topographic Maps	0.268	Weak
Statistical Maps	0.056	Weak
Spatial Data	0.591	Moderate
Web Browsers	−0.067	Weak
English as language	0.314	Moderate
Google Maps	−0.486	Moderate

A moderate linear relationship exists between the average percentage accuracy of respondents and their self-perceived level of expertise in the following categories:

- Map reading,
- Statistics,
- GIS,
- Planning,
- Spatial data,
- English, and
- Google Maps.

The strongest linear relationship was measured for spatial data (0.591), followed by Google Maps (−0.486, indicating a negative relationship) and statistics (0.426). Statistical maps were ranked lowest with a very weak Pearson product moment correlation coefficient of 0.056.

4.4.2 Comparison of Data Classification Methods

This section assesses the accuracy of responses, specifically the percentage of correct and incorrect answers among respondents, in relation to the four data classification methods. The first part includes a series of data tests to determine whether the data are fit for purpose and

statistically relevant for analysis. This is followed by a results section in which response accuracies were evaluated based on the four data classification methods.

Although the goal and primary objective of this research were to evaluate the suitability and effectiveness of different data classification methods for visualising population demand with choropleth maps, the researcher also sought to determine whether other variables are significantly associated with response accuracies. Hence, Section 4.4.3 focuses on response accuracies in relation to secondary predictor variables. Apart from the four data classification methods, other predictor variables include the type of geographic accessibility question, study areas, geographic units, respondents' perceived difficulty scores for each question, and confidence ratings.

4.4.2.1 Data Inspection

Shapiro–Wilk Test

Firstly, a Shapiro–Wilk test (Kwak & Park, 2019) was conducted to test for normality in the data. This test compared the percentage of correct answers given by each respondent to the 48 choropleth map questions in the questionnaire. The questions assessed respondents' ability to solve one of the four geographic accessibility problems and depict a specific data classification method, geographic unit, and study area. Different study areas and geographic units were presented to limit a possible learning effect while respondents completed the survey.

The purpose of the test was to determine whether the data were parametric (normally distributed) or nonparametric (skewed). Which, in turn, determined the appropriate statistical methods for further comparative analyses. Results from the Shapiro–Wilk test revealed a not normal distribution (sig. < 0.001) for all the questions except for Question 38 (see Appendix B). This implies a non-normal distribution for the predictor variables.

Kruskal–Wallis Test

Secondly, a Kruskal–Wallis test was conducted to determine whether there were statistically significant differences between the accuracy scores and the four data classification methods. This test is usually performed on nonparametric data. The results from the Kruskal–Wallis test (*Figure 45*) confirmed that there are statistically significant differences in accuracy scores among the four data classification methods, p-value (< 0.001).

Test Statistics[a,b]

	Percentage Accuracy
Kruskal-Wallis H	22,239
df	3
Asymp. Sig.	<,001

a. Kruskal Wallis Test
b. Grouping Variable: Class_NR

Figure 45: Kruskal–Wallis test

Friedman Test

Unfortunately, although a significance score of less than 0.001 was obtained based on the Kruskal–Wallis test, the questionnaire was designed so that all 107 respondents answered questions related to each of the four data classification methods. This violates the third assumption of the Kruskal–Wallis test: "You should have independence of observations" (Laerd Statistics, n.d.). This means that there should be different participants for each group, as defined by the data classification method used in this research.

Hence, the Friedman test was done (*Figure 46*) to test for differences among the four data classification methods. This test is also recommended for nonparametric data distributions. The statistics table confirms a significance score of less than 0.001, indicating that there are statistically significant differences between the data classification methods.

Test Statistics[a]

N	107
Chi-Square	35,817
df	3
Asymp. Sig.	<,001

a. Friedman Test

Figure 46: Friedman test

4.4.2.2 Response Accuracy of the Four Data Classification Methods

Although the percentage accuracy for each data classification method was notably high, respondents were more likely to provide correct answers for the maps depicting the quantiles data classification method (92.3%). This was followed by natural breaks (Jenks) and geometric interval with 91.2% and 88.8%, respectively. The logarithmic scale was ranked lowest, with an accuracy of 87.2%. Descriptive statistics (Table *17*) highlights the data distribution, specifically the percentage accuracy of participants, for each data classification method.

The median percentage accuracy for both quantiles and natural breaks (Jenks) was 95.8%, while it was 91.7% for the geometric interval and logarithmic scale.

Table 17: Descriptive statistics based on responses per data classification method

Geometric interval	Mean	88.82%
	Median	91.67%
	Minimum	54.2%
	Maximum	100.0%
	Skewness	−0.938
	Kurtosis	−0.148
Logarithmic scale	Mean	87.19%
	Median	91.67%
	Minimum	62.5%
	Maximum	100.0%
	Skewness	−0.682
	Kurtosis	−0.660
Natural breaks (Jenks)	Mean	91.24%
	Median	95.83%
	Minimum	58.3%
	Maximum	100.0%
	Skewness	−1.319
	Kurtosis	0.777
Quantiles	Mean	92.29%
	Median	95.83%
	Minimum	50.0%
	Maximum	100.0%
	Skewness	−1.947
	Kurtosis	3.280

The box plot in *Figure 47* shows the percentage accuracy per respondent for the four data classification methods. Quantiles and natural breaks (Jenks) exhibit less variation, albeit with fewer outliers (mild outliers are marked with black dots and extreme outliers are marked with stars) than the other two methods.

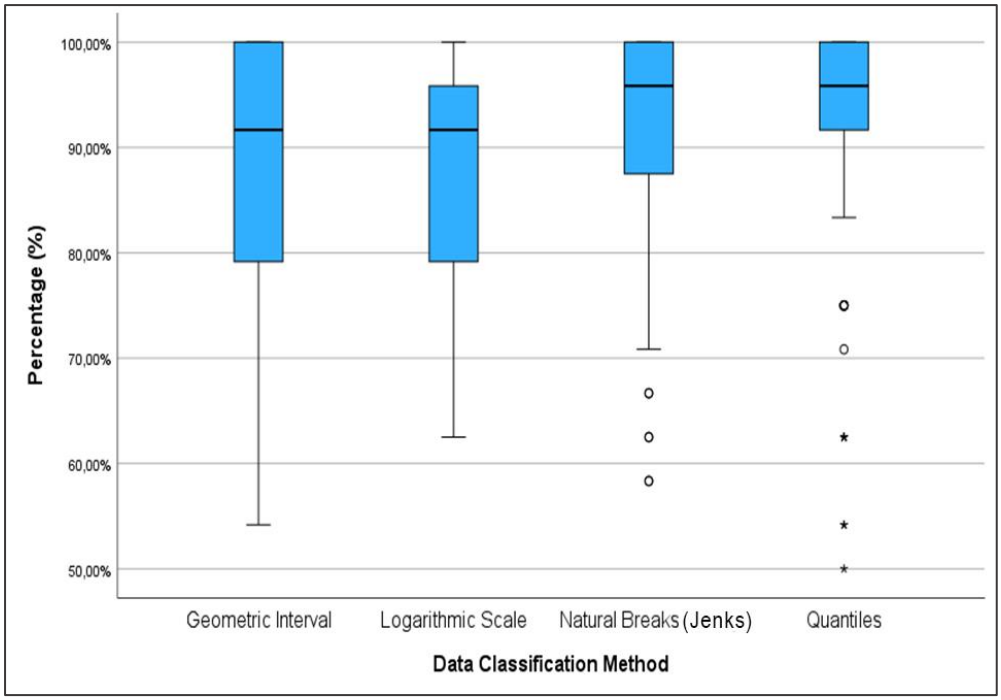


Figure 47: Accuracy score per data classification method

The skewness and kurtosis of the data classification methods showed significant differences. The skewness for both natural breaks (Jenks) and quantiles was below −1, with values of −1.319 and −1.947, indicating a strongly negatively skewed distribution. The geometric interval and logarithmic scale had a skewness of −0.938 and −0.682, respectively, depicting a slightly more normal distribution. See *Figure 48*.

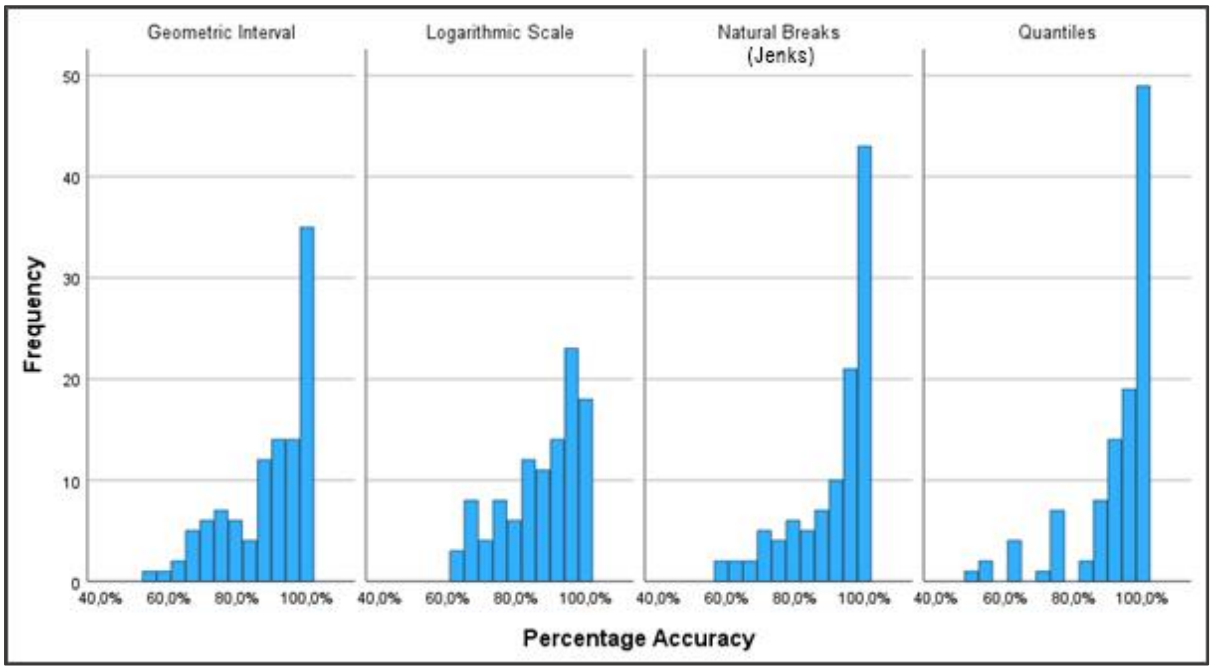


Figure 48: Histogram showing data distribution per data classification method

4.4.3 Significance of Predictor Variables

This section focuses on response accuracies relative to secondary predictor variables. Although the goal and main objective of this research were to evaluate the suitability and effectiveness of different data classification methods to visualise population demand with choropleth maps, the researcher also wanted to test whether secondary predictor variables, as defined in the user study design, could be significantly associated with response accuracies in general, as well as for each data classification method individually. These variables include:

- Question – the four geographic accessibility questions,
- Location – the four study areas,
- GeoUnit – the defined geographic units (hexagons, small area layer, and sub-places),
- Confidence – the confidence level in answering a specific question, and
- Level – the rating of the perceived difficulty level of a question.

A logistic regression analysis was used to test the significance of all predictor variables against a single dichotomous dependent variable, which represented respondents' answers as either correct or incorrect responses. In this question, 1 indicated a correct response, while 0 indicated an incorrect response. These predictor variables were tested against the entire data set with all responses combined, as well as individually for each data classification method. A logistic regression analysis is frequently used when the predicted outcome is binary, for example on/off or pass/fail (Healy, 2006). In this case, it was applied to distinguish between respondents' correct and incorrect answers. Calculations were performed in RStudio using the car (companion to applied regression) library. For the first iteration, which used the entire data set, a statistically significant score of less than 0.001 was achieved for the following predictor variables (Table *18*):

- Question – the four geographic accessibility questions,
- Location – the study areas, and
- Level – the rating of perceived difficulty level of a question.

R script

```
Model <- glm(Answer ~ Question +
         Location + GeoUnit + Confidence + Level, family = "binomial",data = Survey)
Anova(Model)
```

Table 18: Significance score of predictor variables

```
Analysis of Deviance Table (Type II tests)

Response: Answer
           LR Chisq Df Pr(>Chisq)
Question    278.397  3  < 2.2e-16 ***
Location     28.662  3  2.637e-06 ***
GeoUnit       0.245  2    0.88491
Confidence   20.801 10    0.02253 *
Level        75.865  1  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the second iteration, four subsets of the data were created, each representing responses for a specific data classification method. Additionally, response accuracies were tested against the defined predictor variables. Table *19* shows the significance scores per data classification method for each variable.

Overall, for the entire data set and for each data classification method, geographic accessibility questions were found to be a statistically significant predictor (sig. < 0.001). However, some variation exists between the individual data classification methods and the remaining predictor variables.

The perceived difficulty level of a question was significant (sig. < 0.001) for all data classification methods, except for quantiles. Furthermore, location (study areas) was only significant for the geometric interval and logarithmic scale. Geographic units and confidence were not statistically significant. Only geographic accessibility questions proved to be significant for responses based on the quantiles data classification method.

Table 19: Significance score of predictor variables for each data classification method

Geometric Interval

```
Analysis of Deviance Table (Type II tests)

Response: Answer
           LR Chisq Df Pr(>Chisq)
Question     94.615  3  < 2.2e-16 ***
Location     16.942  3  0.0007266 ***
GeoUnit       1.022  2  0.5999761
Confidence    8.738 10  0.5571284
Level        53.373  1  2.759e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logarithmic Scale

```
Analysis of Deviance Table (Type II tests)

Response: Answer
           LR Chisq Df Pr(>Chisq)
Question     41.045  3  6.396e-09 ***
Location     34.626  3  1.461e-07 ***
GeoUnit       3.915  2  0.1412222
Confidence   13.622 10  0.1909619
Level        11.383  1  0.0007413 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Natural Breaks (Jenks)

```
Analysis of Deviance Table (Type II tests)

Response: Answer
           LR Chisq Df Pr(>Chisq)
Question     94.134  3  < 2.2e-16 ***
Location      6.740  3    0.08066 .
GeoUnit      13.624  2    0.00110 **
Confidence   18.549 10    0.04637 *
Level        21.740  1  3.123e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Quantiles

```
Response: Answer
           LR Chisq Df Pr(>Chisq)
Question     61.623  3  2.645e-13 ***
Location      5.531  3    0.13680
GeoUnit       4.771  2    0.09205 .
Confidence   11.461  9    0.24546
Level         0.996  1    0.31821
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The following section compares the accuracy of both correct and incorrect responses with the four predictor variables, examining the results overall and for each data classification method individually.

4.4.3.1 Geographic Accessibility Questions

As described in the study design (Section 4.2), four geographic accessibility questions were designed based on real-world scenarios. In theory, these questions also varied in increasing levels of difficulty. Question 1 was the easiest, while Question 4 was considered the most difficult. The results confirmed our theory (Table *20*). Most participants answered Question 1 correctly (95.3%), followed by Questions 2, 3, and 4 with 94.0%, 88.2%, and 81.5%, respectively. This was also true for each data classification method except quantiles, where respondents performed slightly better on Question 2 than Question 1 (96.6% vs 95.0%).

Based on these results and the fact that geographic accessibility questions proved to be statistically significant for all data classification methods, the target audience, including decision makers and map users, should not only be familiar with choropleth maps in a real-world scenario, but they must also understand the principles of geographic accessibility analysis, as described in Chapter 2, to effectively study and interpret choropleth maps that depict population demand.

Table 20: Percentage accuracy per geographic accessibility question type and data classification method

Question	Geometric Interval	Logarithmic Scale	Natural Breaks (Jenks)	Quantiles	Overall Average
Q1	96.4%	91.9%	97.8%	95.0%	95.3%
Q2	92.8%	91.3%	95.3%	96.6%	94.0%
Q3	84.9%	85.0%	88.0%	94.7%	88.2%
Q4	79.4%	80.2%	82.6%	83.6%	81.5%

4.4.3.2 Study Areas

Study areas proved to be statistically significant overall, as well as for maps depicting geometric interval and logarithmic scale classification methods. Respondents provided the most correct answers for choropleth maps showing the Mangaung Metropolitan Municipality (92.2%). Upon visual inspection, the municipality appears to have large, sparsely populated areas, with only a few high-density locations. The City of Tshwane Metropolitan Municipality ranked second, followed closely by the Buffalo City Metropolitan Municipality, which achieved scores of 89.9% and 89.5%, respectively. The most incorrect answers were for the Polokwane Local Municipality, with 87.9% (*Figure 49*).

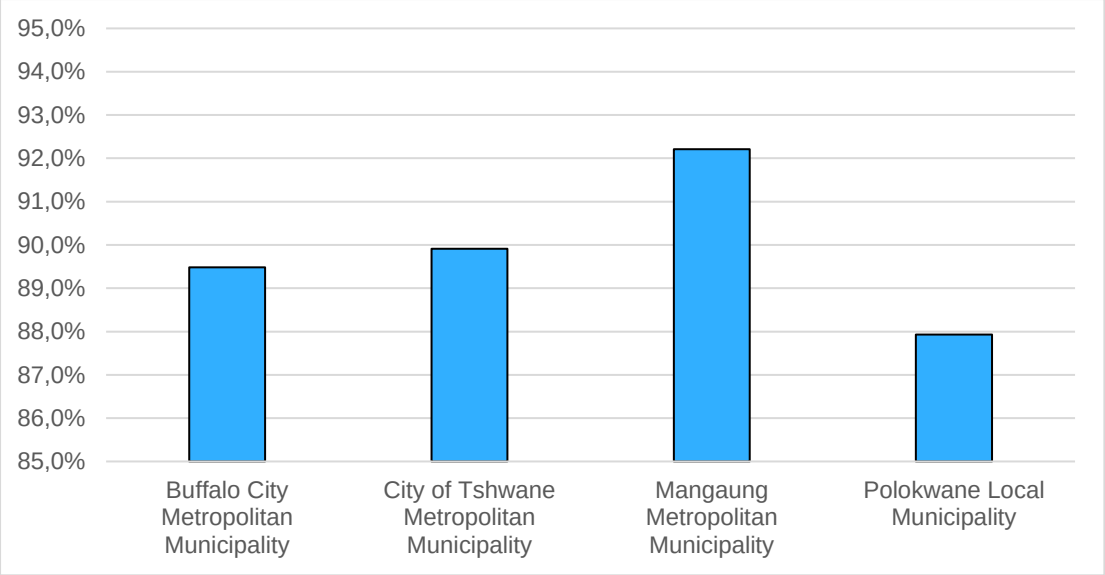


Figure 49: Percentage accuracy per study area

Table *21* shows the percentage accuracy per study area for maps depicting the four data classification methods. The table highlights variations in response accuracy between study areas and data classification methods. Hence, study areas influence respondents' interpretation of choropleth maps. For maps showing the Buffalo City Metropolitan Municipality, respondents provided the most accurate answers for the quantiles data classification method, followed by the geometric interval method. Results for the City of Tshwane Metropolitan Municipality are notably different. Most correct answers were based on natural breaks (Jenks) and logarithmic scale. Quantiles and natural breaks (Jenks) were best interpreted for maps showing the Mangaung Metropolitan Municipality, with natural breaks (Jenks) being the highest for the Polokwane Local Municipality, followed by quantiles.

Table 21: Percentage accuracy per study area and data classification method

Study Area	Geometric Interval	Logarithmic Scale	Natural Breaks (Jenks)	Quantiles
Buffalo City Metropolitan Municipality	92.1%	85.4%	86.2%	93.2%
City of Tshwane Metropolitan Municipality	85.2%	91.3%	94.1%	89.9%
Mangaung Metropolitan Municipality	88.8%	91.1%	93.3%	95.3%
Polokwane Local Municipality	89.9%	79.1%	90.7%	90.2%

4.4.3.3 Geographic Units

Geographic units were not identified as a statistically significant predictor; neither overall nor for any of the four data classification methods. Respondents provided more correct answers (90%) for maps showing population density by sub-place, followed by hexagon and small area layers, with 89.9% and 89.7% correct answers, respectively.

Table *22* shows the percentage accuracy per data classification method and geographic unit. Quantiles scored the highest percentage accuracy for both hexagons and small area layers, with scores of 93.7% and 91.9%, respectively. On the sub-place level, natural breaks (Jenks) was first, followed by quantiles. Natural breaks (Jenks) was also the second-highest for maps showing hexagons and third based on small area layer polygons. Although geographic units were not identified as a statistically significant predictor, the results suggest some variation in response accuracies between the four data classification methods and geographic units.

Table 22: Percentage accuracy by geographic unit and data classification method

Data Classification Method	Hexagon	Small Area Layer	Sub-Place
Geometric interval	88.4%	88.3%	89.7%
Logarithmic scale	84.1%	89.8%	87.6%
Natural breaks (Jenks)	93.3%	88.8%	91.6%
Quantiles	93.7%	91.9%	91.2%

4.4.3.4 Self-perceived Confidence Level

Self-perceived confidence level was also not identified as a statistically significant predictor. For each of the 48 map questions, respondents were asked to rate their confidence in their answers on a scale from 0 to 10, with 10 indicating complete confidence. The average and median confidence level per question was 8 overall, as well as for each data classification method individually, except for the logarithmic scale, which had an average confidence level of 7. Additionally, minimum and maximum confidence levels of 0 and 10 were recorded for all data classification methods.

4.4.3.5 Self-Perceived Difficulty Rate

A significant relationship was found between the percentage accuracy of respondents and their self-perceived difficulty rate per question for all data classification methods, except for quantiles (Table *23*). Respondents were asked to indicate a difficulty level per question ranging from very easy, easy, neutral, difficult, to very difficult. Questions rated as very easy would likely have the highest percentage of accuracy, followed by a gradual decrease, with questions rated as very difficult having the lowest percentage of accuracy. Results generally followed this pattern except for questions rated as very difficult. A percentage accuracy of 95% was calculated for questions considered to be very easy. This was followed by easy, neutral, and difficult, with 92.1%, 85.6%, and 80.4%, respectively. The percentage accuracy of questions that respondents considered very difficult was, however, higher at 83.8% compared to those considered difficult (80.4%).

Table 23: Percentage accuracy and self-perceived difficulty rate

Rate	% Accuracy
Very Easy	95.0%
Easy	92.1%
Neutral	85.6%
Difficult	80.4%
Very Difficult	83.8%

Figure 50 shows the percentage accuracy of responses per data classification method. A similar pattern was observed in which the percentage accuracy gradually decreased with each perceived difficulty level, ranging from very easy to very difficult. The overall percentage accuracy for quantiles was very high, regardless of the perceived difficulty level. The lowest percentage accuracy was recorded for the difficult and very difficult levels based on the geometric interval data classification method, with 69% and 67%, respectively. The highest percentage of accuracy (97%) was calculated for natural breaks (Jenks) when the questions were considered to be very easy.

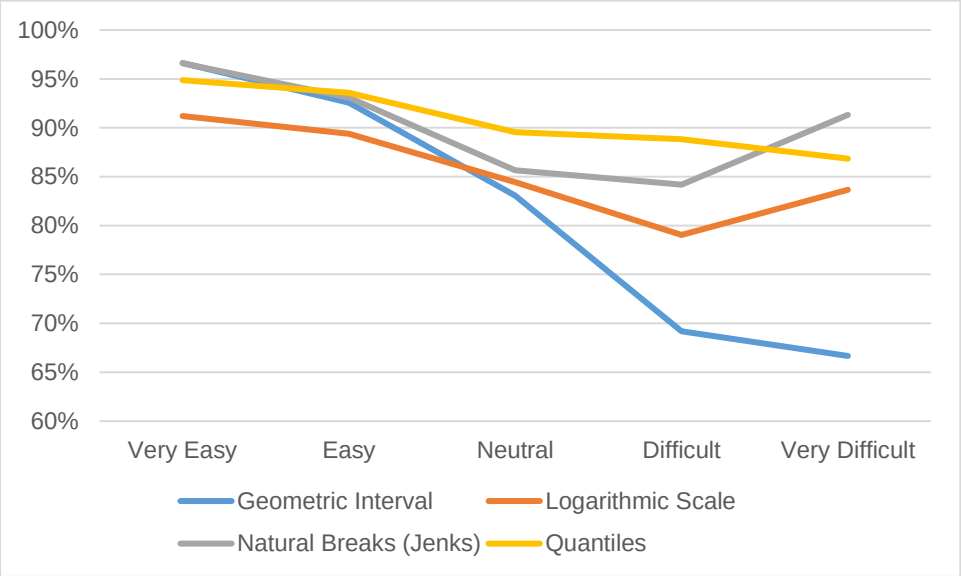


Figure 50: Percentage accuracy and self-perceived difficulty rate per data classification method

4.5 Discussion

Four key topics emerged from the findings of this chapter that require further discussion. The first is the use of choropleth maps to visualise population demand in South Africa. The second area of discussion focuses on the respondents' interpretation of the various data classification methods for choropleth maps. The results from the user study revealed that certain methods were easier or preferred for interpreting population demand compared than others. The third area describes the demographic, academic, and technical efficiency characteristics of the respondents. The final part of the discussion focuses on various predictor variables, as defined

in the user study design, which were significantly associated with response accuracies, including both correct and incorrect answers from respondents.

Choropleth Maps

Choropleth maps are “One of the oldest, and still one of the most frequently used techniques to visualise quantitative data in a GIS” (Tyner, 2014). This is also true in South Africa. Choropleth maps are frequently used by the IEC to visualise voting and registration patterns. The IEC’s (2019) Atlas of Results allows users to analyse political party support using pregenerated choropleth maps for different geographic units, including voting districts, wards, and municipalities. Statistics South Africa (2023b) uses choropleth maps to display population distribution and growth patterns across the country. Other examples include a study by Public Health that used choropleth maps to analyse cervical cancer screening (Makura et al., 2016), and examine the relationship between population distribution and economic activity across the Gauteng city region (Mosiane & Murray, 2021). In a study conducted by the University of Pretoria, choropleth maps were used to display both isochrones (travel time to cath lab facilities) and population density by ward. The purpose of the study was to identify populated areas across South Africa where access to cath labs exceeds specified travel-time thresholds (Coetzee et al., 2021).

To design a choropleth map, information about an area is grouped or categorised into classes and visually represented using symbols, such as the shading of colour (De Smith et al., 2018). Slocum et al. (2014) noted that a choropleth map is “the most commonly used (and abused) method of thematic mapping”. One disadvantage of choropleth maps for visualising statistical data is the tendency to “overemphasize large, yet often sparsely populated, administrative areas because of their strong visual weight” (Besançon et al., 2020). In South Africa, this is also true, as population data shows that rural areas occupy a larger geographic space than smaller, densely populated urban areas (Harris et al., 2017). The four study areas identified for this study comprise a variety of settlement typologies, including sparsely populated rural areas. The analysis confirmed that the population is not normally distributed in any of the four selected study areas; instead, it is highly skewed.

Comparison of Data Classification Methods

In this research, it was essential to first assess whether choropleth maps are useful for visualising population distribution in South Africa, and secondly, whether choropleth maps are useful to identify over- and underserved areas for geographic accessibility analysis. The results from this chapter suggest that choropleth maps are indeed an effective and easy-to-use technique for visualising population demand. The average accuracy percentage for the

107 participants who completed the online questionnaire, based on the 48 geographic accessibility questions, was 90%. Four respondents scored 100%, and the lowest recorded accuracy was 65% among three individuals.

In another map reading experiment regarding the interpretation of COVID-19 data, choropleth map visualisations were compared to graduated symbols. The results from an online questionnaire revealed that choropleth maps were much easier to read than graduated symbols (Sukraini et al., 2022). These results, along with the high percentage of accuracy achieved in the user study, likely explain why choropleth maps are popular and widely used.

Before comparing respondents' interpretations of choropleth maps depicting specific data classification methods, the researcher wanted to test for differences within groups, specifically the percentage accuracy of responses for each of the four data classification methods. Results confirmed a significance score of less than 0.001, suggesting that statistically significant differences exist between groups. The percentage accuracy of responses for each of the four data classification methods was notably high. Respondents were, however, more likely to provide correct answers for maps depicting the quantiles data classification method (92%), followed by natural breaks (Jenks) and geometric interval, with 91.2% and 88.8% accuracy, respectively. The logarithmic scale was ranked the lowest, with an accuracy of 87.2%. In a previous study, Brewer and Pickle (2002) assessed seven data classification methods for classifying epidemiological data using choropleth maps. These methods included hybrid equal intervals, quantiles, box plots, standard deviation, natural breaks (Jenks), minimum boundary error, and shared area. Their map interpretation questions were designed in such a way that certain question types were more difficult than others. The findings from their study also indicate that quantiles, followed by minimum boundary error classification methods, provided the most accurate results, which participants best interpreted. This was followed by natural breaks (Jenks) and a hybrid version of equal intervals (Brewer & Pickle, 2002). These findings could encourage future research to better understand which data classification methods work best for different data types and visualisation tasks.

Demographic, Academic, and Technical Efficiency Characteristics of Respondents

The literature widely supports the notion of a male advantage in spatial abilities. For example, a general map reading skills test conducted by Albert et al. (2016), which focused on orientation, distance, topographic elements, geographic names, map symbols, and hypsography, showed that map reading competency varied based on gender, nationality, and age group. A significantly higher percentage of males provided correct answers regarding the interpretation of hypsography compared to females. In another example related to spatial

tasks, the slowest mean reaction time and lowest mean accuracy were recorded for females (Lloyd & Bunch, 2008).

For this study, however, a comparison of percentage accuracy based on gender did not reveal any significant differences. Females performed slightly better, with a rate of 90%, followed by males at 89%. It is also worth noting that of the 107 respondents, 64% were female, while 35% were male; hence, the gender distribution was unequal. One student preferred not to indicate a specific gender. Third-year students outperformed second- and first-year students, achieving a rate of 94%, compared to 89.4% and 88.8%, respectively. Only three fourth-year students participated in the study. Two of them achieved a percentage accuracy above 98% and one scored 64.6%.

A moderate linear relationship exists between the average percentage accuracy of respondents and their self-perceived level of expertise in the following categories: map reading, statistics, GIS, planning, spatial data, English, and Google Maps. The strongest linear relationship was measured for spatial data (0.591), followed by Google Maps (−0.486, indicating a negative relationship) and statistics (0.426). Statistical maps were ranked lowest with a very weak Pearson product moment correlation coefficient of 0.056, which is ironic since choropleth maps are indeed statistical maps. The average percentage accuracy for the user study was 90%. Hegarty et al. (2002) noted that although evidence suggests that people tend to “overestimate their abilities” in areas such as logic, grammar, and humour, observations from their study show that people are “somewhat truthful and accurate in estimating their environmental spatial abilities”. However, a map literacy experiment by Rautenbach et al. (2014) highlighted that there was no significant correlation between participants’ self-rated level of experience and their performance accuracy. Further research on the subject could include experiments to confirm whether these relationships exist factually.

Significance of Predictor Variables

The last part of the discussion focuses on predictor variables that were significantly associated with response accuracies in general, as well as each with data classification method individually. These predictor variables include the four geographic accessibility questions, the four study areas, defined geographic units (hexagons, small area layer, and sub-places), participants’ confidence levels in answering a specific question, and their perceived difficulty levels of a question.

Overall, for each data classification method, the geographic accessibility questions proved to be a statistically significant predictor (sig. < 0.001). Geographic accessibility questions were designed based on real-world scenarios and varied in increasing levels of difficulty. Question 1

was the easiest, while Question 4 was considered the most difficult. The results from the user study show that most participants answered Question 1 correctly (95%), followed by Questions 2, 3, and 4, with 94%, 88%, and 82% accuracy, respectively. This is also true for each of the four data classification methods, except for quantiles, where respondents performed slightly better on Question 2 than Question 1 (97% versus 95%). Based on these results and the fact that geographic accessibility questions proved to be statistically significant for all the data classification methods, the target audience, decision maker, or map user in a real-world scenario should not only be familiar with choropleth maps; they must also understand the principles of geographic accessibility analysis, as described in Chapter 2, to effectively study and interpret choropleth maps depicting population demand.

The four study areas were found to be significant overall, as well as when using both geometric interval and logarithmic scale data classification methods. It is worth mentioning that the study areas were not statistically significant predictors for the quantiles and natural breaks (Jenks) methods. However, respondents preferred these methods for interpreting population demand compared to the other two methods.

Overall, respondents provided the most correct answers for choropleth maps depicting the Mangaung Metropolitan Municipality (92%). Based on the visual inspection, the municipality is characterised by large, sparsely populated areas with only a few high-density locations, making it easier for respondents to identify the correct locations, regardless of the data classification method used. The City of Tshwane Metropolitan Municipality was second, followed by Buffalo City Metropolitan Municipality which achieved scores of 89.9% and 89.5%, respectively. The most incorrect answers were for the Polokwane Local Municipality, with 87.9%. One would expect respondents to achieve the highest accuracy percentage for maps depicting the City of Tshwane Metropolitan Municipality since the University of Pretoria is located within the metro. Therefore, the respondents would be familiar with the area (Boscoe & Pickle, 2003). Although a high percentage of accuracy was recorded for all study areas, these findings indicate that different data classification methods are sensitive to the selection of study area. Another possibility to consider is that the maps (study areas) were not presented to respondents in a consistent order, which may have helped prevent a potential learning effect.

The size and shape of areas or polygons, also referred to as geographic units, were identified by Robinson (1995) as important elements for choropleth maps. Boscoe and Pickle (2003) further defined the ideal characteristics of geographic units, among others, as: “a high degree of resolution, homogeneity of population size and land area, minimum population thresholds and land area thresholds, compactness of shape, audience familiarity and functional

relevance". Based on the results of the user study, geographic units were not found to be statistically significant predictors, either overall or in relation to any of the four data classification methods. Quantiles scored the highest percentage accuracies for both hexagon and small area layer, with scores of 93.7% and 91.9%, respectively. At the sub-place level, natural breaks (Jenks) was first, followed by quantiles. Natural breaks (Jenks) was also the second-highest for maps showing hexagons and third for small area layer polygons. Although geographic units were not identified as a statistically significant predictor, the results suggest some variation in response accuracies between the four data classification methods and geographic units, indicating a need for further research on the effective use of geographic units.

There was a significant relationship between the percentage accuracy of respondents and their self-perceived difficulty rate per question based on all the data classification methods except for quantiles. For each of the 48 map questions, respondents were asked to indicate a difficulty level ranging from very easy, easy, neutral, difficult, to very difficult. Questions rated as very easy generally had the highest percentage accuracy, followed by a gradual decrease, with questions rated as very difficult having the lowest percentage accuracy. The results generally followed this pattern, except for questions rated as very difficult. A percentage accuracy of 95% was calculated for questions considered very easy. This was followed by easy, neutral, and difficult, with 92.1%, 85.6%, and 80.4%, respectively. The percentage accuracy of questions that respondents considered to be very difficult was, however, higher at 83.8% compared to those considered to be difficult (80.4%). The results confirm what Rautenbach et al. (2017) found when developing and evaluating a task taxonomy for spatial planning through a map literacy experiment using topographic maps; namely, that the accuracy score of participants correlated with their self-perceived difficulty level of a question.

5. ERROR CALCULATION OF CLASS BREAKS

5.1 Introduction

The results from the user study discussed in Chapter 4 show that respondents were more likely to provide correct answers for maps depicting the quantiles data classification method, followed by the natural breaks (Jenks) method, and then the geometric interval method. The logarithmic scale was the least preferred method.

Chapter 5 continues to evaluate the suitability of these four data classification methods by means of a mathematical equation. The equation calculates the error between class breaks, as described in Chapter 2. By doing so, the accuracy of the data classification method for a specific data set was measured.

The purpose of this chapter was simply to compare the human interpretation (target audience) of the data classification methods for choropleth maps (user study) with a recommended mathematical equation, which in theory, could be considered less subjective, to highlight potential similarities and differences derived from the two approaches. Subsequently, a more comprehensive understanding of the application of data classification methods was obtained, specifically within a South African context where the population is not normally distributed.

For an accuracy assessment of data classification methods, the goodness of variance fit (GVF) measurement is considered a suitable (Chandra & Mistri, 2011) and optimal measurement technique (Declerq, 1995, Smith, 1986) to measure the error between class breaks. The technique is also frequently described in the literature (Armstrong et al., 2003; Chandra & Mistri, 2011; Declerq, 1995; Golian et al., 2010; Robinson, 1984; Slocum et al., 2014; Smith, 1986). The GVF measures the “sum of squared deviations about the class mean” (Slocum et al., 2014). By calculating the error between class breaks, the accuracy score was obtained. The accuracy score calculation was based on the data field (population and household densities for this study), number of classes, study areas and data classification methods. One limitation of the GVF measure is its limited spatial proximity expressiveness. The calculation does not consider neighbouring polygons.

An accuracy score ranges from 0 to 1, where 1 represents the highest accuracy, meaning there is no generalisation or distortion, so each polygon belongs to a separate class. An accuracy score of 0 indicates complete generalisation, where all the polygons belong to a single class. Since the data are grouped into five classes based on a data classification method, an accuracy score of either 0 or 1 would not be possible.

5.2 Goodness of Variance Fit

The GVF was calculated for each of the 48 choropleth maps used for the user study, as described in Chapter 3. Statistical capabilities of RStudio 2023.03.1, Build 446[13] were utilised for all the accuracy score calculations, which include the following libraries: sp, sply, data.table, and readxl.

Figure 51 shows the R script created by the author to calculate the GVF score for the Buffalo City Metropolitan Municipality using each of the four data classification methods based on the hexagon geographic unit. Appendix D includes all the R scripts that depict the study areas and geographic units for each data classification method.

The upper- and lower-class breaks derived from the maps produced for the user study were used in each R script. Additionally, the GVF results were validated against an existing R package called Jenks71 (described in Chapter 2), which automatically calculated the GVF.

```
# HEX - Buffalo City Metropolitan Municipality - Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=2]<-1
dfdata$Group[dfdata$Join_Count>2 & dfdata$Join_Count<=18]<-2
dfdata$Group[dfdata$Join_Count>18 & dfdata$Join_Count<=118]<-3
dfdata$Group[dfdata$Join_Count>118 & dfdata$Join_Count<=785]<-4
dfdata$Group[dfdata$Join_Count>785]<-5

# HEX - Buffalo City Metropolitan Municipality - Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=1]<-1
dfdata$Group[dfdata$Join_Count>1 & dfdata$Join_Count<=10]<-2
dfdata$Group[dfdata$Join_Count>10 & dfdata$Join_Count<=100]<-3
dfdata$Group[dfdata$Join_Count>100 & dfdata$Join_Count<=1000]<-4
dfdata$Group[dfdata$Join_Count>1000]<-5

# HEX - Buffalo City Metropolitan Municipality - Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=146]<-1
dfdata$Group[dfdata$Join_Count>146 & dfdata$Join_Count<=493]<-2
dfdata$Group[dfdata$Join_Count>493 & dfdata$Join_Count<=1069]<-3
dfdata$Group[dfdata$Join_Count>1069 & dfdata$Join_Count<=1892]<-4
dfdata$Group[dfdata$Join_Count>1892]<-5

# HEX - Buffalo City Metropolitan Municipality - Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=1]<-1
dfdata$Group[dfdata$Join_Count>1 & dfdata$Join_Count<=11]<-2
dfdata$Group[dfdata$Join_Count>11 & dfdata$Join_Count<=45]<-3
dfdata$Group[dfdata$Join_Count>45 & dfdata$Join_Count<=161]<-4
dfdata$Group[dfdata$Join_Count>161]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(Join_Count)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(Join_Count)) %>%
  mutate(Class_Diff = abs(Join_Count-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(Join_Count-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

Figure 51: GVF script used for the Buffalo City Metropolitan Municipality for each of the four data classification methods based on the hexagon geographic unit

[13] https://posit.co/download/rstudio-desktop/

5.3 Results and Discussion

Jenks specified various algorithms to measure the error between class breaks, including the GVF, which was used in this study (Robinson, 1984). Subsequently, Jenks also introduced a data classification algorithm in 1977, which was proposed by Fisher, to identify data belonging to the same class (Dent et al., 2009). It was initially named the optimal method but is now more commonly referred to in current GIS software applications as natural breaks (Jenks). The objective was to minimise data classification errors.

Accordingly, it was expected that the results from the GVF calculation for this study would identify natural breaks (Jenks) as the optimal data classification method compared to the other methods. Nevertheless, the researcher wanted to confirm whether this is true, specifically in a South African context when visualising population demand, and also concerning the four different study areas and geographic units.

The results firstly highlight the overall GVF calculated mean per data classification method (Table *24*), followed by a breakdown per study area and the three geographic units: hexagons, small area layer, and sub-places. As expected, the highest accuracy score – indicating the lowest error between class breaks – was recorded for natural breaks (Jenks), with a value of 0.931. The geometric interval ranked second with a score of 0.773, followed by the logarithmic scale at 0.641. The lowest accuracy score recorded was for quantiles, which was 0.590.

Based on the choropleth maps with five classes showing population distribution in South Africa, the natural breaks (Jenks) method was considered the most optimal, followed by the geometric interval method. Similar results were found in a previous study by Smith (1986) who compared five traditional data classification methods (quartile, equal interval, standard deviation, natural breaks, and an optimisation method) based on the GVF measurement. The study found that only the optimisation method, now also referred to as natural breaks (Jenks), produced accurate results. For his study, a sample of 117 data sets was derived from the 1977 County and City Data Book, which includes data such as birth rates, population densities, and incomes. Natural breaks (Jenks) is further described in the literature as an appropriate method for visualising data that do not follow a normal distribution (Całka, 2018).

Table 24: GVF per data classification method

Data Classification Method	GVF (Mean)
Geometric interval	0.773
Logarithmic scale	0.641
Natural breaks (Jenks)	0.931
Quantiles	0.590

Tables 25, *26*, *27*, and *28* provide a detailed breakdown of the GVF accuracy score per data classification method by including the different study areas and geographic units. The goal was to determine whether either study areas or geographic units would influence the GVF accuracy score of a data classification method compared to the overall derived mean (Table *24*). Each table shows the results for a specific study area and includes the three geographic units. The mean accuracy score for each geographic unit was also recorded. The results were subsequently ranked with colour codes ranging from green to red. Green indicates the highest accuracy score, followed by yellow and orange. Red represents the lowest score.

The natural breaks (Jenks) data classification method once again outperformed the other methods, achieving the highest accuracy score, regardless of the study area or geographic unit. The second-highest score, based on mean accuracy, was calculated for the geometric interval.

Accuracy scores for the logarithmic scale and quantiles were not consistent across the different study areas and geographic units. Quantiles performed poorly, recording the lowest calculated accuracy score in the Buffalo City Metropolitan Municipality across all three geographic units. In the City of Tshwane Metropolitan Municipality, quantiles ranked lowest for both hexagons and the small area layer, but second lowest for sub-places. The ranking order for hexagons remained consistent across all the study areas.

The highest overall accuracy score of 0.963 was recorded for hexagons in the Mangaung Metropolitan Municipality. This was followed by hexagons in the Polokwane Local Municipality and sub-places in the Mangaung Metropolitan Municipality, which had scores of 0.954 and 0.948, respectively. All scores were based on the natural breaks (Jenks) data classification method. The lowest accuracy score was measured for hexagons, in the Mangaung Metropolitan Municipality, with a value of 0.250 based on quantiles.

Table 25: GVF – Buffalo City Metropolitan Municipality

Buffalo City Metropolitan Municipality				
Data classification method	Hexagon	Small area layer	Sub-place	Mean
Geometric interval	0.775	0.858	0.686	0.773
Logarithmic scale	0.782	0.719	0.652	0.718
Natural breaks (Jenks)	0.926	0.933	0.930	0.930
Quantiles	0.450	0.653	0.445	0.516

Table 26: GVF – City of Tshwane Metropolitan Municipality

City of Tshwane Metropolitan Municipality				
Data classification method	Hexagon	Small area layer	Sub-place	Mean
Geometric interval	0.757	0.758	0.877	0.797
Logarithmic scale	0.758	0.581	0.655	0.665
Natural breaks (Jenks)	0.947	0.907	0.928	0.928
Quantiles	0.480	0.550	0.732	0.587

Table 27: GVF – Mangaung Metropolitan Municipality

Mangaung Metropolitan Municipality				
Data classification method	Hexagon	Small area layer	Sub-place	Mean
Geometric interval	0.679	0.865	0.914	0.819
Logarithmic scale	0.819	0.505	0.565	0.630
Natural breaks (Jenks)	0.963	0.888	0.948	0.933
Quantiles	0.250	0.863	0.912	0.675

Table 28: GVF – Polokwane Local Municipality

Polokwane Local Municipality				
Data classification method	Hexagon	Small area layer	Sub-place	Mean
Geometric interval	0.659	0.649	0.807	0.705
Logarithmic scale	0.828	0.579	0.248	0.552
Natural breaks (Jenks)	0.954	0.916	0.933	0.935
Quantiles	0.506	0.559	0.676	0.580

The results from the GVF indicate that the error between class breaks measurement is sensitive to both study areas and geographic units. Hence, the choice of a geographic unit for a study area should be considered carefully, specifically in a South African context.

6. CONCLUSION

6.1 Summary of Findings

The concluding chapter summarises the results and provides recommendations related to the three research questions defined in Chapter 1. These questions contribute to the overall aim of assessing the suitability of data classification methods for choropleth maps to visualise population demand in South Africa effectively. Additionally, the limitations identified during this study are listed and described. Finally, potential topics for further research are highlighted.

6.1.1 Question 1 – Which Data Classification Methods Are Frequently Used to Visualise Statistical Data with Choropleth Maps?

Kraak and Ormeling (2020) mentioned that “it is a good cartographic practice to conveniently arrange the data before displaying them. This process is called classification”, which is a form of generalisation. Data classification methods for choropleth maps are used to group data into classes. These methods are generally used to analyse univariate data, which include a single variable, such as population density or mortality rate. Generally, a data classification method will determine the upper- and lower-class limits based on a specified number of classes. Since each data classification method produces notably different spatial patterns, the data classification method should be chosen carefully. Numerous data classification methods for choropleth maps are described in the literature and are available in most GIS applications. Some of the most documented methods include equal intervals, quantiles, mean-standard deviation, maximum breaks, natural breaks, and the optimal method. Arithmetic progression, nested mean, and geometric progression are less common but are also described in the literature. Besides these methods, users can design their own class breaks and manually adjust upper- and lower-class limits to accentuate a particular phenomenon.

Since ArcGIS Pro and QGIS are considered the most popular and frequently used applications in South Africa and globally, the data classification methods offered by both these applications were considered for this research. In addition to manual and defined intervals (where the user is required to set custom limits for each class), these methods include equal interval, geometric interval, logarithmic scale, natural breaks (Jenks), pretty breaks, quantiles, and standard deviation.

6.1.2 Question 2 – Which Measurement Techniques Are Used to Evaluate the Suitability and Effectiveness of These Data Classification Methods?

One of the key considerations, as recommended in the literature, for selecting a data classification method for a specific data set is data distribution. “Different frequency distributions suggest different class interval systems” (Evans, 1977).

In addition to data distribution, the literature identifies two effective ways for measuring or assessing data classification methods used in choropleth maps. These include, firstly, a user study in the form of a survey or questionnaire, and secondly, mathematical equations that measure the error between class breaks. This is useful for understanding the level of data generalisation within class breaks.

In user studies, questionnaires are frequently used to assess respondents’ interpretations of choropleth maps that depict different data classification methods. A user study typically comprises a list of map-specific questions, where correct and incorrect answers are used to determine which method(s) are better or more easily understood by the target audience, specifically the respondents participating in the user study.

For the second part, various mathematical equations for determining the error between class breaks are described in the literature, dating back to the 1970s (see Chapter 2). These equations are used to calculate an accuracy score or index indicating the level of data generalisation within each class break. Jenks and Caspall (1971) highlighted that since a choropleth map is a generalisation of reality based on an aerial or locational distribution, it must include some degree of error. Algorithms, as described by Jenks and Caspall (1971), include the overview accuracy index, the tabular accuracy index, and the boundary accuracy index. Additionally, the literature frequently describes two techniques: the GVF equation, which measures the “sum of squared deviations about the class mean” (Slocum et al., 2014), and the GADF, which is used to measure deviations about the class median.

When calculating the error between class breaks, an accuracy score is obtained. The accuracy score calculation is based on the data field, specifically the population and household densities in this study, and the number of classes. An accuracy score ranges from 0 to 1 where 1 represents the highest accuracy, meaning there is no generalisation or distortion; therefore, each polygon will belong to a separate class.

6.1.3 Question 3 – For a Geographic Accessibility Analysis of Service Centres In South Africa, Which Data Classification Method(s) for Choropleth Maps Are Best Interpreted by a Target Audience and Also Statistically Proven to Be Effective?

To determine which data classification method(s) for choropleth maps are best interpreted by a target audience and statistically proven to be effective for geographic accessibility analysis in South Africa, both measurement techniques described in Section 6.1.2 were implemented:

- Conducting user study
- Calculating the error between class breaks

From the seven data classification methods available in either ArcGIS Pro or QGIS, preliminary data evaluation and visual inspection were conducted to identify potential methods for testing in the user study and to calculate the errors between class breaks. In preparation, four representative municipalities were selected as study areas where people are distributed across different enumerator area types, reflecting the unique geographic landscape in South Africa. Additionally, three geographic units were identified, each demonstrating population densities for both equal-sized and varied-sized polygons. The histograms depicting population densities in these study areas and geographic units reveal that the population in South Africa is not normally distributed; instead, it is highly skewed. Some methods, such as standard deviation, are more effective when the data are evenly or normally distributed, as opposed to natural breaks, which are often used when the data distribution is skewed. Furthermore, the equal interval (and pretty breaks) method works best if the data are uniformly distributed, meaning the data distribution has no peak and is consistent (Kraak et al., 2021). Hence, the standard deviation, equal interval, and pretty breaks data classification methods, which are available in both ArcGIS Pro and QGIS, were subsequently excluded from further analysis.

The remaining data classification methods tested and evaluated in this research include geometric interval, logarithmic scale, natural breaks (Jenks), and quantiles.

6.1.3.1 User Study

The user study included the design of an online questionnaire. Respondents were asked to interpret choropleth maps by answering a series of questions about supply (service centres) and demand (population distribution) in four different municipalities across South Africa. Their responses helped inform decisions aimed at optimising service delivery through geographic accessibility analysis. The questionnaire was designed in Qualtrics[14] and consisted of 48 map-

[14] https://www.qualtrics.com/uk/

specific questions. For each question, respondents were required to identify one or more locations on a map.

Respondents

Participation was completely voluntary. The respondents were students from the University of Pretoria, enrolled in various programmes at the Department of GGM, both with and without prior experience in geography and GIS. Students were selected for the user study because they represent, in one way or another, the upcoming workforce. In total, 107 students participated. The majority were females (64%). Males comprised 35%, and one student preferred not to specify their gender. Additionally, most respondents were registered for the BEd programme (Education) at the time of the survey (40%), followed by BSc Geography and Environmental Science and BSc Geoinformatics with 14% and 11%, respectively.

Percentage Accuracy of Responses

Overall, respondents performed well in the user study, achieving an average percentage accuracy of 89.9% based on the 48 geographic accessibility questions. Four respondents scored 100%, while the lowest recorded accuracy was 64.6% among three individuals.

Respondents with prior knowledge or experience in geography performed slightly better, achieving a score of 92.0%, compared to those without such experience, who scored 88.8%. Students who indicated that they use maps either once a week or every day performed well (91.8% and 90.3%) compared to those who use them only once a month (87.8%) or rarely (86.2%).

A moderately strong linear relationship exists between the average percentage accuracy of respondents and their self-perceived level of expertise in the following categories: map reading, statistics, GIS, planning, spatial data, English and Google Maps. The strongest linear relationship was measured for spatial data (0.591), followed by Google Maps (−0.486), indicating a negative relationship, and statistics (0.426).

Data Classification Methods

Before comparing the results of the four data classification methods, a statistical test was conducted to assess the differences within groups, specifically focusing on the percentage accuracy of responses for each method. The results confirmed a significance score of less than 0.001, indicating statistically significant differences between the groups.

The percentage accuracy of responses for each of the four data classification methods was notably high. Respondents were more likely to provide correct answers for maps that used the

quantiles data classification method (92.3%), followed by the natural breaks (Jenks) method and the geometric interval, which had accuracies of 91.2% and 88.8%, respectively. Logarithmic scale was ranked lowest at 87.2% accuracy.

Statistical Significance

Although the overall goal and main objective of this research were to assess the suitability and effectiveness of different data classification methods to visualise population demand in South Africa, the researcher also wanted to test whether additional predictor variables, as defined in the user study design, could be significantly associated with response accuracies in general, as well as for each data classification method individually. These variables include four recurring questions related to geographic accessibility, four study areas, three geographic units, a self-perceived confidence level in answering a specific question and, lastly, a self-perceived difficulty level of a question.

A logistic regression analysis was used to test the significance of all predictor variables against a single dichotomous dependent variable, which represented respondents' answers as either correct or incorrect responses. (1 indicated a correct response, while 0 indicated an incorrect response). These predictor variables were tested against the entire data set with all responses combined, as well as individually for each data classification method. A significance score of < 0.001 was considered to be statistically significant.

Overall and for each data classification method, the geographic accessibility questions proved to be a statistically significant predictor (sig. < 0.001). This highlights the importance of understanding a specific research topic, such as geographic accessibility analysis, regardless of the data visualisation technique used. These geographic accessibility questions were designed based on real-world scenarios. Questions ranged in increasing difficulty levels. Question 1 was the easiest, and Question 4 was considered the most difficult.

Most participants answered Question 1 correctly (95.3%), followed by Questions 2, 3, and 4 with 94.0%, 88.2% and 81.5%, respectively. This is also true for all four data classification methods, except for quantiles, where respondents performed slightly better on Question 2 than on Question 1 (96.6% compared to 95.0%).

The four study areas were found to be significant overall, as well as for both the geometric interval and logarithmic scale methods. Overall, participants provided the most correct answers for choropleth maps depicting the Mangaung Metropolitan Municipality (92.2%). From a visual inspection, the municipality consists of large, sparsely populated areas with only a few high-density locations. The City of Tshwane Metropolitan Municipality was in second place, followed by the Buffalo City Metropolitan Municipality with scores of 89.9% and 89.5%,

respectively. The most incorrect answers were for the Polokwane Local Municipality, with 87.9%. Although a high percentage of accuracy was recorded for all study areas, the results suggest that the interpretation of data classification methods, or choropleth maps in general, is sensitive to specific localities or study areas.

The geographic units, hexagons, small area layers and sub-places, along with self-perceived confidence levels, were not found to be statistically significant predictors overall, nor for the interpretation of any of the four data classification methods. Although geographic units were not identified as statistically significant predictors, results suggest some variation in response accuracies between the four data classification methods and geographic units. Quantiles achieved the highest percentage accuracy for both the hexagon and small area layers, with 93.7% and 91.9%, respectively. On the sub-place level, natural breaks (Jenks) was first, followed by quantiles.

Self-perceived difficulty levels were tested and found to be significant (sig. < 0.001) for all data classification methods except for quantiles. For each of the 48 map questions, respondents were asked to indicate a difficulty level ranging from very easy, easy, neutral, difficult to very difficult. A percentage accuracy of 95% was calculated for questions considered to be very easy. This was followed by easy, neutral, and difficult, with percentages of 92.1%, 85.6%, and 80.4%, respectively. The percentage accuracy of questions that respondents considered very difficult was, however, higher at 83.8% compared to those considered difficult (80.4%).

6.1.3.2 Error Between Class Breaks

Lastly, the error between class breaks was calculated for each data classification method, depicting the three geographic units and four study areas. The purpose was to compare the human interpretation (the target audience) of data classification methods for choropleth maps (user study) with the recommended mathematical equation, which, in theory, could be considered less subjective. However, it is important to note that these equations were also developed by humans. The aim of the comparison was to highlight potential similarities and differences derived from the two approaches. The intent was to provide a more comprehensive view of the use of data classification methods for choropleth maps to visualise population demand in South Africa.

In this research, the GVF equation was used to calculate the error between class breaks. The results from the GVF calculation show that the natural breaks (Jenks) method achieved the highest accuracy score of 0.931, meaning the lowest error between class breaks, regardless of the study area or geographic unit. The geometric interval was ranked second with a score of 0.773, followed by the logarithmic scale at 0.641. The lowest accuracy score was measured for quantiles (0.590).

Accuracy scores for the logarithmic scale and quantiles were not as consistent across the different study areas and geographic units. Quantiles performed poorly, with the lowest calculated accuracy score for the Buffalo City Metropolitan Municipality across all three geographic units. In the City of Tshwane Metropolitan Municipality, quantiles ranked lowest for both hexagons and small area layers.

6.2 Concluding Remarks

The high percentage accuracy achieved during the user study demonstrates that choropleth maps are an effective and easy-to-use technique for visualising population demand in South Africa. It also highlights the fact that professionals can consider any of the four selected data classification methods to analyse supply and demand for the optimal positioning of service centres. With that in mind, respondents were more likely to provide correct answers for maps depicting the quantiles and natural breaks (Jenks) data classification methods. This suggests that these methods are easier to interpret and analyse in relation to population distribution in South Africa than the other methods.

Furthermore, the mathematical equation used to calculate the error between class breaks yielded somewhat different results. Based on these calculations, natural breaks (Jenks) and geometric intervals were considered the optimal data classification methods, while the logarithmic scale and quantiles ranked lowest.

The results from both the user study and error calculation provided a more comprehensive view of the use of data classification methods. This research also emphasises the importance of human interpretation, and the inclusion thereof, when assessing methods or techniques that represent spatial phenomena.

Although both the user study and GVF measures are recommended in the literature to assess the effectiveness of data classification methods for choropleth maps, the results between these two techniques were somewhat different. One significant difference is that respondents best interpreted quantiles in the user study, but it was ranked lowest based on the GVF measure. Results from the user study align with Brewer and Pickle's mortality rate findings, where quantiles was also the preferred method (Brewer and Pickle, 2002). Also, their study ranked natural breaks (Jenks) third overall.

As discussed in Chapter 2, quantiles group an equal number of features in each class (De Smith et al., 2018). As a result, class intervals are usually not consistent. This method is effective if the data are not normally distributed. This means that significant variations

(minimum and maximum values) within a class could occur, which is problematic for the GVF technique, which measures the “sum of squared deviations about the class mean”.

Both techniques considered natural breaks (Jenks) an effective data classification method; ranked 1st for GVF and 2nd in the user study. In addition, logarithmic scale was not considered effective, ranked lowest in the user study and 3rd for the GVF measure. Data distribution is probably one of the contributing factors causing variation in results between the two techniques, which are worth further research and testing, specifically on more normally distributed data sets.

6.3 Limitations

During this research, several limitations were identified. Statistics South Africa is the custodian of demographic data in the country. Although a national census survey was conducted in 2022, demographic data has not yet been released at a more granular level than local and metropolitan municipalities. Hence, for this research, demographic (or population) data from Census 2011 were used, as the data are available at more granular geographic units, such as small area layers and sub-places.

While interactive maps depicting data at various geographic scales are becoming increasingly popular, only static choropleth maps were designed and presented to the respondents. This research focused on the visual interpretation of choropleth maps depicting population demand in South Africa for geographic accessibility analysis. Although this is a significant and novel contribution to science, it should be viewed in the context of its broader implications. Other factors not included in this research, such as the incorporation of a routable travel network depicting actual travel distances, as well as the capacity, availability, and attractiveness of service centres, are also significant and essential for conducting a comprehensive geographic accessibility analysis.

The user study involved the voluntary participation of students from the University of Pretoria enrolled in programmes offered by the Department of GGM. Ideally, this study could also include participation from the current workforce who are, in one way or another, working with geospatial information.

As with many surveys or questionnaires, respondent fatigue is a reality. Respondent fatigue occurs when respondents become disengaged or they lose focus in answering questions, leading to a potential decrease in data quality.

Validating click-based responses as correct or incorrect was done manually. Although the validation was done thoroughly and systematically, human error is always possible.

6.4 Further Research and Scope for Future Studies

- Although literature recommends using five to seven classes for choropleth maps, this research tested only five class intervals to ensure consistency and maintain a manageable questionnaire size. Further research could include varying the number of class intervals assessed with choropleth maps. One of the benefits of the error calculations between class breaks is the potential to determine an optimal number of class breaks for a specific data set and classification method.

- In addition to the GVF error measurement used in this research, several other calculations are also available and described in the literature. These include the boundary accuracy index, overview accuracy index, tabular accuracy index, and the GADF. Future research could involve comparing the GVF results with those obtained from other measurement techniques. One limitation of the GVF measure is its limited spatial proximity expressiveness. The calculation does not consider neighbouring polygons. Future research could investigate the possibility of including spatial proximity of polygons (neighbourhoods) as a variable.

- The results of this study could be valuable for professionals who prepare choropleth maps in other applications, such as the IEC for voter registration and election results and Statistics South Africa for census results. Additionally, the results could be tested in other countries with similar demographic characteristics. Results from this study are transferable and could be utilised in other applications such as the spatial visualisation of building densities, animal or species densities, crime pattern densities, or disease mapping.

- Future research could investigate why some classification methods performed worse, and also to investigate variances in the underlying data.

- For both the user study and the calculation of error between class breaks, four municipalities were identified, representing the unique population distribution in South Africa. Further research could include other municipalities, focusing on population distribution that depicts both urban and rural areas.

- Although geographic units were not identified as a statistically significant predictor, the results suggest some variation in response accuracies between the four data classification methods and geographic units. Follow-up investigations could be useful in understanding the impact of different geographic units on visualising choropleth

maps. Tests could include different localities and geographic units such as voting districts, wards, or municipalities.

- The results emanating from this study could be tested in other countries with similar population distribution characteristics to determine possible similarities and differences based on the use of these data classification methods.
- During the user study, respondents' self-perceived levels of efficiency in different categories were compared to the accuracy of their responses in the questionnaire. The goal was to determine whether specific skills would increase the probability of respondents correctly interpreting choropleth maps. A moderately strong linear relationship was observed between the average percentage accuracy of respondents and their self-perceived level of expertise in the following categories: map reading, statistics, GIS, planning, spatial data, English, and Google Maps. Further research could include experiments to confirm whether these relationships exist factually.
- Future studies could include qualitative research by means of user interviews, allowing respondents to describe their interpretation of choropleth maps based on different data classification methods.

REFERENCES

AB HAMID, J., JUNI, M. H., ABDUL MANAF, R., SYED ISMAIL, S. N. & LIM, P. Y. 2023. Spatial accessibility of primary care in the dual public-private health system in rural areas, Malaysia. *International Journal of Environmental Research and Public Health,* 20(4), 3147.

AFIFAH, Z. 2019. Effectiveness of classification method and color symbol scheme on choropleth map of population density in special region of Yogyakarta. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, XLII-4/W16, 55–61. https://doi.org/10.5194/isprs-archives-XLII-4-W16-55-2019

ALBERT, G., ILYÉS, V., KIS, D., SZIGETI, C. & VÁRKONYI, D. 2016. Testing the map reading skills of university students. In *Proceedings of the 6th International Conference on Cartography and GIS*, 188–199.

ARMSTRONG, M. P., XIAO, N. & BENNETT, D. A. 2003. Using genetic algorithms to create multicriteria class intervals for choropleth maps. *Annals of the Association of American Geographers,* 93(3), 595–623.

ASHIAGBOR, G., OFORI-ASENSO, R., FORKUO, E. K. & AGYEI-FRIMPONG, S. 2020. Measures of geographic accessibility to health care in the Ashanti Region of Ghana. *Scientific African,* 9, e00453.

BARROZO, L. V., PÉREZ-MACHADO, R. P., SMALL, C. & CABRAL-MIRANDA, W. 2016. Changing spatial perception: Dasymetric mapping to improve analysis of health outcomes in a megacity. *Journal of Maps,* 12(5), 1242–1247.

BASKARADA, S. & KORONIOS, A. 2013. Data, information, knowledge, wisdom (DIKW): A semiotic theoretical and empirical exploration of the hierarchy and its quality dimension. *Australasian Journal of Information Systems,* 18(1). https://doi.org/10.3127/ajis.v18i1.748

BEKKER, M. 2023. Depends on how you count them: The value of general propensity choropleth maps for visualising databases of protest incidents. *Journal of Maps,* 19(1), 2064778.

BELLINGER, G., CASTRO, D. & MILLS, A. 2004. Data, information, knowledge, and wisdom. https://homepages.dcc.ufmg.br/~amendes/SistemasInformacaoTP/TextosBasicos/Data-Information-Knowledge.pdf

BESANÇON, L., COOPER, M., YNNERMAN, A. & VERNIER, F. 2020. An evaluation of visualization methods for population statistics based on choropleth maps. *arXiv:2005.00324*. https://doi.org/10.48550/arXiv.2005.00324

BOLSTAD, P. 2012. *GIS fundamental.*, White Bear Lake, MN, Eider Press.

BOSCOE, F. P. & PICKLE, L. W. 2003. Choosing geographic units for choropleth rate maps, with an emphasis on public health applications. *Cartography and Geographic Information Science,* 30(3), 237–248.

BREWER, C. 2015. *Designing better maps: A guide for GIS users*. Redlands, CA, ESRI Press.

BREWER, C. A. 2006. Basic mapping principles for visualizing cancer data using geographic information systems (GIS). *American Journal of Preventive Medicine,* 30(2), S25–S36.

BREWER, C. A. & PICKLE, L. 2002. Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers,* 92(4), 662–681.

BUFFALO CITY METROPOLITAN MUNICIPALITY. n.d. Overview of Buffalo City Metropolitan Municipality. https://www.buffalocity.gov.za/about.php

BROWN, L. A. & CHUNG, S. Y. 2006. Spatial segregation, segregation indices and the geographical perspective. *Population, Space and Place,* 12(2), 125–143.

CAŁKA, B. 2018. Comparing continuity and compactness of choropleth map classes. *Geodesy and Cartography,* 67(1), 21–34.

CHANDRA, S. & MISTRI, B. 2011. Approach to select suitable choropleth mapping. *Journal of Interacademicia*, 15(2), 240–258.

CHEN, L., CHEN, T., LAN, T., CHEN, C. & PAN, J. 2023. The contributions of population distribution, healthcare resourcing, and transportation infrastructure to spatial accessibility of health care. *Inquiry,* 60, 469580221146041.

CHURCH, R. L. & MURRAY, A. T. 2009. *Business site selection, location analysis, and GIS*. Hoboken, NJ, John Wiley & Sons.

CITY OF TSHWANE. n.d. Tshwane at a Glance. https://www.tshwane.gov.za/?page_id=683

CLIFFORD, N., COPE, M., GILLESPIE, T. & FRENCH, S. 2016. *Key methods in geography*. Los Angeles, CA, SAGE.

COETZEE, S., SNYMAN, L. & DELPORT, R. 2021. Revealing the value of geospatial information with isochrone maps for improving the management of heart attacks in South Africa. *International Journal of Cartography,* 7(2), 184–189.

CROMLEY, G. A. 2019. Measuring differential access to facilities between population groups using spatial Lorenz curves and related indices. *Transactions in GIS,* 23(6), 1332–1351.

DE SMITH, M. J., GOODCHILD, P. M. F. & LONGLEY, P. P. A. 2018. *Geospatial analysis: A comprehensive guide to principles techniques and software tools.* East Sussex, The Winchelsea Press.

DECLERQ, F. 1995. Choropleth map accuracy and the number of class intervals. In *Proceedings of the 17th Conference and the 10th General Assembly of the International Cartographic Association*, Barcelona, 918–22.

DENT, B. D., TORGUSON, J. & HODLER, T. W. 2009. *Cartography: Thematic map design.* New York, NY, McGraw-Hill Higher Education.

DEPARTMENT OF PUBLIC SERVICE AND ADMINISTRATION (DPSA). 2012. *Guideline: Improving geographic access to government service points*. Pretoria. https://www.dpsa.gov.za/dpsa2g/documents/sdot/2012/DPSA_Guideline%20on%20Access%20to%20Service%20Points.pdf

DEPARTMENT OF PUBLIC SERVICE AND ADMINISTRATION (DPSA). 2013. *Geographic accessibility study of social facility and government service points for the metropolitan cities of Johannesburg and eThekwini 2011/12: Part C*. Pretoria.

DEPARTMENT OF PUBLIC SERVICE AND ADMINISTRATION (DPSA). 2021. *Guideline: Improving geographic access to government service points*. Pretoria. https://www.dpsa.gov.za/dpsa2g/documents/service_access/Guideline%20On%20Improving%20Geographic%20Access%20To%20Goverment%20Service%20Points%20-%202021.pdf

ELECTORAL COMMISSION OF SOUTH AFRICA (IEC). 2019. *Atlas of results: National & provincial elections*. Pretoria. https://atlas.elections.org.za/npeatlas/#

EVANS, I. S. 1977. The selection of class intervals. *Transactions of the Institute of British Geographers*, 2(1), 98–124.

FRIENDLY, M. 2008. A brief history of data visualization. In C. Chen, W. Härdle & A. Unwin (Eds.), *Handbook of data visualization*, pp. 15–56. Berlin, Springer.

FRIESEN, L., SCHAAB, G., COETZEE, S., RAUTENBACH, V., MARCUS, T. & HUGO, J. 2018. Community-oriented primary care (COPC) in the city of Tshwane (South Africa): A web map application in support of responsive and dynamic health care. *Kartographische Nachrichten*, 68, 173–182.

GIS GEOGRAPHY. 2019. 25 map types for building unbeatable maps. https://gisgeography.com/map-types/

GIS GEOGRAPHY. 2022. Mapping out the GIS software landscape. https://gisgeography.com/best-gis-software/

GOLIAN, S., SAGHAFIAN, B., SHESHANGOSHT, S. & GHALKHANI, H. 2010. Comparison of classification and clustering methods in spatial rainfall pattern recognition at Northern Iran. *Theoretical and Applied Climatology,* 102(3), 319–329.

GREEN, C. A. 2012. *CSIR guidelines for the provision of social facilities in South African settlements*. Pretoria, CSIR Built Environment.

HARRIS, R., CHARLTON, M., BRUNSDON, C. & MANLEY, D. 2017. Balancing visibility and distortion: Remapping the results of the 2015 UK General Election. *Environment and Planning A: Economy and Space,* 49(9), 1945–1947.

HATEM, G., ZEIDAN, J., GOOSSENS, M. & MOREIRA, C. 2022. Normality testing methods and the importance of skewness and kurtosis in statistical analysis. BAU Journal-Science and Technology, 3, 7.

HEALY, L. M. 2006. *Logistic regression: An overview*. Ypsilanti, MI, Eastern Michigan College of Technology.

HEGARTY, M., RICHARDSON, A. E., MONTELLO, D. R., LOVELACE, K. & SUBBIAH, I. 2002. Development of a self-report measure of environmental spatial ability. *Intelligence,* 30(5), 425–447.

HENDERSON, A. R. 2006. Testing experimental data for univariate normality. *Clinica Chimica Acta,* 366(1–2), 112–129.

HEY, J. 2004. The data, information, knowledge, wisdom chain: The metaphorical link. *Intergovernmental Oceanographic Commission,* 26, 1–18.

HO, A. D. & YU, C. C. 2015. Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects. *Educational and Psychological Measurement*, 75, 365-388.

HOGRÄFER, M., HEITZLER, M. & SCHULZ, H.-J. 2020. The state of the art in map-like visualization. *Computer Graphics Forum,* 39(3), 647–674.

JENKS, G. F. 1963. Generalization in statistical mapping. *Annals of the Association of American Geographers,* 53(1), 15–26.

JENKS, G. F. & CASPALL, F. C. 1971. Error on choroplethic maps: Definition, measurement, reduction. *Annals of the Association of American Geographers,* 61(2), 217–244.

JIANG, B. 2012. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *Professional Geographer,* 65(3). http://dx.doi.org/10.1080/00330124.2012.700499

JONES, W. D. 1930. Ratios and isopleth maps in regional investigation of agricultural land occupance. *Annals of the Association of American Geographers,* 20, 177–195.

JUERGENS, C. 2020. Trustworthy COVID-19 mapping: Geo-spatial data literacy aspects of choropleth maps. *KN – Journal of Cartography and Geographic Information,* 70, 155–161.

KHAN, S. & MOHIUDDIN, K. 2018. Evaluating the parameters of ArcGIS and QGIS for GIS applications. *International Journal of Advance Research in Science and Engineering,* 7(3), 582–594.

KHUMALO, G. E., NTULI, S., LUTGE, E. & MASHAMBA-THOMPSON, T. P. 2022. Geo-analysis: The distribution of community health workers in relation to the HIV prevalence in KwaZulu-Natal province, South Africa. *BMC Health Services Research,* 22, 326.

KOSARA, R., HEALEY, C. G., INTERRANTE, V., LAIDLAW, D. H. & WARE, C. 2003. Thoughts on user studies: Why, how, and when. *IEEE Computer Graphics and Applications,* 23(4), 20–25.

KRAAK, M.-J. & ORMELING, F. 2011. *Cartography: Visualization of spatial data*. New York, NY, Guilford Publications.

KRAAK, M.-J. & ORMELING, F. 2020. *Cartography: Visualization of geospatial data*. Boca Raton, FL, CRC Press.

KRAAK, M.-J., ROTH, R. E., RICKER, B., KAGAWA, A. & LE SOURD, G. 2021. *Mapping for a sustainable world*. New York, NY, United Nations.

KWAK, S. G. & PARK, S.-H. 2019. Normality test in clinical research. *Journal of Rheumatic Diseases,* 26(1), 5–11.

LAERD STATISTICS. n.d. *Kruskal–Wallis H Test using SPSS Statistics*. https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

LEE, J. & BEDNARZ, R. 2012. Components of spatial thinking: Evidence from a spatial thinking ability test. *Journal of Geography,* 111(1), 15–26.

LLOYD, R. & STEINKE, T. 1977. Visual and statistical comparison of choropleth maps. *Annals of the Association of American Geographers,* 67(3), 429–436.

LLOYD, R. E. & BUNCH, R. L. 2008. Explaining map-reading performance efficiency: gender, memory, and geographic information. *Cartography and Geographic Information Science,* 35, 171–202.

MA, L., LUO, N., WAN, T., HU, C. & PENG, M. 2018. An improved healthcare accessibility measure considering the temporal dimension and population demand of different ages. *International Journal of Environmental Research and Public Health,* 15(11), 2421.

MAKURA, C. B., SCHNIPPEL, K., MICHELOW, P., CHIBWESHA, C. J., GOEIEMAN, B., JORDAAN, S. & FIRNHABER, C. 2016. Choropleth mapping of cervical cancer screening in South Africa using healthcare facility-level data from the national laboratory network. *AIMS Public Health,* 3(4), 849–862.

MOETI, T., MOKHELE, T., WEIR-SMITH, G., DLAMINI, S. & TESFAMICHEAL, S. 2023. Factors affecting access to public healthcare facilities in the City of Tshwane, South Africa. *International Journal of Environmental Research and Public Health,* 20(4), 3651.

MONMONIER, M. 1993. *Mapping it out: expository cartography for the humanities and social sciences*. Chicago, IL, University of Chicago Press.

MONMONIER, M. 2018. *How to lie with maps*. Chicago, IL, University of Chicago Press.

MONTELLO, D. & SUTTON, P. 2012. *An introduction to scientific research methods in geography and environmental studies*. Thousand Oaks, CA, SAGE.

MOSIANE, N. & MURRAY, J. 2021. *Distribution of population vs economic activity across the GCR*. Gauteng City-Region Observatory, Johannesburg.

MOTLANA, M. K. T. N., GININDZA, T. G., MITKU, A. A. & JAFTA, N. 2021. Spatial distribution of cancer cases seen in three major public hospitals in KwaZulu-Natal, South Africa. *Cancer Informatics,* 20. https://doi.org/10.1177/11769351211028194

O'SULLIVAN, D. & UNWIN, D. 2010. *Geographic information analysis*. Hoboken, NJ, John Wiley & Sons.

O'SULLIVAN, D. & UNWIN, D. 2014. *Geographic information analysis*, 2nd ed. New York, NY, Wiley.

PETERSON, G. N. 2015. *GIS cartography: A guide to effective map design.* Boca Raton, FL, CRC Press.

RAUTENBACH, V., COETZEE, S. & ÇÖLTEKIN, A. 2017. Development and evaluation of a specialized task taxonomy for spatial planning–A map literacy experiment with topographic maps. *ISPRS Journal of Photogrammetry and Remote Sensing,* 127, 16–26.

RAUTENBACH, V., COETZEE, S. M. & ÇÖLTEKIN, A. 2014. Towards evaluating the map literacy of planners in 2D maps and 3D models in South Africa. 2014. *AfricaGEO 2014 Conference Proceedings*.

REGMI, P. R., WAITHAKA, E., PAUDYAL, A., SIMKHADA, P. & VAN TEIJLINGEN, E. 2016. Guide to the design and application of online questionnaire surveys. *Nepal Journal of Epidemiology,* 6(4), 640–644.

ROBINSON, A. H. 1984. *Elements of cartography.* New York, NY, Wiley.

ROBINSON, A. H. 1995. *Elements of cartography*. 6th ed. New York, NY, Wiley.

RODRIGUE, J.-P, COMTOIS, C. & SLACK, B. 2009. *The geography of transport systems.* New York, NY, Routledge.

SCHIEWE, J. 2019. Empirical studies on the visual perception of spatial patterns in choropleth maps. *KN – Journal of Cartography and Geographic Information,* 69, 217–228.

SCHIEWE, J. 2023. Preserving change information in multi-temporal choropleth maps through an extended data classification method. *The Cartographic Journal*, 61(2), 1–14.

SCHIEWE, J. 2024. Task-oriented and change-preserving data classification for multi-temporal choropleth maps. *KN – Journal of Cartography and Geographic Information,* 74, 17–27.

SCHULTZ, G. M. 1961. An experiment in selecting value scales for statistical distribution maps. *Surveying and Mapping,* 21, 224–230.

SHAITO, M. & ELMASRI, R. 2021, Map visualization using spatial and spatio-temporal data: Application to Covid-19 data. In *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference*, 284–291.

SLOCUM, T. A., MCMASTER, R. B., KESSLER, F. C. & HOWARD, H. H. 2014. *Thematic cartography and geovisualization*. Cambridge, Pearson.

SMITH, R. M. 1986. Comparing traditional methods for selecting class intervals on choropleth maps. *The Professional Geographer,* 38(1), 62–67.

SNYMAN, L. & COETZEE, S. 2024. Measuring geographic accessibility in data poor rural areas by augmenting the road network with a triangular irregular network: A case study in the O.R. Tambo District Municipality of the Eastern Cape, South Africa. *Journal of Transport Geography,* 115, 103808.

STATISTICS SOUTH AFRICA. n.d. *Updating geo-frame for census*. Pretoria. https://www.statssa.gov.za/?page_id=11703

STATISTICS SOUTH AFRICA. 2012a. *Census 2011: How the count was done*. Pretoria. http://statssa.gov.za/census/census_2011/census_products/Census_2011_How_the_count_was_done.pdf

STATISTICS SOUTH AFRICA. 2012b. *Census 2011: Metadata*. Pretoria. https://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Metadata.pdf

STATISTICS SOUTH AFRICA. 2012c. *Census 2011: Statistical release P0301.4*. Pretoria. https://www.statssa.gov.za/publications/P03014/P030142011.pdf

STATISTICS SOUTH AFRICA. 2012d. *Statistics by place: City of Tshwane*. Pretoria. https://www.statssa.gov.za/?page_id=993&id=city-of-tshwane-municipality

STATISTICS SOUTH AFRICA. 2012e. *Statistics by place: Mangaung*. Pretoria. https://www.statssa.gov.za/?page_id=993&id=mangaung-municipality

STATISTICS SOUTH AFRICA. 2023a. *Census 2022: Statistical release P0301.4*. Pretoria. https://census.statssa.gov.za/assets/documents/2022/P03014_Census_2022_Statistical_Release.pdf

STATISTICS SOUTH AFRICA. 2023b. *Census portal*. https://census.statssa.gov.za/#/

STRAUSS, M. 2019. A historical exposition of spatial injustice and segregated urban settlement in South Africa. *Fundamina*, 25(2), 135–168.

SUKRAINI, T. T., YASA, I. & WIGUNA, P. P. K. 2022. Comparing choropleth and graduated symbols: How different map types affect public understanding in Covid-19 map reading in Badung Regency, Bali, Indonesia. *Geographia Technica,* 17(1), 150–166.

SUPERCROSS 8.0.2. 2012. [Database]. Space-Time Research.

THE EUROPEAN SPACE AGENCY. n.d. *Spot 5 instruments.* Paris. https://earth.esa.int/eogateway/missions/spot-5

TOMASZEWSKI, B., VODACEK, A., PARODY, R. & HOLT, N. 2015. Spatial thinking ability assessment in Rwandan secondary schools: Baseline results. *Journal of Geography,* 114(2), 39–48.

TRAUN, C. & AND LOIDL, M. 2012. Autocorrelation-based regioclassification: A self-calibrating classification approach for choropleth maps explicitly considering spatial autocorrelation. *International Journal of Geographical Information Science,* 26(5), 923–939.

TYNER, J. A. 2014. *Principles of map design*. New York, NY, Guilford Publications.

UN-HABITAT. n.d. South Africa. https://unhabitat.org/south-africa

VASILCA, D. 2019. How to create an effective thematic map. *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture* 76(2), 258.

WEIR-SMITH, G. & DLAMINI, S. 2024. Hit the road: Spatial characteristics of labor absorption in South Africa. *The Professional Geographer,* 76(3), 331–342.

WORLD BANK. 2018. *Overcoming poverty and inequality in South Africa: An assessment of drivers, constraints, and opportunities.* World Bank Publications, Washington, DC.

WRIGHT, J. K. 1942. Map makers are human: Comments on the subjective in maps. *Geographical Review,* 32, 527–544.

XIAO, C., YE, J., ESTEVES, R. M. & RONG, C. 2016. Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience,* 28(14), 3866–3878.

ZHANG, J., XU, L., ZHANG, Y., LIU, G., ZHAO, L. & WANG, Y. 2019. An on-demand scalable model for geographic information system (GIS) data processing in a cloud GIS. *ISPRS International Journal of Geo-Information,* 8(9), 392.

ZHAO, L., CHEN, L., RANJAN, R., CHOO, K.-K. R. & HE, J. 2016. Geographical information system parallelization for spatial big data processing: A review. *Cluster Computing,* 19, 139–152.

APPENDICES

APPENDIX A: ETHICS APPROVAL

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Natural and Agricultural Sciences

Deputy Dean of Research and Post Graduate Studies Preliminary Approval

02 May 2023

Mr LF Snyman
Department of Geography Geoinformatics and Meteorology
Faculty of Natural and Agricultural Sc
University of Pretoria

Dear Mr LF Snyman

PERMISSION FROM DEAN'S OFFICE FOR RESEARCH PROJECT NAS021/2023

The letter serves to confirm that I am supportive of the following Doctoral research project:

EVALUATING DATA CLASSIFICATION METHODS FOR CHOROPLETH MAPS TO ANALYSE GEOGRAPHIC ACCESSIBILITY IN SOUTH AFRICA

I have no objection to the research team requesting the staff/students from the Faculty of Natural and Agricultural Sciences to participate in this research project, **subject to ethics approval by the Faculty of Natural and Agricultural Sciences Research Ethics Committee.**

Kind regards

Prof Vinesh Maharaj
Deputy Dean: Research and Post Graduate Studies
Faculty of Natural and Agricultural Sciences

APPENDIX B: TEST FOR NORMALITY

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Q_Number	Statistic	df	Sig.	Statistic	df	Sig.
Percentage	1	,534	107	<,001	,239	107	<,001
	2	,537	107	<,001	,289	107	<,001
	3	,411	107	<,001	,647	107	<,001
	4	,476	107	<,001	,519	107	<,001
	5	,540	107	<,001	,216	107	<,001
	6	,537	107	<,001	,289	107	<,001
	7	,521	107	<,001	,350	107	<,001
	8	,354	107	<,001	,719	107	<,001
	9	,535	107	<,001	,241	107	<,001
	10	,538	107	<,001	,266	107	<,001
	11	,437	107	<,001	,603	107	<,001
	12	,508	107	<,001	,428	107	<,001
	13	,509	107	<,001	,423	107	<,001
	14	,519	107	<,001	,397	107	<,001
	15	,451	107	<,001	,572	107	<,001
	16	,355	107	<,001	,714	107	<,001
	17	,518	107	<,001	,307	107	<,001
	18	,537	107	<,001	,289	107	<,001
	19	,419	107	<,001	,635	107	<,001
	20	,502	107	<,001	,444	107	<,001
	21	,430	107	<,001	,545	107	<,001
	22	,540	107	<,001	,242	107	<,001
	23	,511	107	<,001	,401	107	<,001
	24	,503	107	<,001	,443	107	<,001
	25	,529	107	<,001	,071	107	<,001
	26	,539	107	<,001	,154	107	<,001
	27	,502	107	<,001	,434	107	<,001
	28	,505	107	<,001	,442	107	<,001
	29	,536	107	<,001	,183	107	<,001
	30	,540	107	<,001	,186	107	<,001
	31	,433	107	<,001	,614	107	<,001
	32	,435	107	<,001	,602	107	<,001
	33	,539	107	<,001	,219	107	<,001
	34	,537	107	<,001	,289	107	<,001
	35	,502	107	<,001	,436	107	<,001
	36	,493	107	<,001	,471	107	<,001
	37	,536	107	<,001	,183	107	<,001
	38	.	107	.	.	107	.
	39	,527	107	<,001	,309	107	<,001
	40	,496	107	<,001	,469	107	<,001
	41	,506	107	<,001	,360	107	<,001
	42	,536	107	<,001	,117	107	<,001
	43	,529	107	<,001	,348	107	<,001
	44	,493	107	<,001	,471	107	<,001
	45	,527	107	<,001	,278	107	<,001
	46	,534	107	<,001	,310	107	<,001
	47	,530	107	<,001	,286	107	<,001
	48	,452	107	<,001	,570	107	<,001

a. Lilliefors Significance Correction

APPENDIX C: USER STUDY QUESTIONNAIRE

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Block1

Dear participant
This survey evaluates the interpretation of different data classification methods for choropleth maps. Results from the survey will be used to develop a set of good practices for professionals who need to visualise population demand with choropleth maps in order to conduct geographic accessibility studies.

What is geographic accessibility and how is it measured?
Geographic accessibility is measured by calculating the physical distance people travel to specific facilities or service centres (such as clinics, community centres, police stations etc.). "Travelling long distances to reach these centres is costly and time consuming, especially to those who suffer the burden of poverty and deprivation". Measuring geographic accessibility enables policy makers to implement effective strategies for the optimal positioning of service centres (close to the people).

Choropleth (or thematic) maps are frequently used to visualise population demand for geographic accessibility studies. It is however noted that these maps are sometimes incorrectly interpreted or misunderstood which leads to ineffective optimisation strategies. These strategies include recommendations for (1) where to open a new service centre, i.e. the 'expansion model' (2) where to close a current service centre, i.e. the 'reduction model' and (3) where to move a current service centre to a different location, i.e. the 'relocation model'.

What you need to do

You will be shown 2 general map reading questions, as well as 48 map interpretation questions specifically related to geographic accessibility (where to open, close or move service centres). Please read each question carefully and then follow the instructions for providing an answer.

Note that consent cannot be withdrawn once the questionnaire is submitted as there is no way to trace the particular questionnaire that has been filled in. Please answer the questions in the questionnaire as completely and honestly as possible. This should not take more than 30 minutes of your time. This study has received written approval from Research Ethics Committees of the Faculty of Natural and Agricultural Sciences (tel: 012 420 4356). The results of the survey may be published in the media and/or an academic journal without identifying any of the participants individually. We will provide you with a summary of our findings on request.
If you have any questions or comments, please do not hesitate to ask the facilitators or contact Lourens Snyman, lourens.snyman@up.ac.za

By clicking 'Next', you agree to the above terms and provide your consent to use these results in this study.

Thank you

Block2

Section 1: General Questions
Please answer all the questions

Question 1 - Have you done this survey before?

- ○ Yes
- ○ No

Question 2 - On which device are you completing the survey? For the best user experience, please use a PC or Laptop

- ○ PC/Laptop
- ○ Tablet
- ○ Mobile smart phone

Question 3 - What is your age?

Question 4 - What gender do you identify as?

- ○ Male
- ○ Female
- ○ Non-binary / third gender
- ○ Prefer not to say

Question 5 - Did you take geography as a subject at school or university?

- ○ Yes
- ○ No

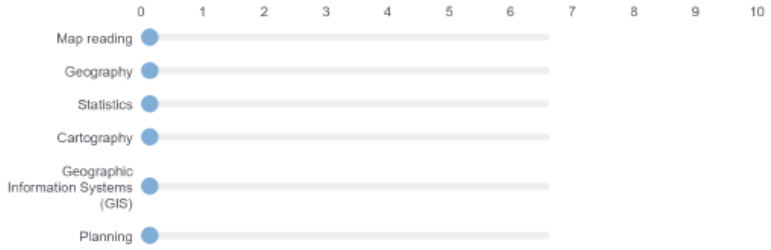
Question 6 - For which academic programme are you registered? Please select one of the following options:

- ○ BSc Geography and Environmental Science
- ○ BSc Geoinformatics
- ○ BSc Meteorology
- ○ BA specialising in Geographical Sciences
- ○ Other

Question 7 - Have you complete or are you currently enrolled for any of the following modules? You can select multiple entries.

- ☐ GGY283
- ☐ GIS310
- ☐ GMC110
- ☐ GGY383
- ☐ GIS221
- ☐ GIS708

Question 8 - Rate your level of training (or expertise) in the following categories



Question 9 - Have you ever lived in any of the following areas? You can select multiple locations

- ☐ Buffalo City (East London)
- ☐ City of Tshwane
- ☐ Mangaung (Bloemfontein)
- ☐ Polokwane
- ☐ None of the above

Question 10 - How often do you use maps?

- ○ Every day
- ○ Once a week
- ○ Once a month
- ○ Rarely

Question 11 - Have you ever been told by a professional that you have imperfect color vision?

- ○ Yes
- ○ No

Question 12 - Use the image below to identify the eight different colors (test colors - color oracle - web

	Blue	Light Blue	Green	Light Green	Red	Light Red	Yellow	Light Yellow
A	○	○	○	○	○	○	○	○
B	○	○	○	○	○	○	○	○
C	○	○	○	○	○	○	○	○
D	○	○	○	○	○	○	○	○
E	○	○	○	○	○	○	○	○
F	○	○	○	○	○	○	○	○
G	○	○	○	○	○	○	○	○
H	○	○	○	○	○	○	○	○

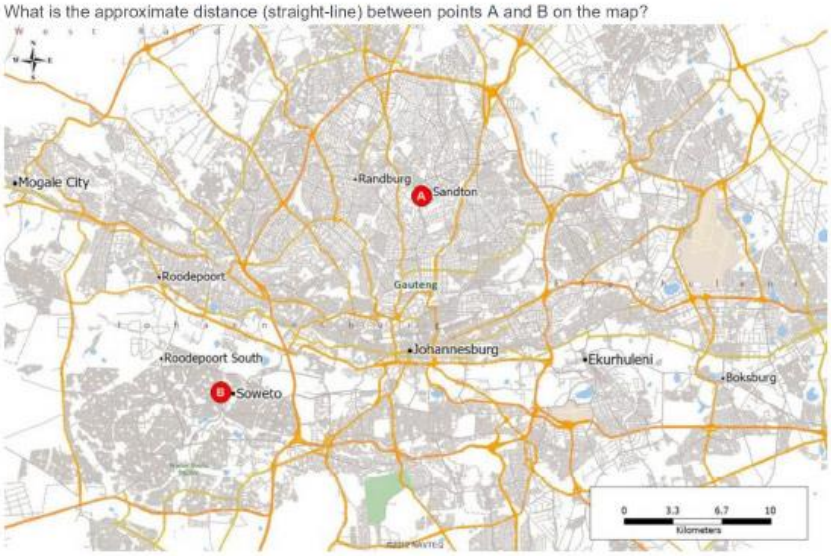
Block3

SECTION 2: MAP READING
Please read the questions carefully before answering.

Block4

Please rate your level of expertise in using:

What is the approximate distance (straight-line) between points A and B on the map?



- ○ 10 Km
- ○ 20 Km
- ○ 30 Km
- ○ 50 Km

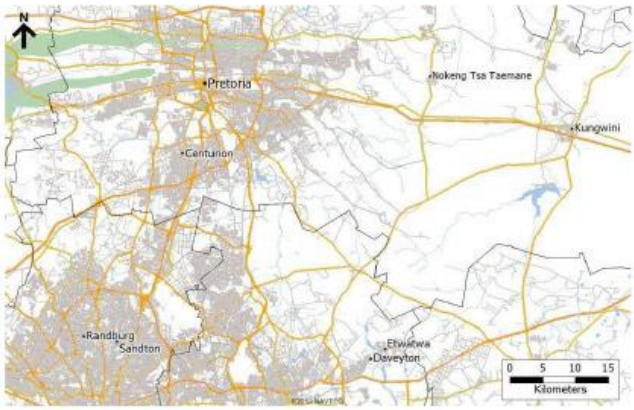
On a scale from 0-10, How confident are you with your answer?

0 1 2 3 4 5 6 7 8 9 10

Confidence level

Block5

If I travel from Pretoria to Kungwini, I will be travelling in a(n) _____ direction.



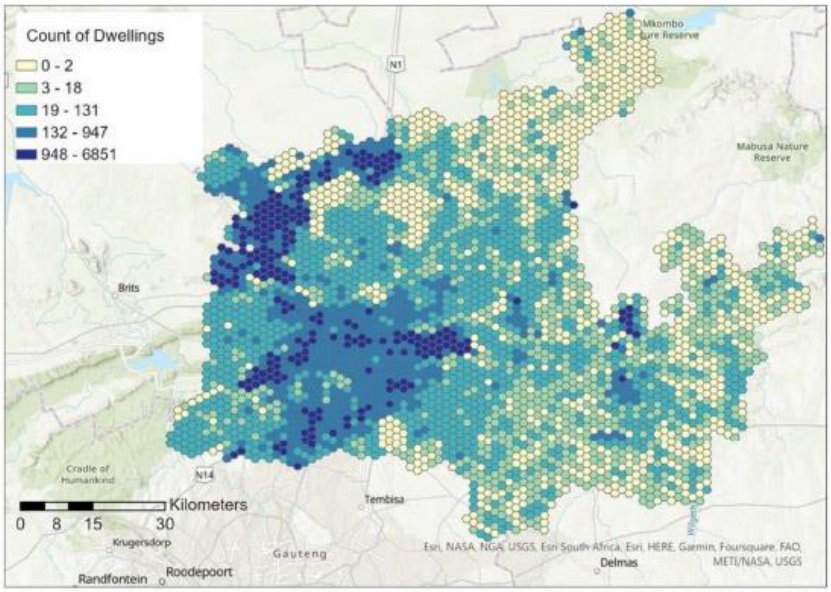
- Northern
- Eastern
- Southern
- Western

Block 54

The next section includes a set of questions specifically related to geographic accessibility. Each question will include a choropleth map where you will need to identify locations (by clicking on the map) about where to to open, close or move service centres. Please read each question carefully.

Slide1

Identify 3 areas on the map where the dwelling count is very low. Click on the relevant areas.



1

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

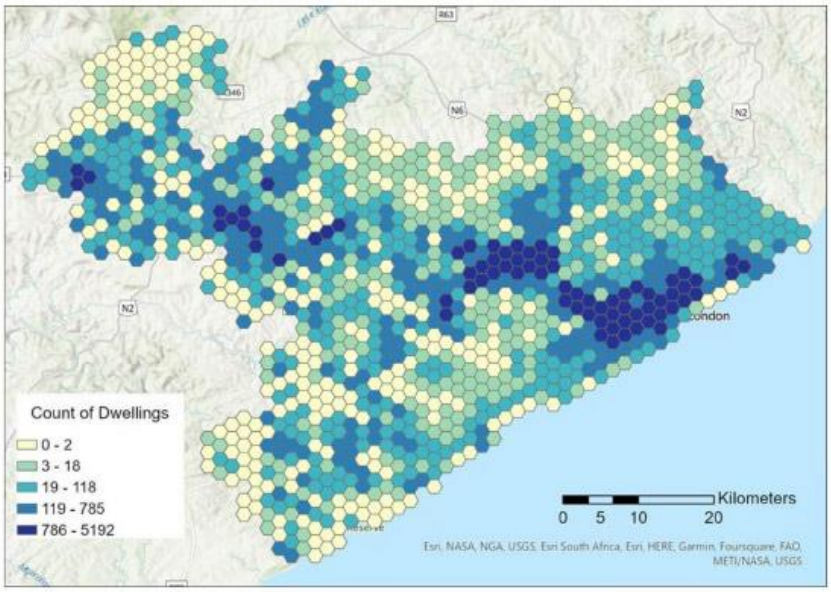
Confidence level

Finding relevant areas on the map was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide2

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster (areas with a high dwelling count). The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



2

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

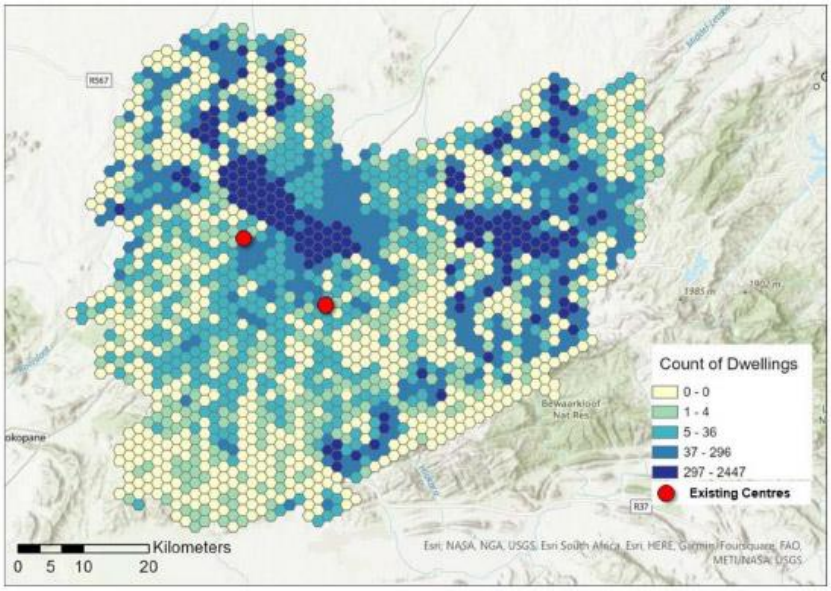
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide3

Identify two locations on the map to add additional service centres. These centres should be located in high density clusters (areas with a high dwelling count) with no other facility nearby (further than 10km away from existing centres).



3

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

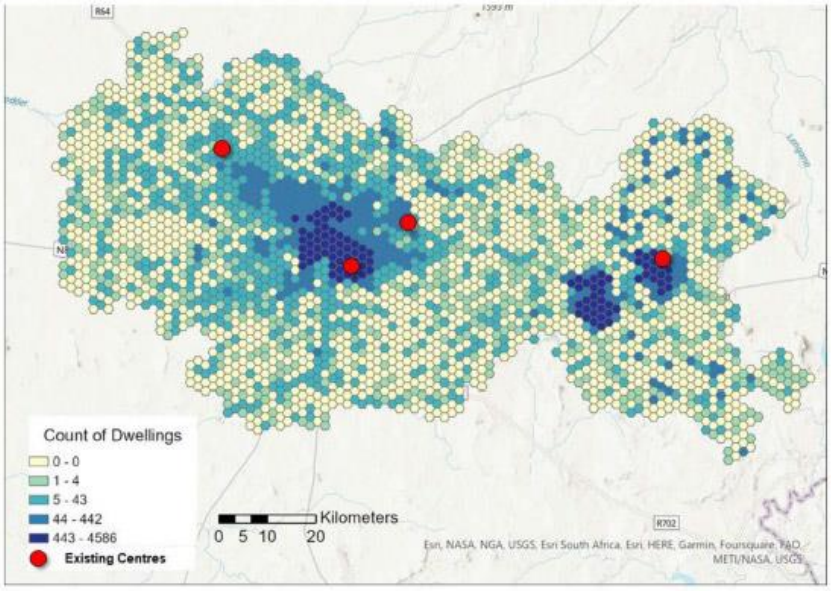
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide4

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high density clusters (areas with a high dwelling count) with no other facility close by (within 10km).



4

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

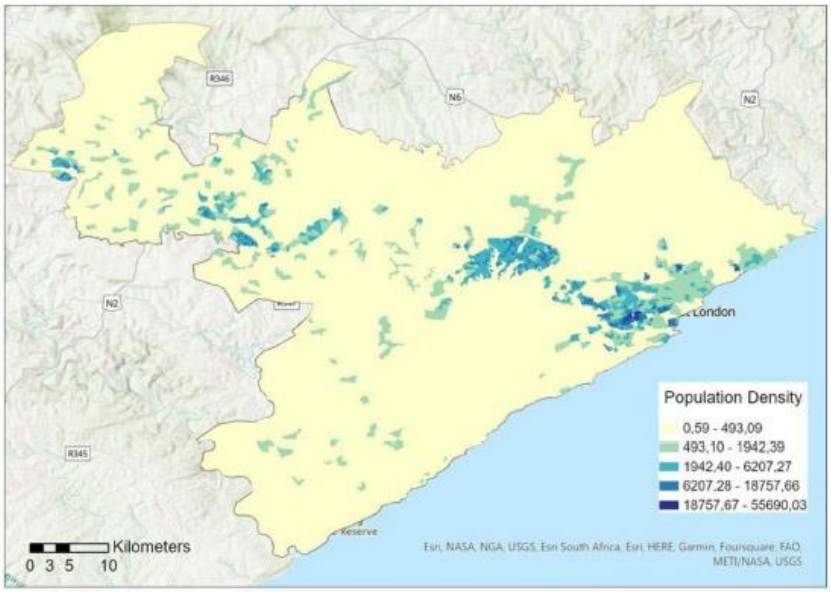
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide5

Identify 3 areas on the map where the population density is very low. Click on the relevant areas.



5

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

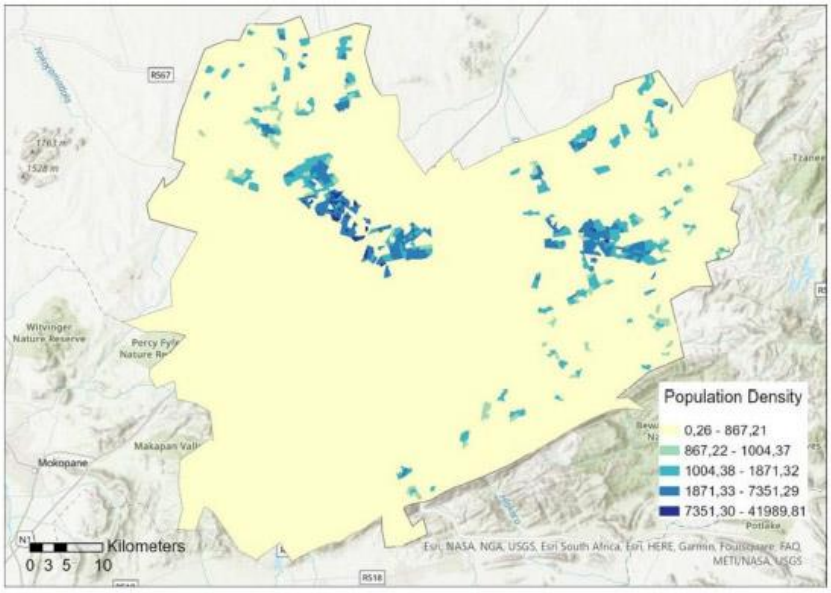
Confidence level

Finding relevant areas on the maps was:

Very Easy | Easy | Neutral | Difficult | Very Difficult

Slide6

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



6

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

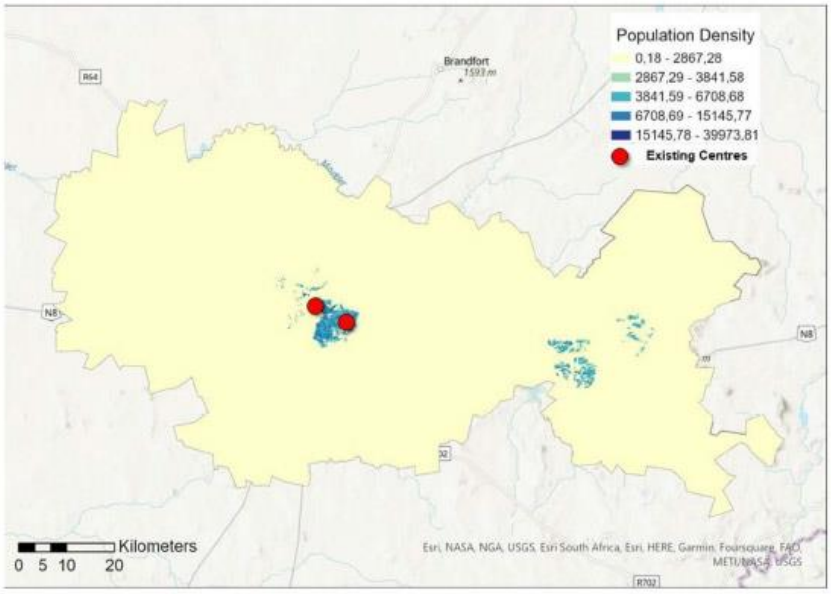
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide7

Identify two locations on the map to add additional service centres. These centres should be located in high population density clusters with no other facility nearby (further than 10km away from existing centres).



7

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

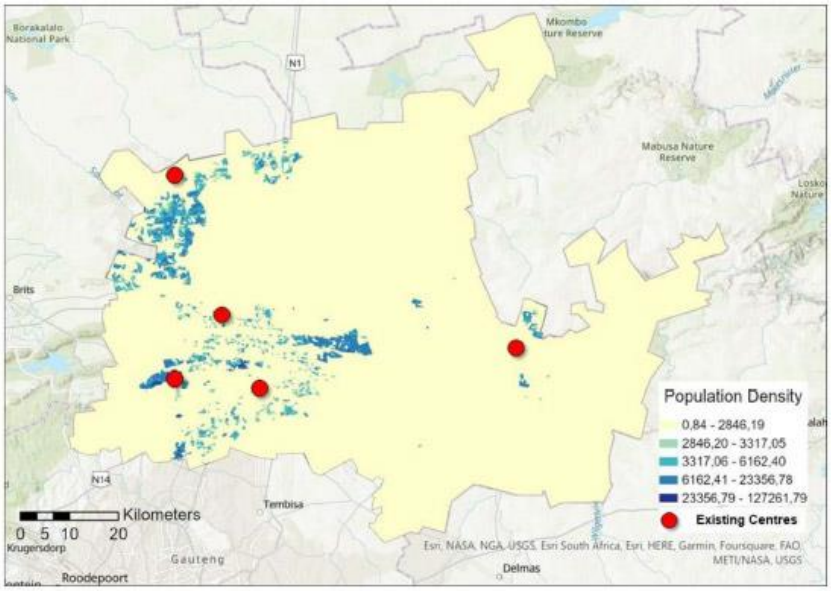
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide8

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high population density clusters with no other facility close by (within 10km).



8

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

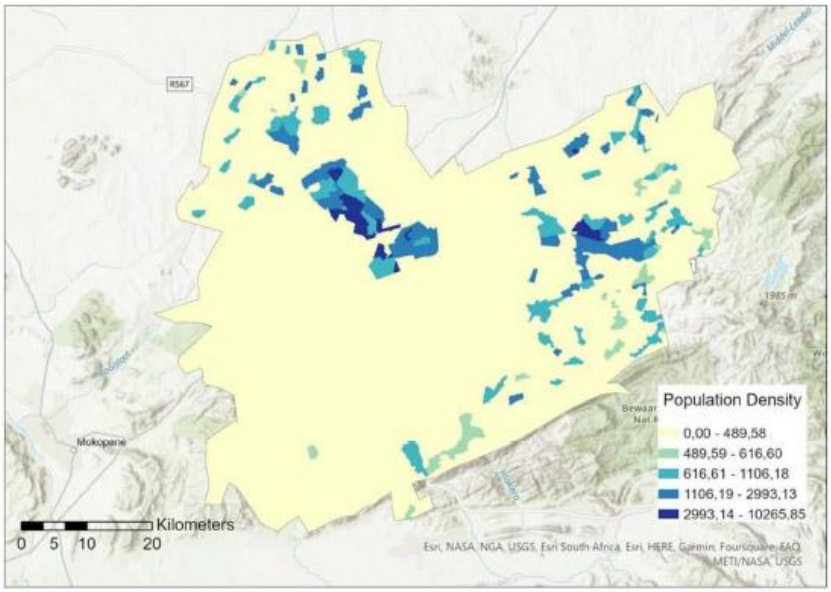
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide9

Identify 3 areas on the map where the population density is very low. Click on the relevant areas.



9

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

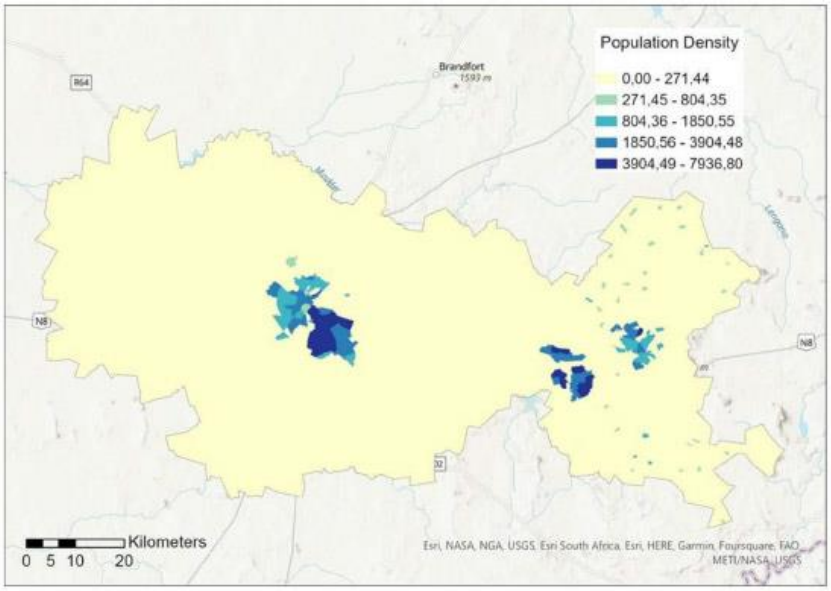
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide10

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



10

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

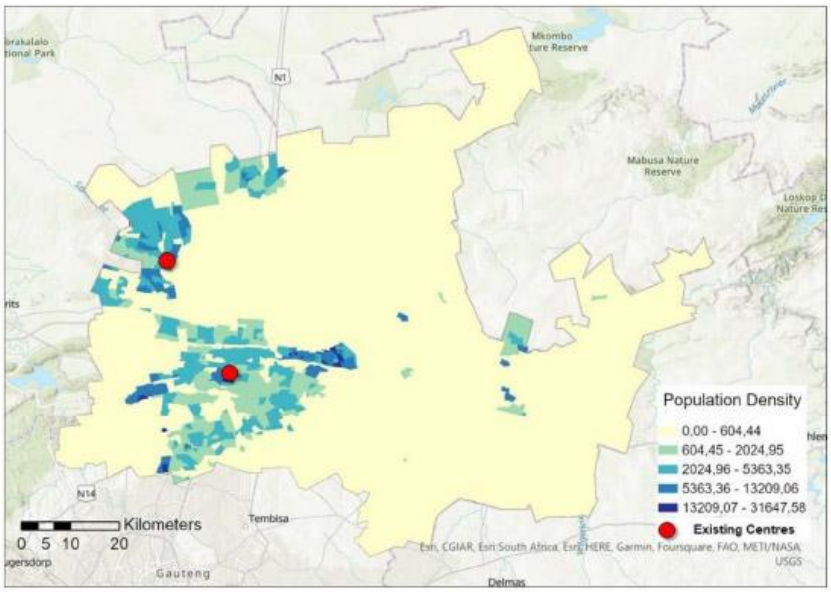
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide11

Identify two locations on the map to add additional service centres. These centres should be located in high population density clusters with no other facility nearby (further than 10km away from existing centres).



11

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

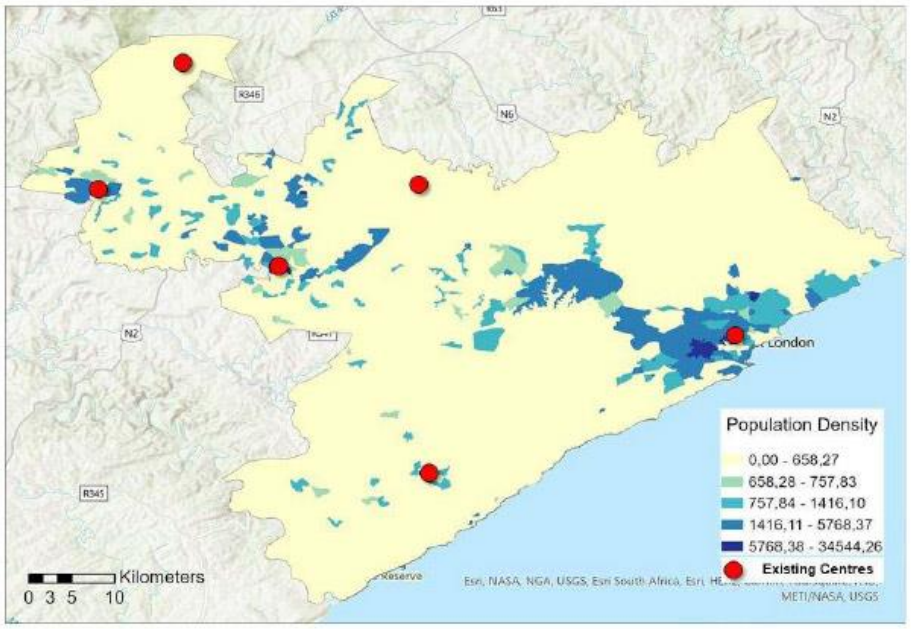
Confidence level

Finding relevant areas on the maps was:

Very Easy | Easy | Neutral | Difficult | Very Difficult

Slide12

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high population density clusters with no other facility close by (within 10km).



12

On a scale from 0-10, How confident are you with your answer of the previous question?

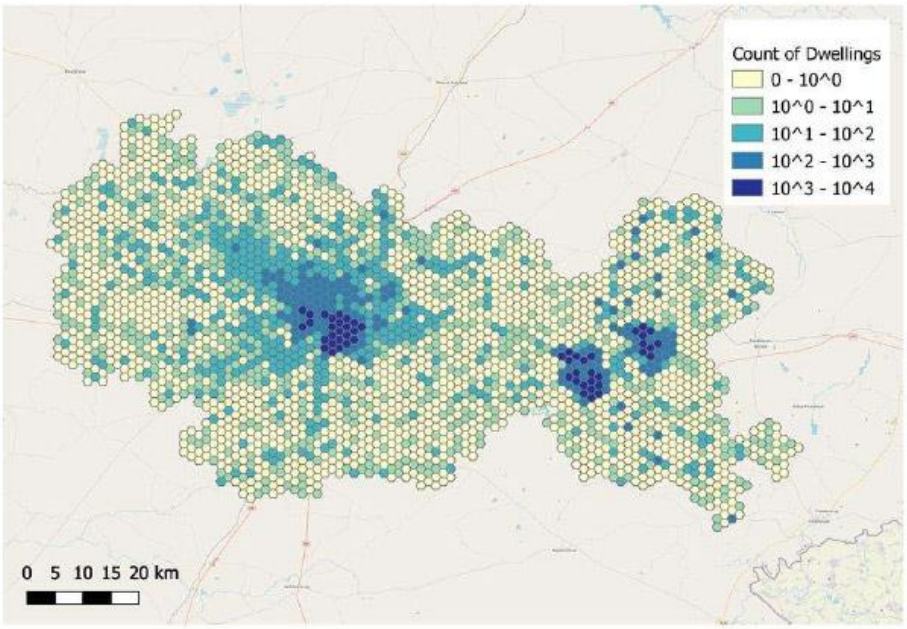
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide13

Identify 3 areas on the map where the dwelling count is very low. Click on the relevant areas.



13

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

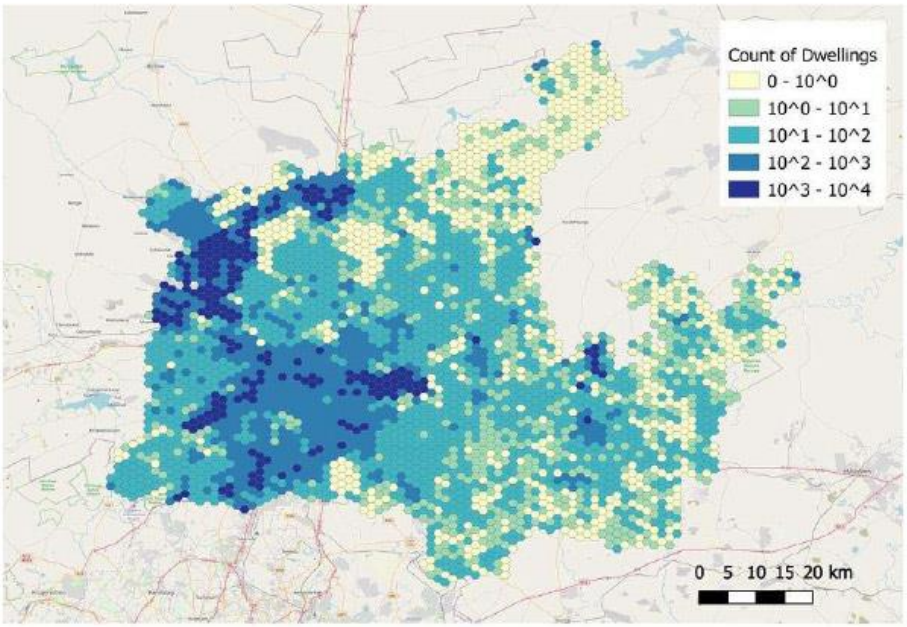
Confidence level

Finding relevant areas on the maps was:

Very Easy ○ Easy ○ Neutral ○ Difficult ○ Very Difficult ○

Slide14

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster (areas with a high dwelling count). The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



14

On a scale from 0-10, How confident are you with your answer of the previous question?

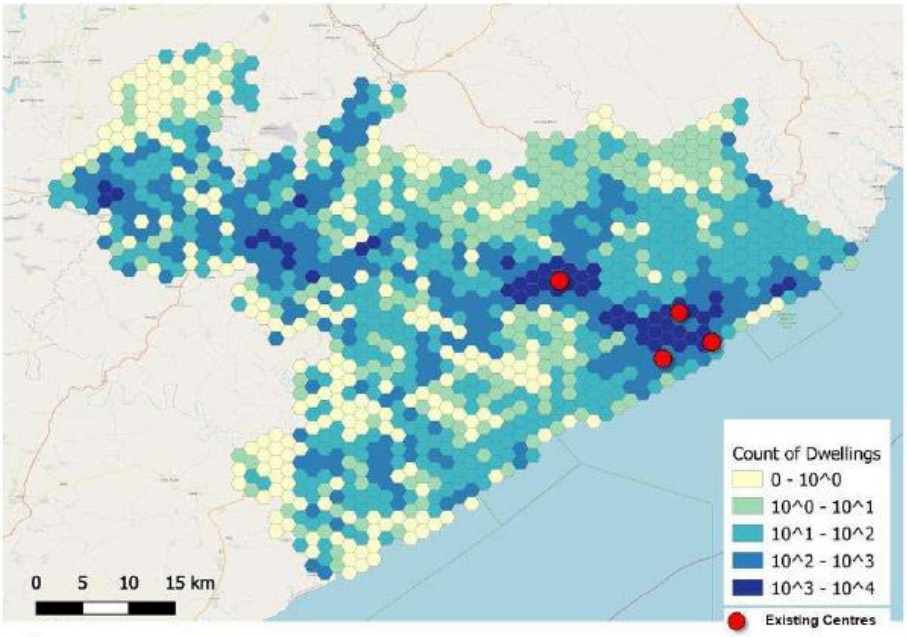
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide15

Identify two locations on the map to add additional service centres. These centres should be located in high density clusters (areas with a high dwelling count) with no other facility nearby (further than 10km away from existing centres).



15

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

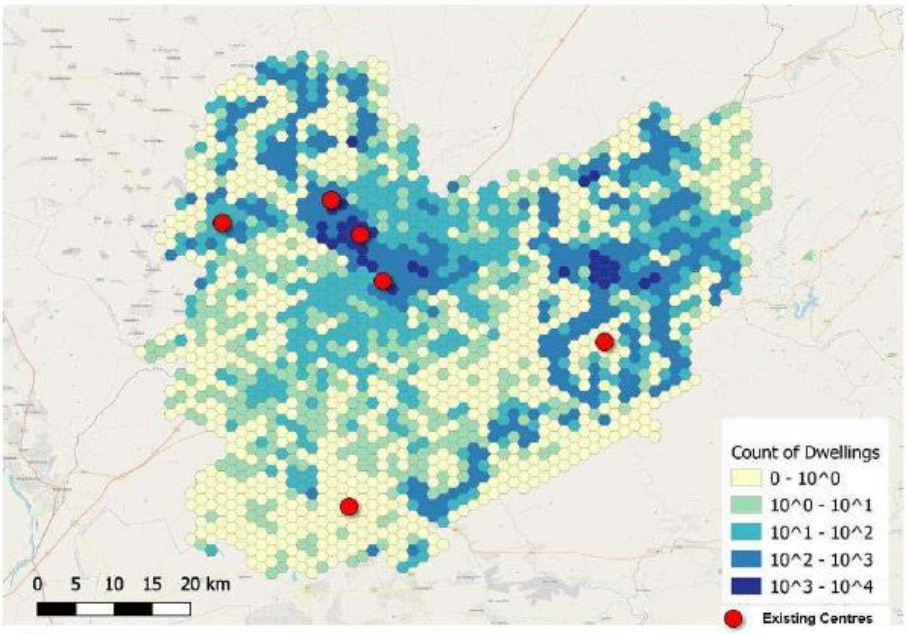
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide16

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high density clusters (areas with a high dwelling count) with no other facility close by (within 10km).



16

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

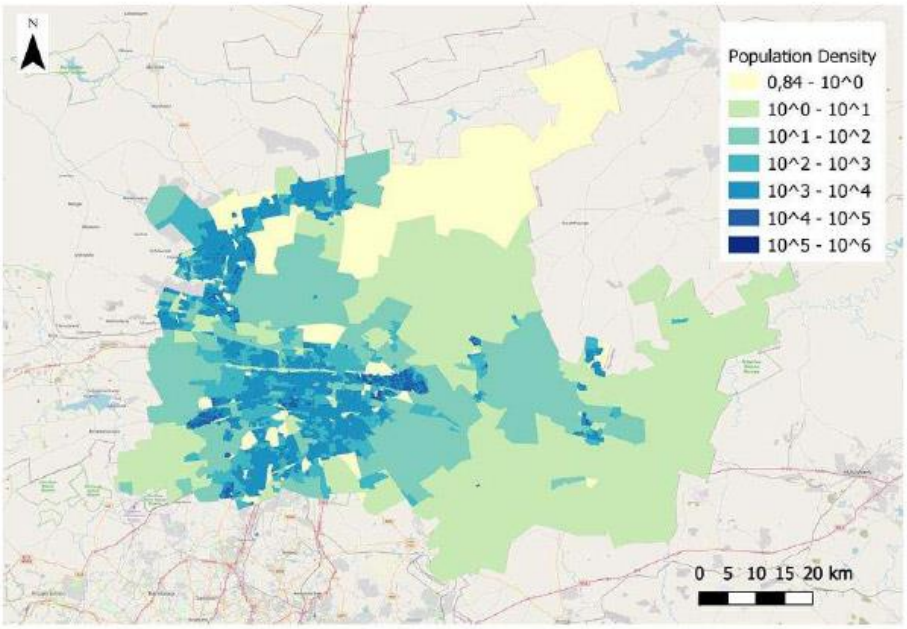
Confidence Level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide17

Identify 3 areas on the map where the population density is very low. Click on the relevant areas.



17

On a scale from 0-10, How confident are you with your answer of the previous question?

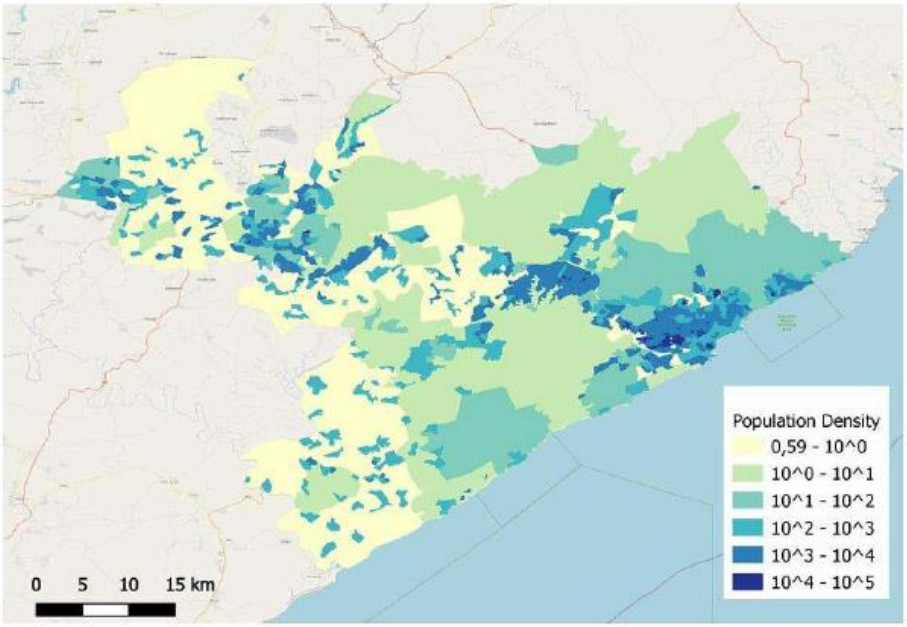
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide18

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



18

On a scale from 0-10, How confident are you with your answer of the previous question?

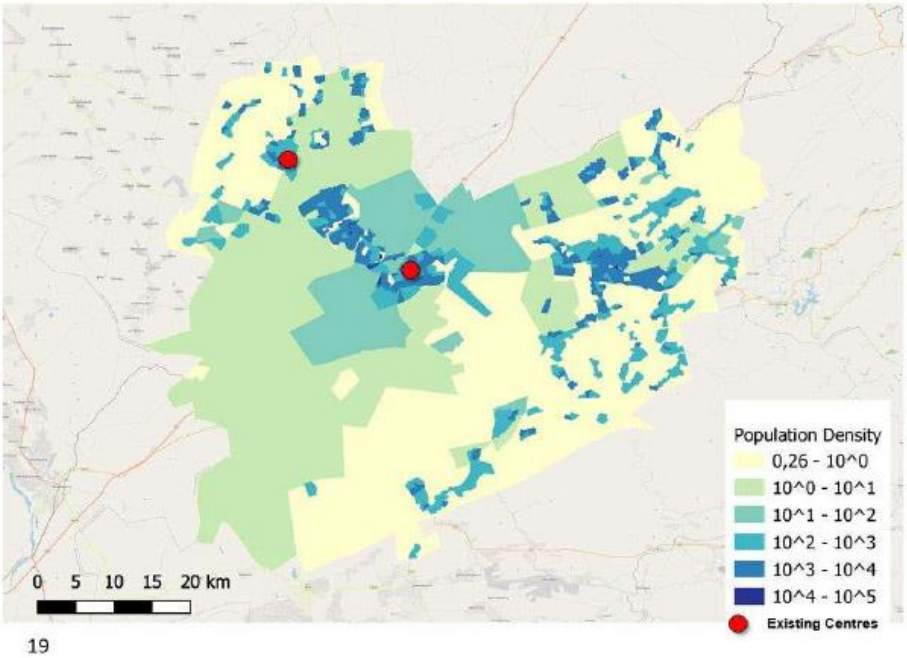
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide19

Identify two locations on the map to add additional service centres. These centres should be located in high population density clusters with no other facility nearby (further than 10km away from existing centres).



19

On a scale from 0-10, How confident are you with your answer of the previous question?

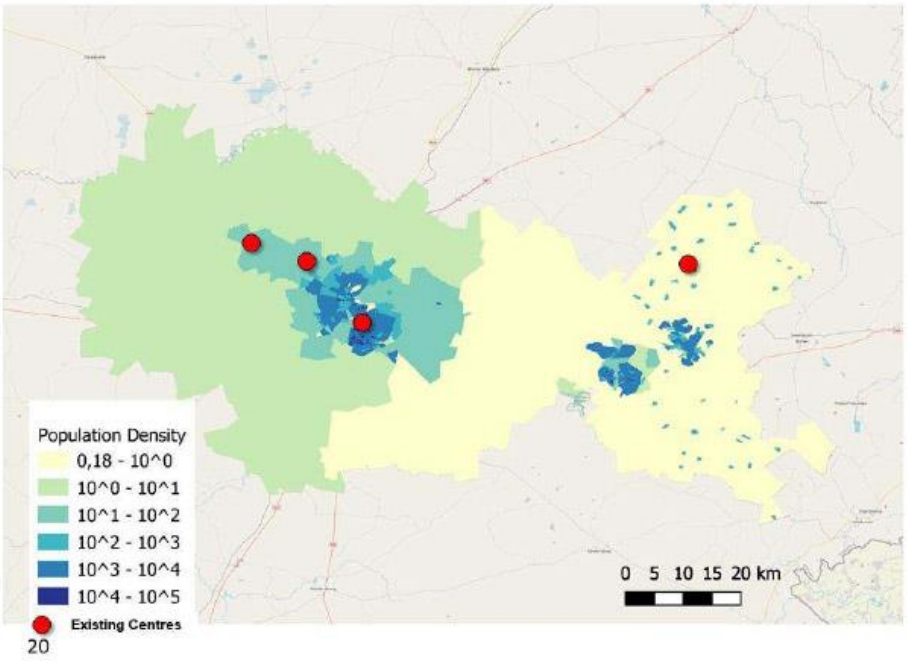
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide20

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high population density clusters with no other facility close by (within 10km).



20

On a scale from 0-10, How confident are you with your answer of the previous question?

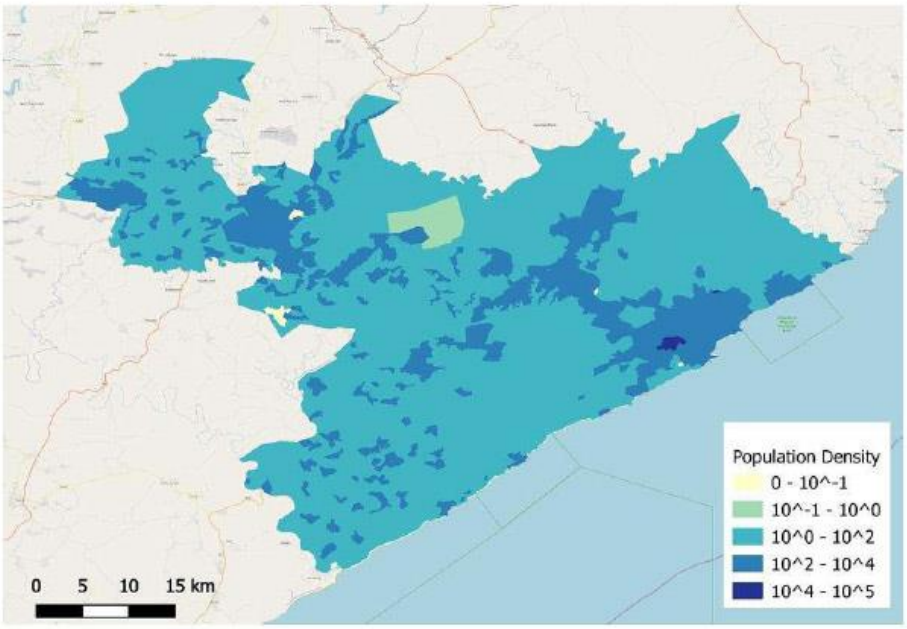
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide21

Identify 3 areas on the map where the population density is very low. Click on the relevant areas.



21

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

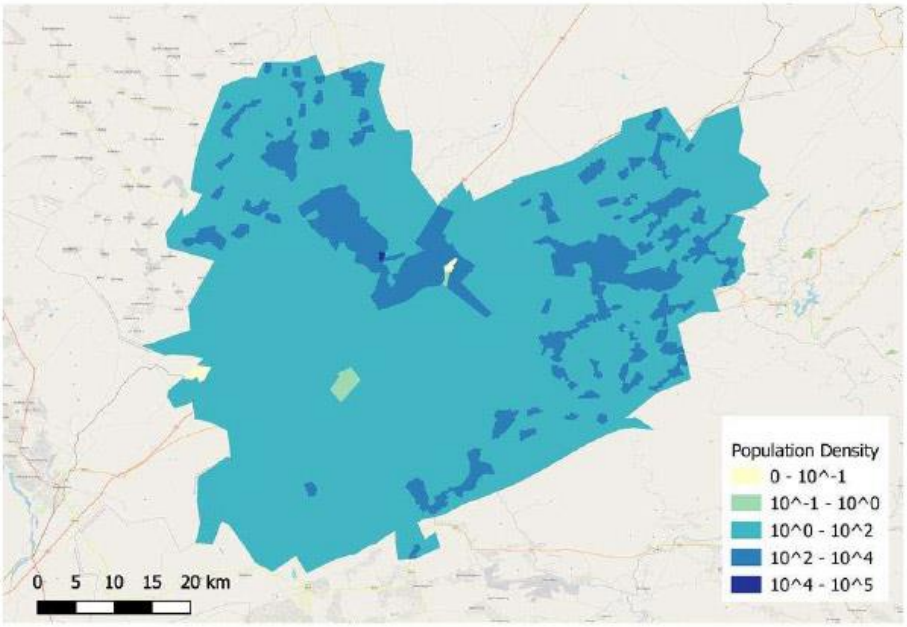
Confidence level

Finding relevant areas on the maps was:

Very Easy ○ Easy ○ Neutral ○ Difficult ○ Very Difficult ○

Slide22

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



22

On a scale from 0-10, How confident are you with your answer of the previous question?

0 1 2 3 4 5 6 7 8 9 10

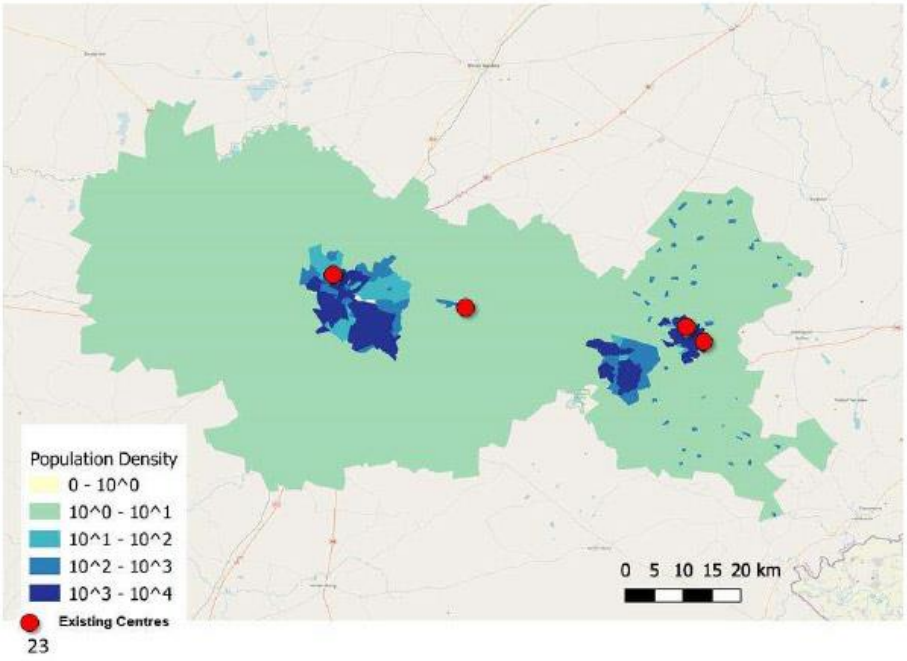
Confidence level

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide23

Identify two locations on the map to add additional service centres. These centres should be located in high population density clusters with no other facility nearby (further than 10km away from existing centres).



On a scale from 0-10, How confident are you with your answer of the previous question?

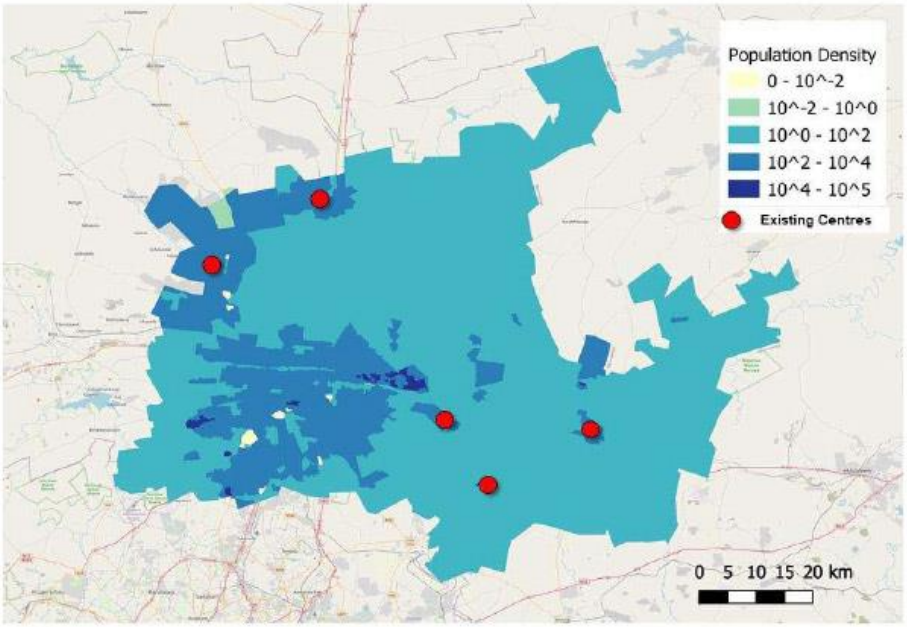
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide24

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high population density clusters with no other facility close by (within 10km).



24

On a scale from 0-10, How confident are you with your answer of the previous question?

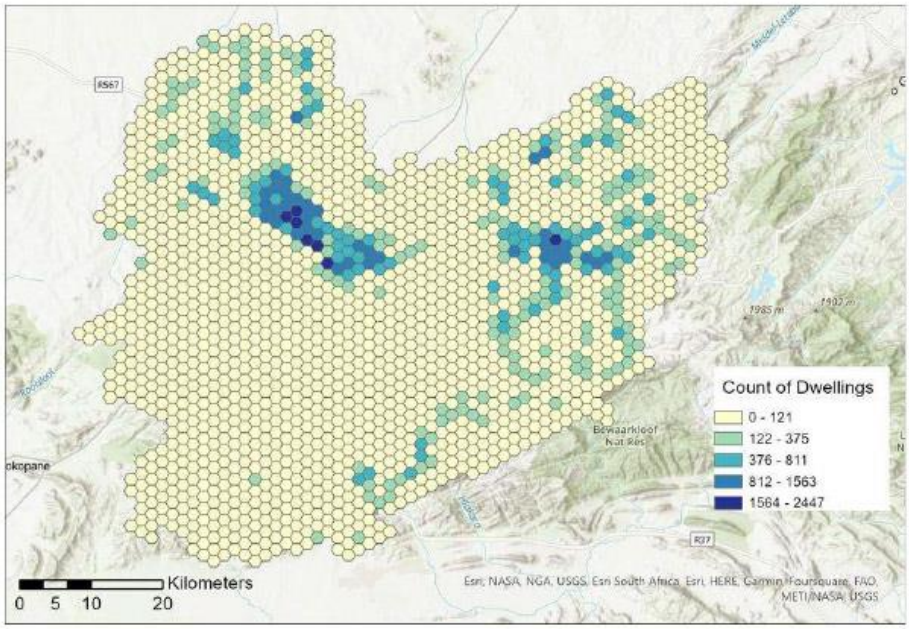
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide25

Identify 3 areas on the map where the dwelling count is very low. Click on the relevant areas.



25

On a scale from 0-10, How confident are you with your answer of the previous question?

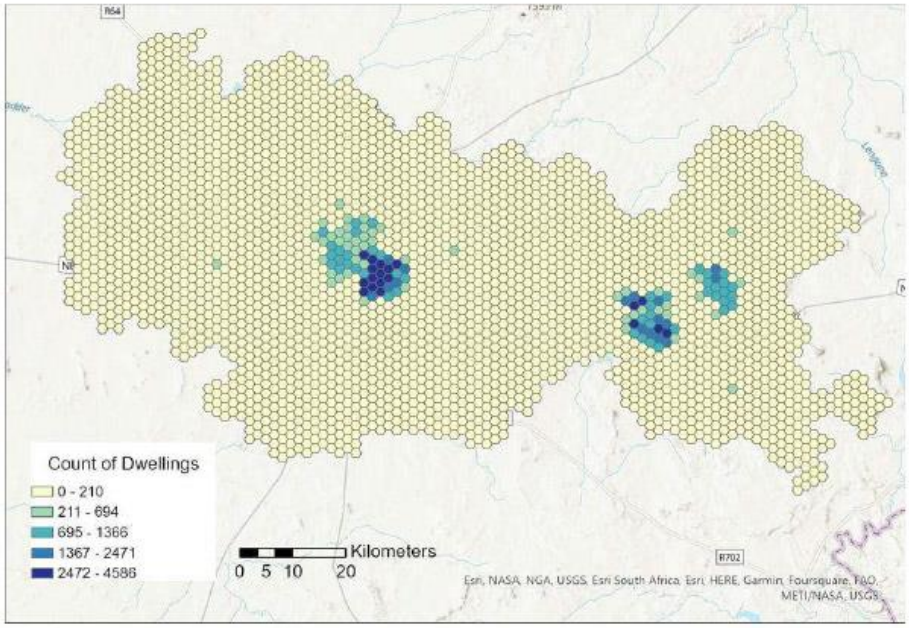
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide26

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster (areas with a high dwelling count). The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



26

On a scale from 0-10, How confident are you with your answer of the previous question?

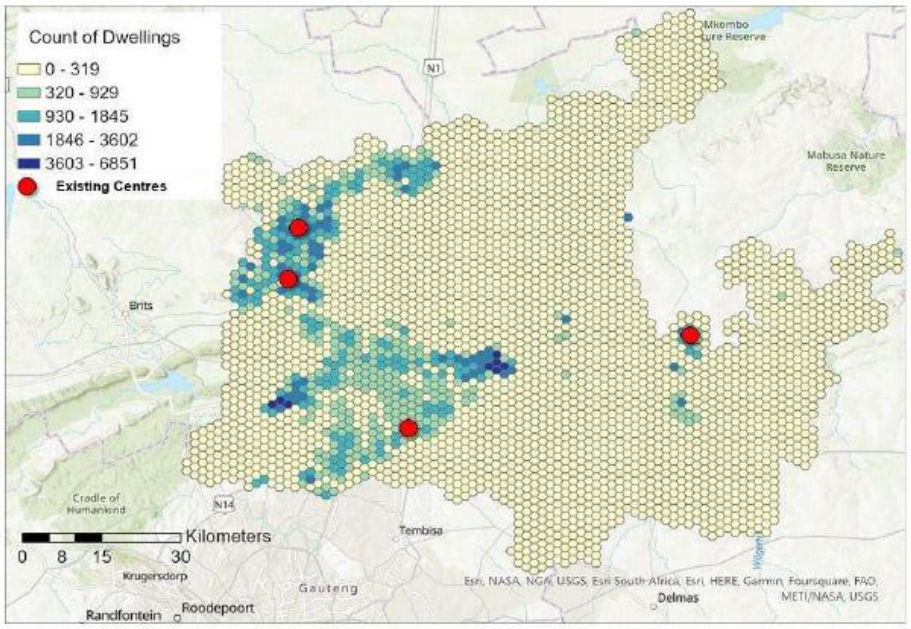
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide27

Identify two locations on the map to add additional service centres. These centres should be located in high density clusters (areas with a high dwelling count) with no other facility nearby (further than 10km away from existing centres).



27

On a scale from 0-10, How confident are you with your answer of the previous question?

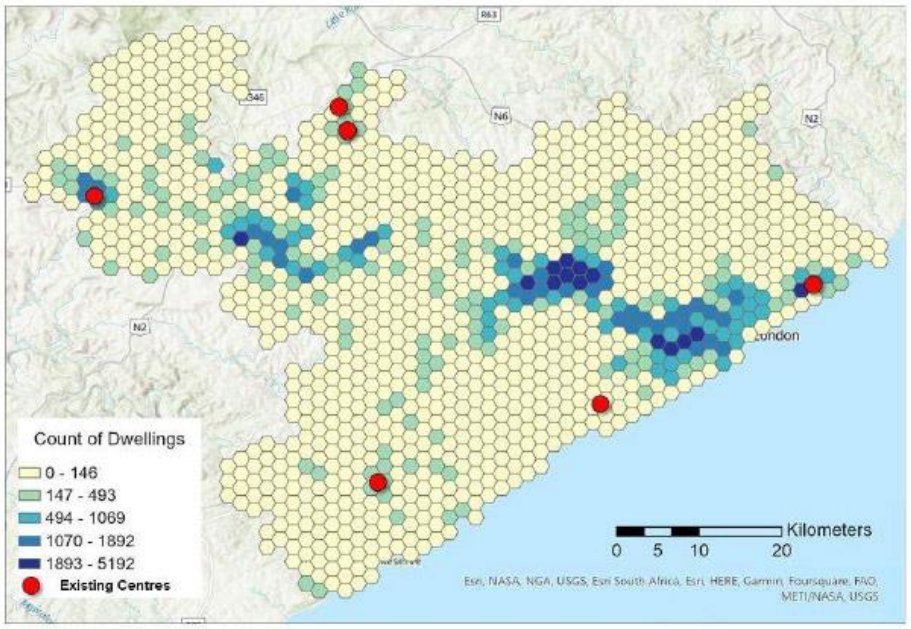
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide28

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high density clusters (areas with a high dwelling count) with no other facility close by (within 10km).



28

On a scale from 0-10, How confident are you with your answer of the previous question?

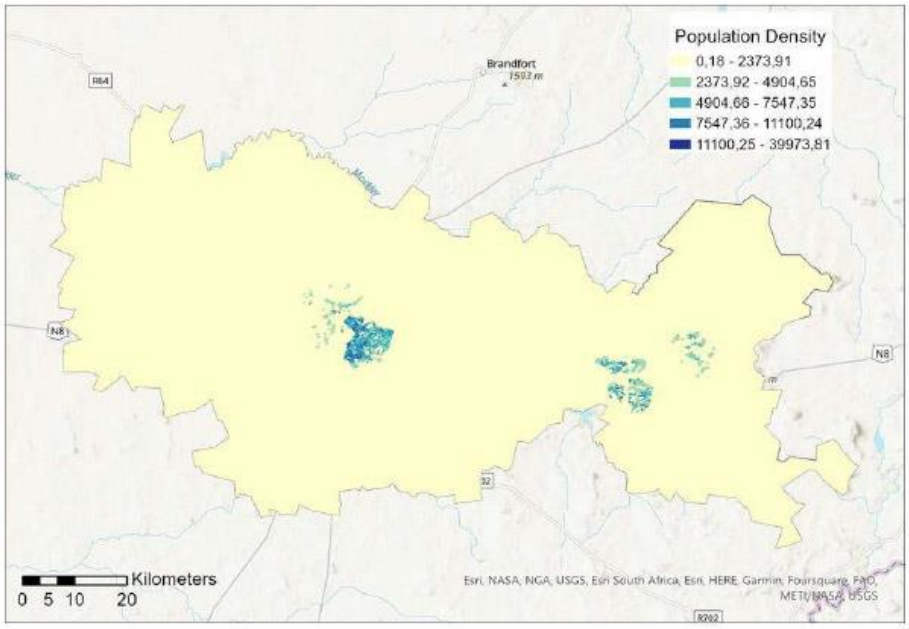
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide29

Identify 3 areas on the map where the population density is very low. Click on the relevant areas.



29

On a scale from 0-10, How confident are you with your answer of the previous question?

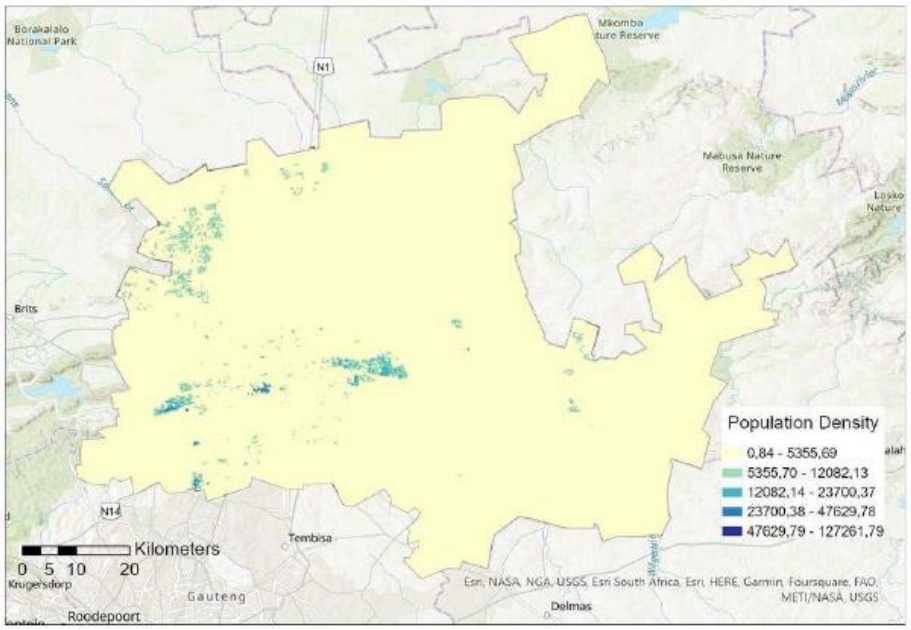
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide30

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



30

On a scale from 0-10, How confident are you with your answer of the previous question?

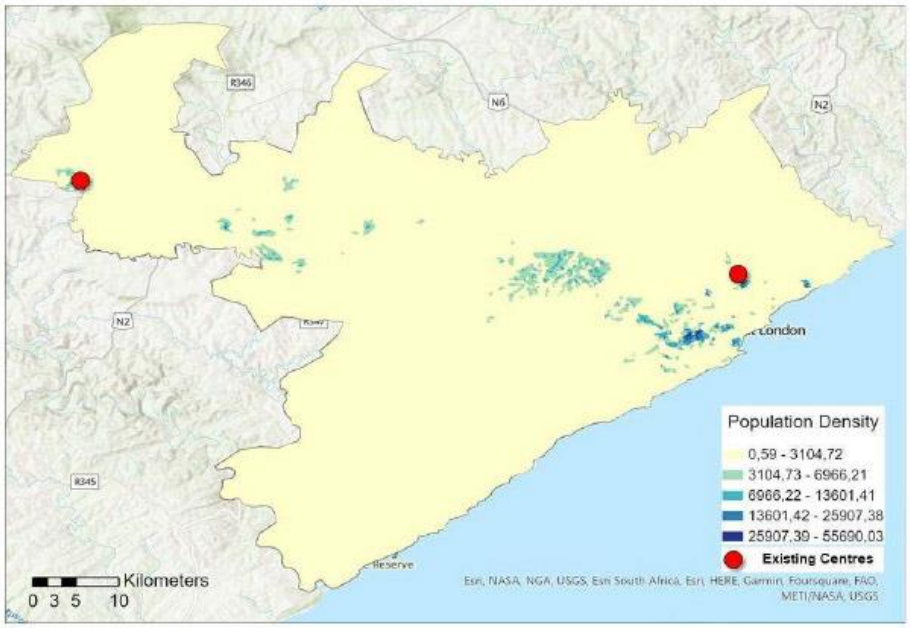
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide31

Identify two locations on the map to add additional service centres. These centres should be located in high population density clusters with no other facility nearby (further than 10km away from existing centres).



31

On a scale from 0-10, How confident are you with your answer of the previous question?

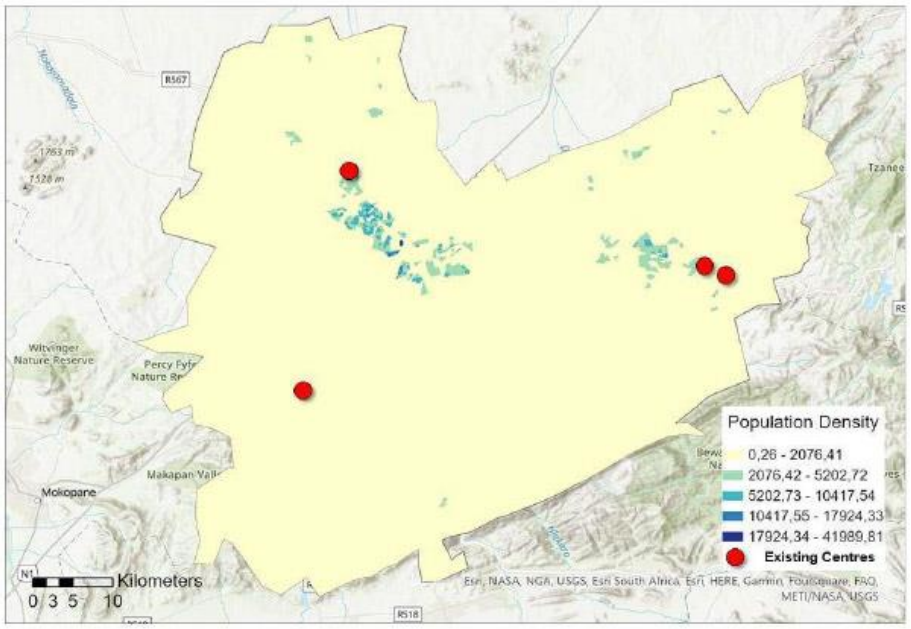
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide32

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high population density clusters with no other facility close by (within 10km).



32

On a scale from 0-10, How confident are you with your answer of the previous question?

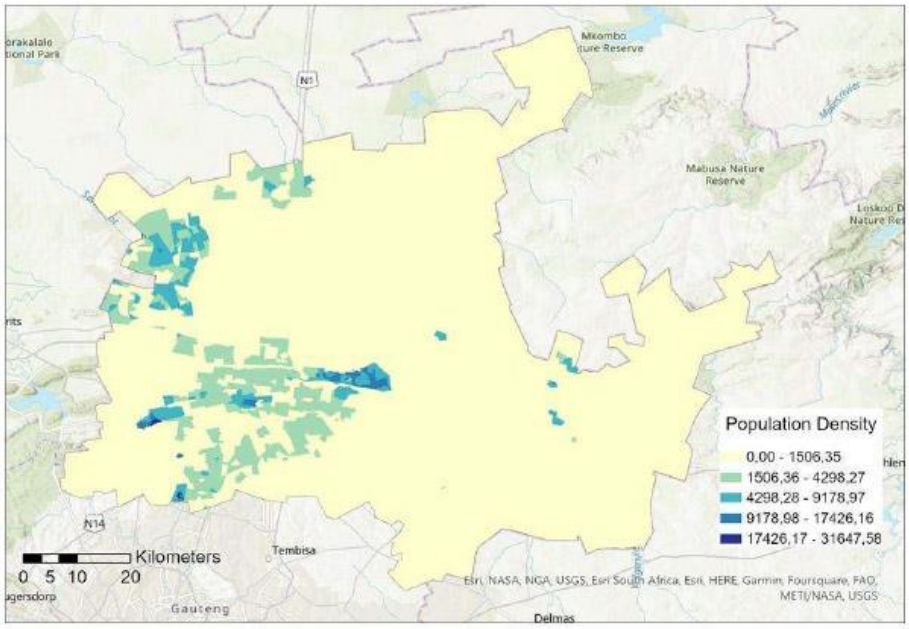
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide33

Identify 3 areas on the map where the population density is very low. Click on the relevant areas.



33

On a scale from 0-10, How confident are you with your answer of the previous question?

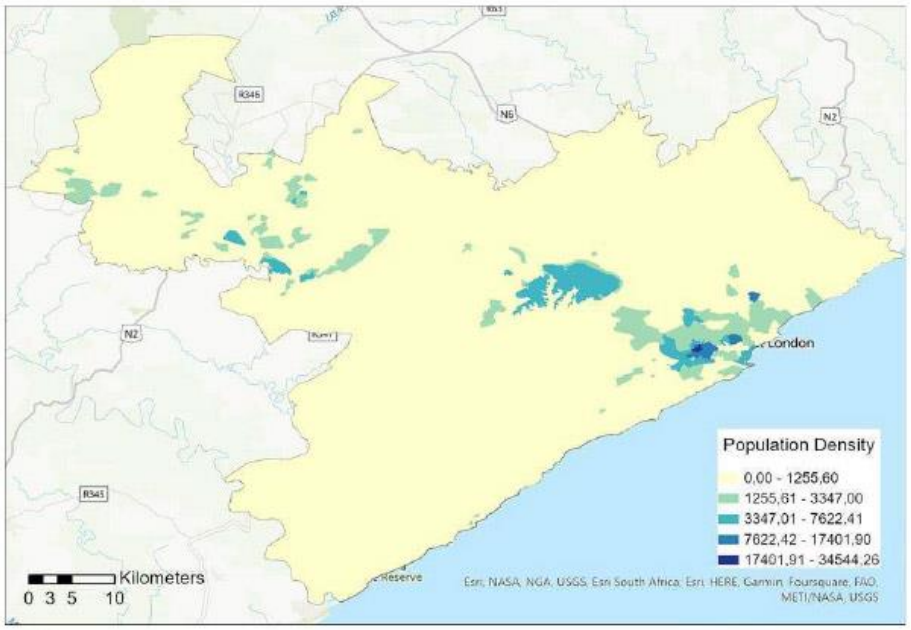
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide34

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



34

On a scale from 0-10, How confident are you with your answer of the previous question?

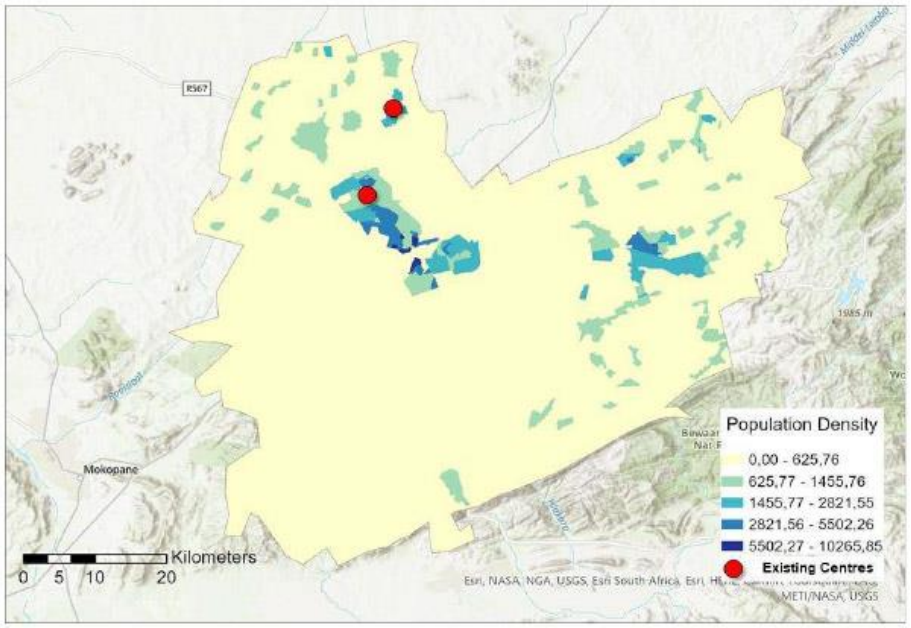
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide35

Identify two locations on the map to add additional service centres. These centres should be located in high population density clusters with no other facility nearby (further than 10km away from existing centres).



35

On a scale from 0-10, How confident are you with your answer of the previous question?

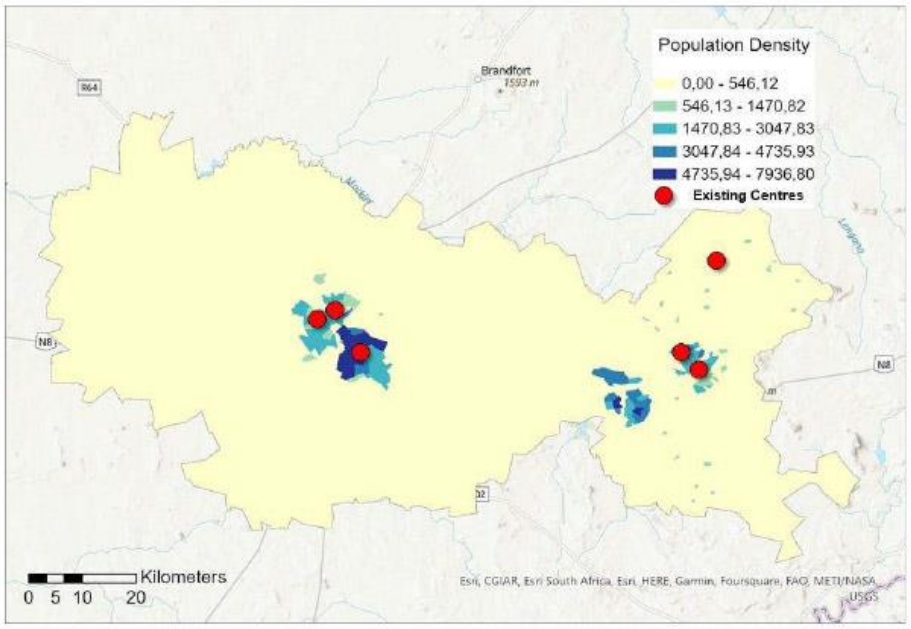
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide36

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high population density clusters with no other facility close by (within 10km).



36

On a scale from 0-10, How confident are you with your answer of the previous question?

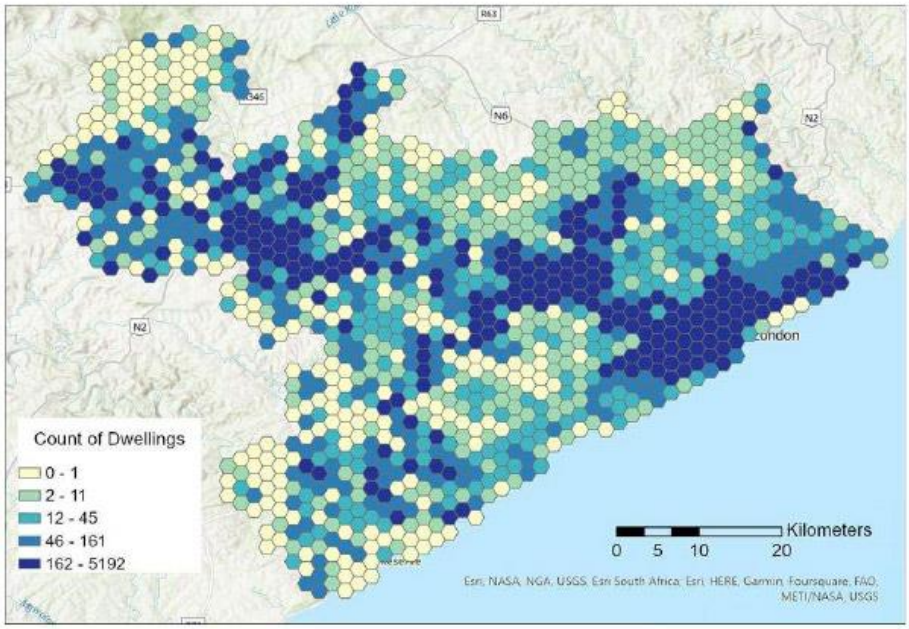
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide37

Identify 3 areas on the map where the dwelling count is very low. Click on the relevant areas.



37

On a scale from 0-10, How confident are you with your answer of the previous question?

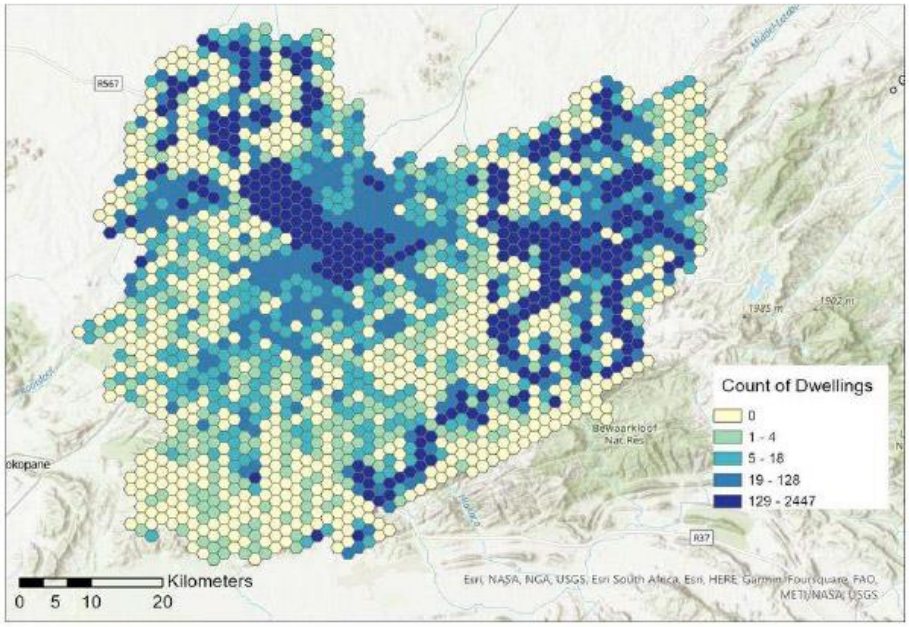
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide38

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster (areas with a high dwelling count). The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



38

On a scale from 0-10, How confident are you with your answer of the previous question?

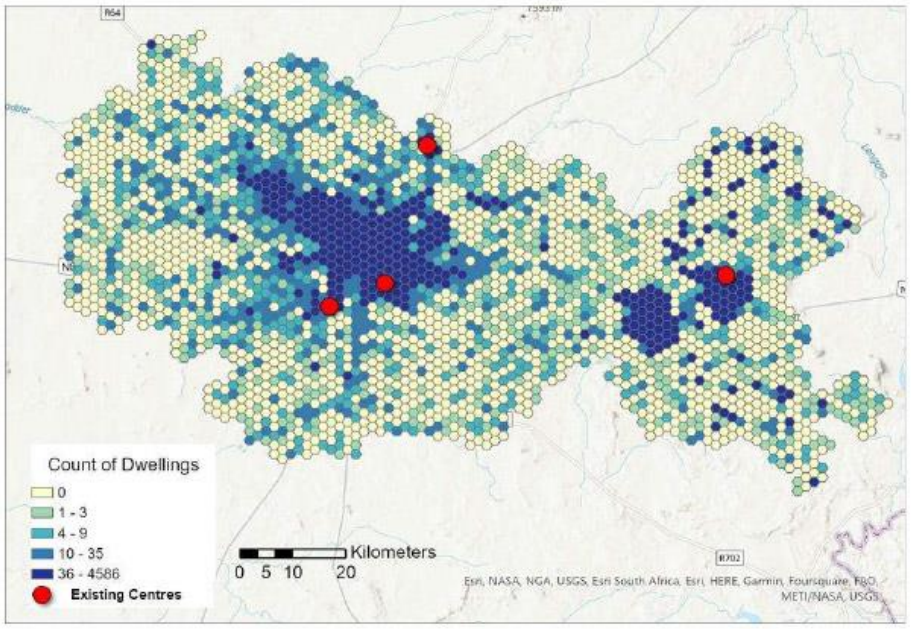
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide39

Identify two locations on the map to add additional service centres. These centres should be located in high density clusters (areas with a high dwelling count) with no other facility nearby (further than 10km away from existing centres).



39

On a scale from 0-10, How confident are you with your answer of the previous question?

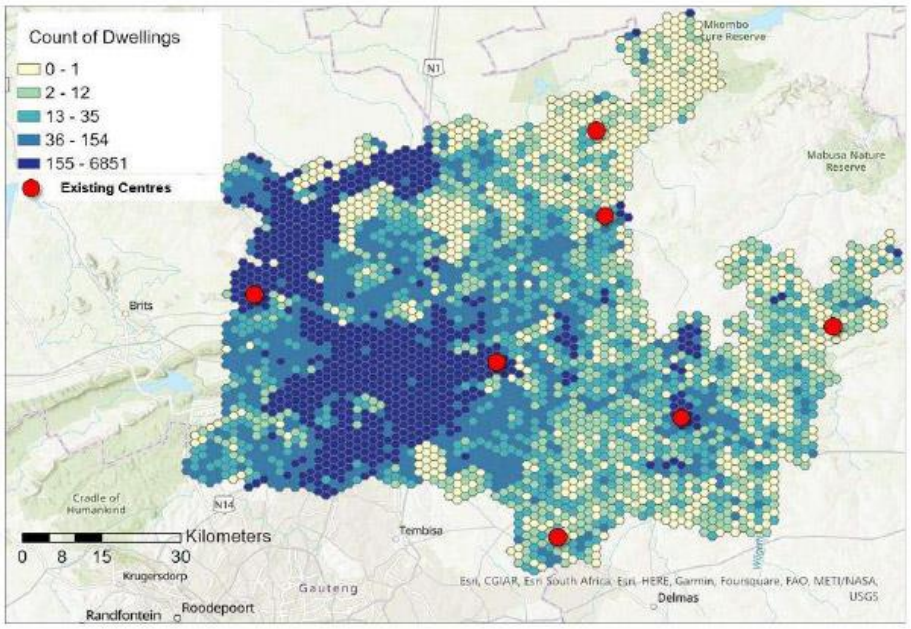
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide40

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high density clusters (areas with a high dwelling count) with no other facility close by (within 10km).



40

On a scale from 0-10, How confident are you with your answer of the previous question?

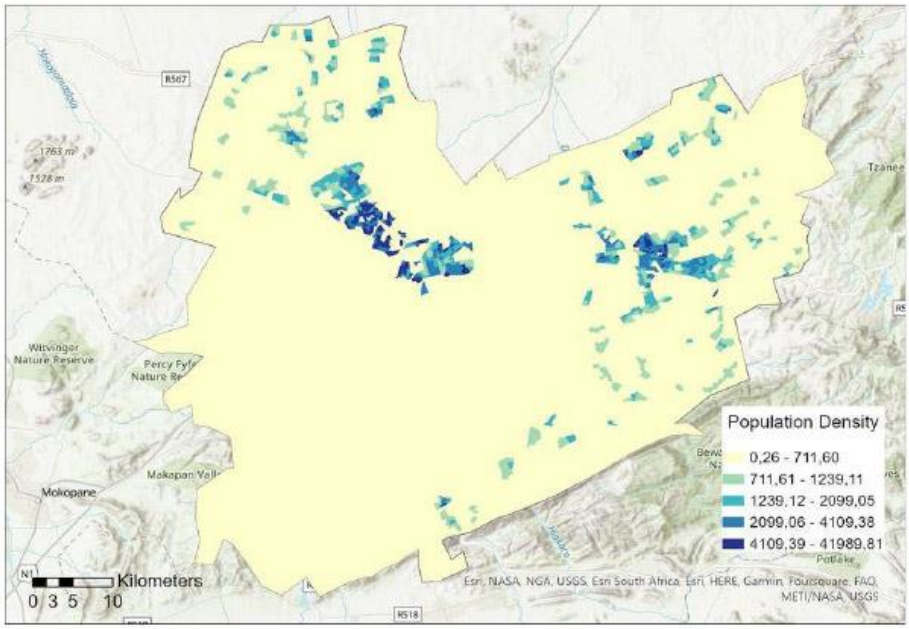
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide41

Identify 3 areas on the map where the population density is very low. Click on the relevant areas.



41

On a scale from 0-10, How confident are you with your answer of the previous question?

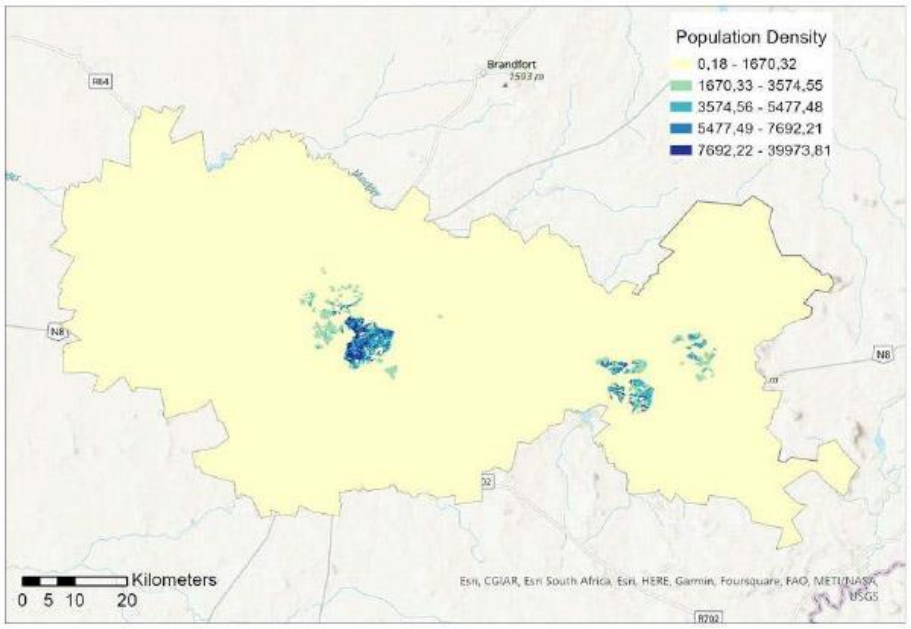
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide42

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



42

On a scale from 0-10, How confident are you with your answer of the previous question?

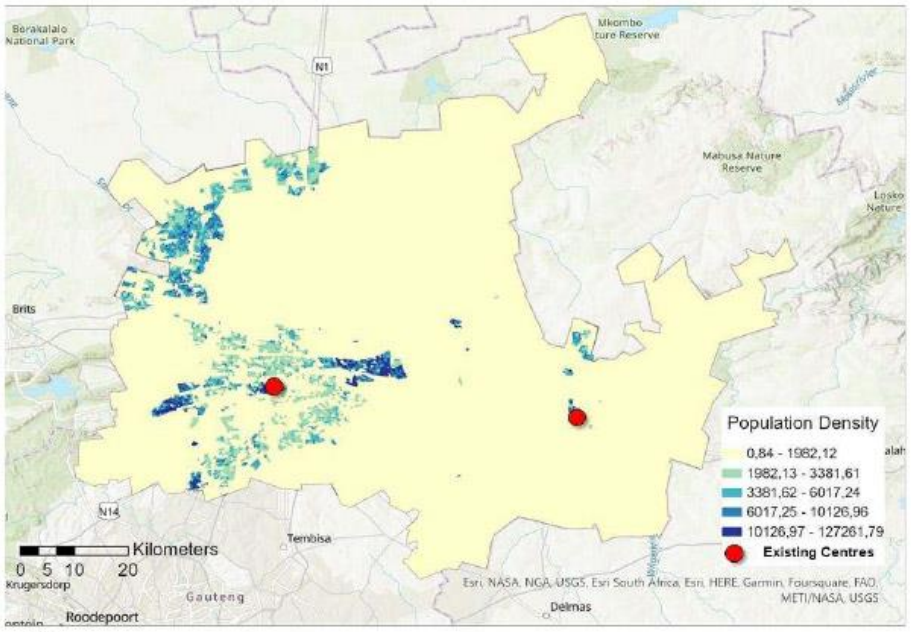
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide43

Identify two locations on the map to add additional service centres. These centres should be located in high population density clusters with no other facility nearby (further than 10km away from existing centres).



43

On a scale from 0-10, How confident are you with your answer of the previous question?

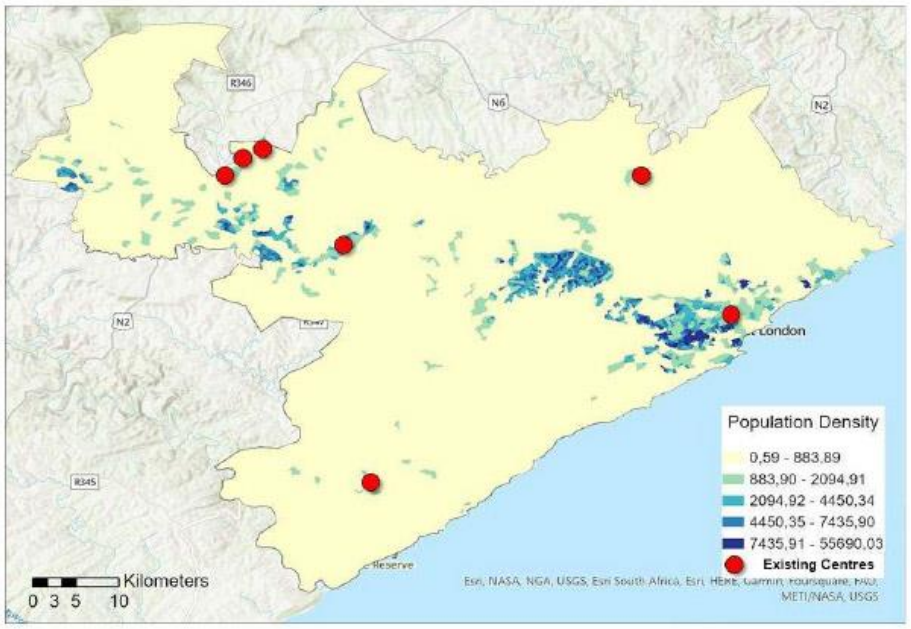
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide44

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high population density clusters with no other facility close by (within 10km).



44

On a scale from 0-10, How confident are you with your answer of the previous question?

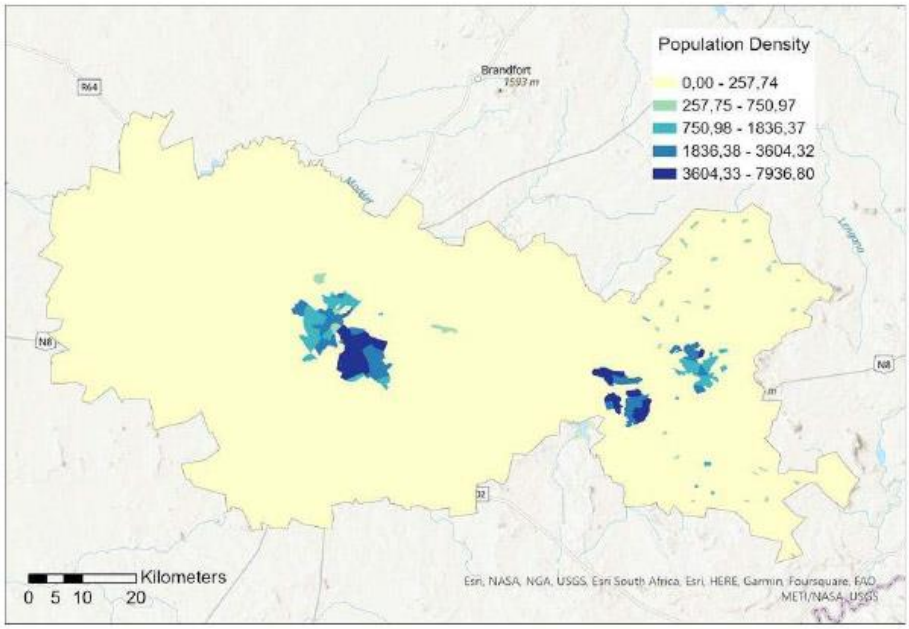
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide45

Identify 3 areas on the map where the population density is very low. Click on the relevant areas.



45

On a scale from 0-10, How confident are you with your answer of the previous question?

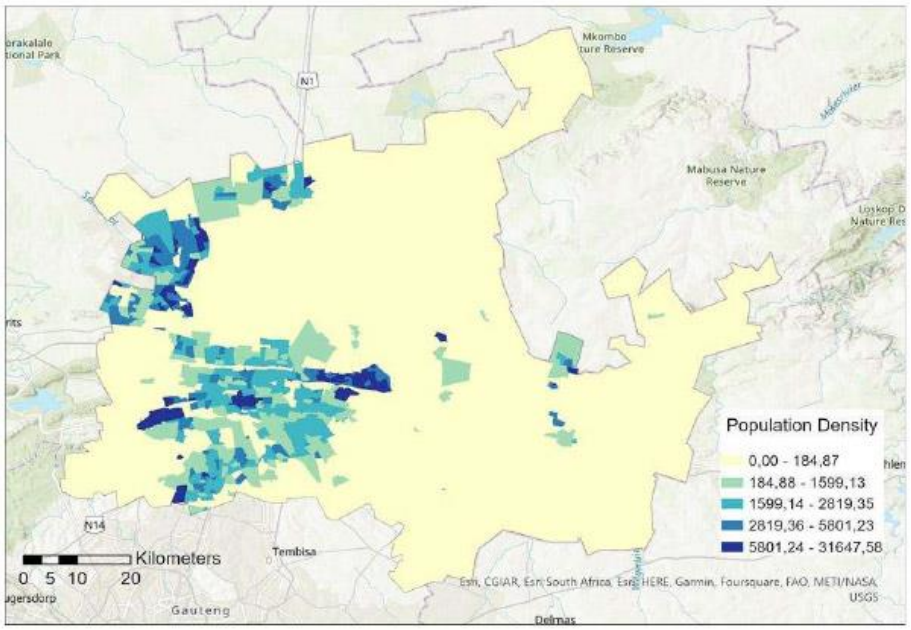
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide46

Identify the best area for opening a new service centre. Ideally the new centre should be located in a densely populated cluster. The maximum threshold (catchment area of a service centre) is 10km, i.e. a centre serves only people within 10km from it. Anybody living further away is not served by that centre.



46

On a scale from 0-10, How confident are you with your answer of the previous question?

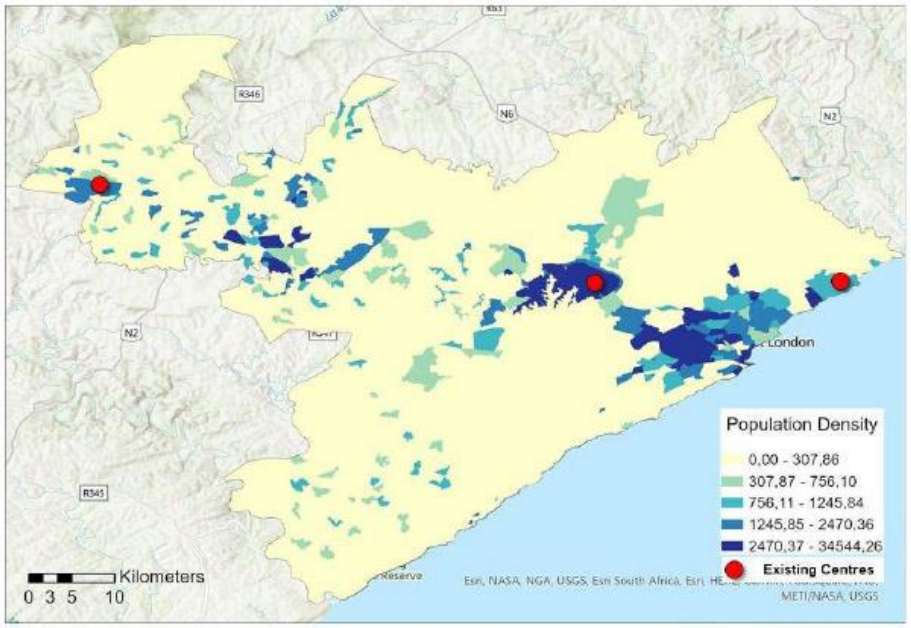
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide47

Identify two locations on the map to add additional service centres. These centres should be located in high population density clusters with no other facility nearby (further than 10km away from existing centres).



47

On a scale from 0-10, How confident are you with your answer of the previous question?

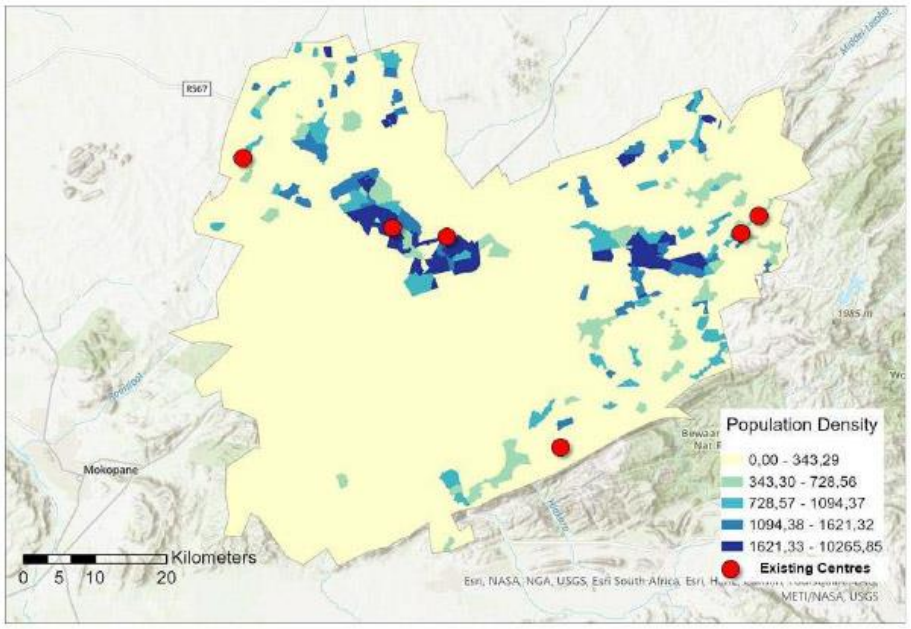
	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Slide48

Identify two service centres that are incorrectly placed and should be relocated to different locations. Remember the ideal location would be high population density clusters with no other facility close by (within 10km).



48

On a scale from 0-10, How confident are you with your answer of the previous question?

	0	1	2	3	4	5	6	7	8	9	10
Confidence level											

Finding relevant areas on the maps was:

Very Easy	Easy	Neutral	Difficult	Very Difficult
○	○	○	○	○

Block 10

Block 55

Thank you for participating in this survey. Please let me know if you have any comments or recommendations.

Powered by Qualtrics

APPENDIX D: R SCRIPTS

This appendix includes the R scripts for calculating the goodness of variance fit for each data classification method by study area and geographic unit.

GVF Hexagon; Buffalo City Metropolitan Municipality

```
# HEX - Buffalo City Metropolitan Municipality - Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=2]<-1
dfdata$Group[dfdata$Join_Count>2 & dfdata$Join_Count<=18]<-2
dfdata$Group[dfdata$Join_Count>18 & dfdata$Join_Count<=118]<-3
dfdata$Group[dfdata$Join_Count>118 & dfdata$Join_Count<=785]<-4
dfdata$Group[dfdata$Join_Count>785]<-5

# HEX - Buffalo City Metropolitan Municipality - Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=1]<-1
dfdata$Group[dfdata$Join_Count>1 & dfdata$Join_Count<=10]<-2
dfdata$Group[dfdata$Join_Count>10 & dfdata$Join_Count<=100]<-3
dfdata$Group[dfdata$Join_Count>100 & dfdata$Join_Count<=1000]<-4
dfdata$Group[dfdata$Join_Count>1000]<-5

# HEX - Buffalo City Metropolitan Municipality - Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=146]<-1
dfdata$Group[dfdata$Join_Count>146 & dfdata$Join_Count<=493]<-2
dfdata$Group[dfdata$Join_Count>493 & dfdata$Join_Count<=1069]<-3
dfdata$Group[dfdata$Join_Count>1069 & dfdata$Join_Count<=1892]<-4
dfdata$Group[dfdata$Join_Count>1892]<-5

# HEX - Buffalo City Metropolitan Municipality - Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=1]<-1
dfdata$Group[dfdata$Join_Count>1 & dfdata$Join_Count<=11]<-2
dfdata$Group[dfdata$Join_Count>11 & dfdata$Join_Count<=45]<-3
dfdata$Group[dfdata$Join_Count>45 & dfdata$Join_Count<=161]<-4
dfdata$Group[dfdata$Join_Count>161]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(Join_Count)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(Join_Count)) %>%
  mutate(Class_Diff = abs(Join_Count-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(Join_Count-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Small Area Layer; Buffalo City Metropolitan Municipality

```
# SAL - Buffalo City Metropolitan Municipality - Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=493.09]<-1
dfdata$Group[dfdata$POPDENS>493.09 & dfdata$POPDENS<=1942.39]<-2
dfdata$Group[dfdata$POPDENS>1942.39 & dfdata$POPDENS<=6207.27]<-3
dfdata$Group[dfdata$POPDENS>6207.27 & dfdata$POPDENS<=18757.66]<-4
dfdata$Group[dfdata$POPDENS>18757.66]<-5

# SAL - Buffalo City Metropolitan Municipality - Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1]<-1
dfdata$Group[dfdata$POPDENS>1 & dfdata$POPDENS<=10]<-2
dfdata$Group[dfdata$POPDENS>10 & dfdata$POPDENS<=100]<-3
dfdata$Group[dfdata$POPDENS>100 & dfdata$POPDENS<=1000]<-4
dfdata$Group[dfdata$POPDENS>1000 & dfdata$POPDENS<=10000]<-5
dfdata$Group[dfdata$POPDENS>10000]<-6

# SAL - Buffalo City Metropolitan Municipality - Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=3104.72]<-1
dfdata$Group[dfdata$POPDENS>3104.72 & dfdata$POPDENS<=6966.21]<-2
dfdata$Group[dfdata$POPDENS>6966.21 & dfdata$POPDENS<=13601.41]<-3
dfdata$Group[dfdata$POPDENS>13601.41 & dfdata$POPDENS<=25907.38]<-4
dfdata$Group[dfdata$POPDENS>25907.38]<-5

# SAL - Buffalo City Metropolitan Municipality - Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=883.89]<-1
dfdata$Group[dfdata$POPDENS>883.89 & dfdata$POPDENS<=2094.91]<-2
dfdata$Group[dfdata$POPDENS>2094.21 & dfdata$POPDENS<=4450.34]<-3
dfdata$Group[dfdata$POPDENS>4450.34 & dfdata$POPDENS<=7435.90]<-4
dfdata$Group[dfdata$POPDENS>7435.90]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(POPDENS)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(POPDENS)) %>%
  mutate(Class_Diff = abs(POPDENS-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(POPDENS-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Sub-Place; Buffalo City Metropolitan Municipality

```
# SP  - Buffalo City Metropolitan Municipality - Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=658.27]<-1
dfdata$Group[dfdata$POPDENS>658.27 & dfdata$POPDENS<=757.83]<-2
dfdata$Group[dfdata$POPDENS>757.83 & dfdata$POPDENS<=1416.10]<-3
dfdata$Group[dfdata$POPDENS>1416.10 & dfdata$POPDENS<=5768.37]<-4
dfdata$Group[dfdata$POPDENS>5768.37]<-5

# SP - Buffalo City Metropolitan Municipality - Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=0.1]<-1
dfdata$Group[dfdata$POPDENS>0.1 & dfdata$POPDENS<=1]<-2
dfdata$Group[dfdata$POPDENS>1 & dfdata$POPDENS<=100]<-3
dfdata$Group[dfdata$POPDENS>100 & dfdata$POPDENS<=10000]<-4
dfdata$Group[dfdata$POPDENS>10000]<-5

# SP - Buffalo City Metropolitan Municipality - Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1255.6]<-1
dfdata$Group[dfdata$POPDENS>1255.6 & dfdata$POPDENS<=3347.00]<-2
dfdata$Group[dfdata$POPDENS>3347.00 & dfdata$POPDENS<=7622.41]<-3
dfdata$Group[dfdata$POPDENS>7622.41 & dfdata$POPDENS<=17401.9]<-4
dfdata$Group[dfdata$POPDENS>17401.9]<-5

# SP - Buffalo City Metropolitan Municipality - Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=307.86]<-1
dfdata$Group[dfdata$POPDENS>307.86 & dfdata$POPDENS<=756.1]<-2
dfdata$Group[dfdata$POPDENS>756.1 & dfdata$POPDENS<=1245.84]<-3
dfdata$Group[dfdata$POPDENS>1245.84 & dfdata$POPDENS<=2470.36]<-4
dfdata$Group[dfdata$POPDENS>2470.6]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(POPDENS)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(POPDENS)) %>%
  mutate(Class_Diff = abs(POPDENS-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(POPDENS-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Hexagon; City of Tshwane Metropolitan Municipality

```
# HEX Tshwane Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=2]<-1
dfdata$Group[dfdata$Join_Count>2 & dfdata$Join_Count<=18]<-2
dfdata$Group[dfdata$Join_Count>18 & dfdata$Join_Count<=131]<-3
dfdata$Group[dfdata$Join_Count>131 & dfdata$Join_Count<=947]<-4
dfdata$Group[dfdata$Join_Count>947]<-5

# HEX Tshwane Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=1]<-1
dfdata$Group[dfdata$Join_Count>1 & dfdata$Join_Count<=10]<-2
dfdata$Group[dfdata$Join_Count>10 & dfdata$Join_Count<=100]<-3
dfdata$Group[dfdata$Join_Count>100 & dfdata$Join_Count<=1000]<-4
dfdata$Group[dfdata$Join_Count>1000]<-5

# HEX Tshwane Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=319]<-1
dfdata$Group[dfdata$Join_Count>319 & dfdata$Join_Count<=929]<-2
dfdata$Group[dfdata$Join_Count>929 & dfdata$Join_Count<=1845]<-3
dfdata$Group[dfdata$Join_Count>1845 & dfdata$Join_Count<=3602]<-4
dfdata$Group[dfdata$Join_Count>3602]<-5

# HEX Tshwane Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=1]<-1
dfdata$Group[dfdata$Join_Count>1 & dfdata$Join_Count<=12]<-2
dfdata$Group[dfdata$Join_Count>12 & dfdata$Join_Count<=35]<-3
dfdata$Group[dfdata$Join_Count>35 & dfdata$Join_Count<=154]<-4
dfdata$Group[dfdata$Join_Count>154]<-5

# HEX Tshwane Standard Deviation Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=446]<-1
dfdata$Group[dfdata$Join_Count>446 & dfdata$Join_Count<=950]<-2
dfdata$Group[dfdata$Join_Count>950 & dfdata$Join_Count<=1454]<-3
dfdata$Group[dfdata$Join_Count>1454]<-4

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(Join_Count)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(Join_Count)) %>%
  mutate(Class_Diff = abs(Join_Count-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(Join_Count-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Small Area Layer; City of Tshwane Metropolitan Municipality

```
# Tshwane SAL Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=2846.19]<-1
dfdata$Group[dfdata$POPDENS>2846.19 & dfdata$POPDENS<=3317.05]<-2
dfdata$Group[dfdata$POPDENS>3317.05 & dfdata$POPDENS<=6162.4]<-3
dfdata$Group[dfdata$POPDENS>6162.4 & dfdata$POPDENS<=23356.78]<-4
dfdata$Group[dfdata$POPDENS>23356.78]<-5

# SAL Tshwane Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1]<-1
dfdata$Group[dfdata$POPDENS>1 & dfdata$POPDENS<=10]<-2
dfdata$Group[dfdata$POPDENS>10 & dfdata$POPDENS<=100]<-3
dfdata$Group[dfdata$POPDENS>100 & dfdata$POPDENS<=1000]<-4
dfdata$Group[dfdata$POPDENS>1000 & dfdata$POPDENS<=10000]<-5
dfdata$Group[dfdata$POPDENS>10000 & dfdata$POPDENS<=100000]<-6
dfdata$Group[dfdata$POPDENS>100000]<-7

# SAL Tshwane Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=5355.69]<-1
dfdata$Group[dfdata$POPDENS>5355.69 & dfdata$POPDENS<=12082.13]<-2
dfdata$Group[dfdata$POPDENS>12082.13 & dfdata$POPDENS<=23700.37]<-3
dfdata$Group[dfdata$POPDENS>23700.37 & dfdata$POPDENS<=47629.78]<-4
dfdata$Group[dfdata$POPDENS>47629.78]<-5

# SAL Tshwane Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1982.12]<-1
dfdata$Group[dfdata$POPDENS>1982.12 & dfdata$POPDENS<=3381.61]<-2
dfdata$Group[dfdata$POPDENS>3381.61 & dfdata$POPDENS<=6017.24]<-3
dfdata$Group[dfdata$POPDENS>6017.24 & dfdata$POPDENS<=10126.96]<-4
dfdata$Group[dfdata$POPDENS>10126.96]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(POPDENS)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(POPDENS)) %>%
  mutate(Class_Diff = abs(POPDENS-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(POPDENS-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Sub-Place; City of Tshwane Metropolitan Municipality

```
# Tshwane SP Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=604.44]<-1
dfdata$Group[dfdata$POPDENS>604.44 & dfdata$POPDENS<=2024.95]<-2
dfdata$Group[dfdata$POPDENS>2024.95 & dfdata$POPDENS<=5363.35]<-3
dfdata$Group[dfdata$POPDENS>5363.35 & dfdata$POPDENS<=13209.06]<-4
dfdata$Group[dfdata$POPDENS>13209.06]<-5

# SP Tshwane Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=0.01]<-1
dfdata$Group[dfdata$POPDENS>0.01 & dfdata$POPDENS<=1]<-2
dfdata$Group[dfdata$POPDENS>1 & dfdata$POPDENS<=100]<-3
dfdata$Group[dfdata$POPDENS>100 & dfdata$POPDENS<=10000]<-4
dfdata$Group[dfdata$POPDENS>10000]<-5

# SP Tshwane Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1506.35]<-1
dfdata$Group[dfdata$POPDENS>1506.35 & dfdata$POPDENS<=4298.27]<-2
dfdata$Group[dfdata$POPDENS>4298.27 & dfdata$POPDENS<=9178.97]<-3
dfdata$Group[dfdata$POPDENS>9178.97 & dfdata$POPDENS<=17426.16]<-4
dfdata$Group[dfdata$POPDENS>17426.16]<-5

# SP Tshwane Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=184.87]<-1
dfdata$Group[dfdata$POPDENS>184.87 & dfdata$POPDENS<=1599.13]<-2
dfdata$Group[dfdata$POPDENS>1599.13 & dfdata$POPDENS<=2819.35]<-3
dfdata$Group[dfdata$POPDENS>2819.35 & dfdata$POPDENS<=5801.23]<-4
dfdata$Group[dfdata$POPDENS>5801.23]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(POPDENS)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(POPDENS)) %>%
  mutate(Class_Diff = abs(POPDENS-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(POPDENS-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Hexagon; Mangaung Metropolitan Municipality

```
# HEX Mangaung Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=0]<-1
dfdata$Group[dfdata$Join_Count>0 & dfdata$Join_Count<=4]<-2
dfdata$Group[dfdata$Join_Count>4 & dfdata$Join_Count<=43]<-3
dfdata$Group[dfdata$Join_Count>43 & dfdata$Join_Count<=442]<-4
dfdata$Group[dfdata$Join_Count>442]<-5

# HEX Mangaung Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=1]<-1
dfdata$Group[dfdata$Join_Count>1 & dfdata$Join_Count<=10]<-2
dfdata$Group[dfdata$Join_Count>10 & dfdata$Join_Count<=100]<-3
dfdata$Group[dfdata$Join_Count>100 & dfdata$Join_Count<=1000]<-4
dfdata$Group[dfdata$Join_Count>1000]<-5

# HEX Mangaung Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=210]<-1
dfdata$Group[dfdata$Join_Count>210 & dfdata$Join_Count<=694]<-2
dfdata$Group[dfdata$Join_Count>694 & dfdata$Join_Count<=1366]<-3
dfdata$Group[dfdata$Join_Count>1366 & dfdata$Join_Count<=2471]<-4
dfdata$Group[dfdata$Join_Count>2471]<-5

# HEX Mangaung Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=0]<-1
dfdata$Group[dfdata$Join_Count>0 & dfdata$Join_Count<=3]<-2
dfdata$Group[dfdata$Join_Count>3 & dfdata$Join_Count<=9]<-3
dfdata$Group[dfdata$Join_Count>9 & dfdata$Join_Count<=35]<-4
dfdata$Group[dfdata$Join_Count>35]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(Join_Count)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(Join_Count)) %>%
  mutate(Class_Diff = abs(Join_Count-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(Join_Count-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Small Area Layer; Mangaung Metropolitan Municipality

```
# Bloem SAL Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=2867.28]<-1
dfdata$Group[dfdata$POPDENS>2867.28 & dfdata$POPDENS<=3841.58]<-2
dfdata$Group[dfdata$POPDENS>3841.58 & dfdata$POPDENS<=6708.68]<-3
dfdata$Group[dfdata$POPDENS>6708.68 & dfdata$POPDENS<=15145.77]<-4
dfdata$Group[dfdata$POPDENS>15145.77]<-5

# SAL Bloem Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1]<-1
dfdata$Group[dfdata$POPDENS>1 & dfdata$POPDENS<=10]<-2
dfdata$Group[dfdata$POPDENS>10 & dfdata$POPDENS<=100]<-3
dfdata$Group[dfdata$POPDENS>100 & dfdata$POPDENS<=1000]<-4
dfdata$Group[dfdata$POPDENS>1000 & dfdata$POPDENS<=10000]<-5
dfdata$Group[dfdata$POPDENS>10000]<-6

# SAL Bloem Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=2373.91]<-1
dfdata$Group[dfdata$POPDENS>2373.91 & dfdata$POPDENS<=4904.65]<-2
dfdata$Group[dfdata$POPDENS>4904.65 & dfdata$POPDENS<=7547.35]<-3
dfdata$Group[dfdata$POPDENS>7547.35 & dfdata$POPDENS<=11100.24]<-4
dfdata$Group[dfdata$POPDENS>11100.24]<-5

# SAL Bloem Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1670.32]<-1
dfdata$Group[dfdata$POPDENS>1670.32 & dfdata$POPDENS<=3574.55]<-2
dfdata$Group[dfdata$POPDENS>3574.55 & dfdata$POPDENS<=5477.48]<-3
dfdata$Group[dfdata$POPDENS>5477.48 & dfdata$POPDENS<=7692.21]<-4
dfdata$Group[dfdata$POPDENS>7692.21]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(POPDENS)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(POPDENS)) %>%
  mutate(Class_Diff = abs(POPDENS-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(POPDENS-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Sub-Place; Mangaung Metropolitan Municipality

```
# Mangaung SP Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=271.44]<-1
dfdata$Group[dfdata$POPDENS>271.44 & dfdata$POPDENS<=804.35]<-2
dfdata$Group[dfdata$POPDENS>804.35 & dfdata$POPDENS<=1850.55]<-3
dfdata$Group[dfdata$POPDENS>1850.55 & dfdata$POPDENS<=3904.48]<-4
dfdata$Group[dfdata$POPDENS>3904.48]<-5

# SP Mangaung Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1]<-1
dfdata$Group[dfdata$POPDENS>1 & dfdata$POPDENS<=10]<-2
dfdata$Group[dfdata$POPDENS>10 & dfdata$POPDENS<=100]<-3
dfdata$Group[dfdata$POPDENS>100 & dfdata$POPDENS<=1000]<-4
dfdata$Group[dfdata$POPDENS>1000]<-5

# SP Mangaung Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=546.12]<-1
dfdata$Group[dfdata$POPDENS>546.12 & dfdata$POPDENS<=1470.82]<-2
dfdata$Group[dfdata$POPDENS>1470.82 & dfdata$POPDENS<=3047.83]<-3
dfdata$Group[dfdata$POPDENS>3047.83 & dfdata$POPDENS<=4735.93]<-4
dfdata$Group[dfdata$POPDENS>4735.93]<-5

# SP Mangaung Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=257.74]<-1
dfdata$Group[dfdata$POPDENS>257.74 & dfdata$POPDENS<=750.97]<-2
dfdata$Group[dfdata$POPDENS>750.97 & dfdata$POPDENS<=1836.37]<-3
dfdata$Group[dfdata$POPDENS>1836.37 & dfdata$POPDENS<=3604.32]<-4
dfdata$Group[dfdata$POPDENS>3604.32]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(POPDENS)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(POPDENS)) %>%
  mutate(Class_Diff = abs(POPDENS-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(POPDENS-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Hexagon; Polokwane Local Municipality

```
# HEX Polokwane Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=0]<-1
dfdata$Group[dfdata$Join_Count>0 & dfdata$Join_Count<=4]<-2
dfdata$Group[dfdata$Join_Count>4 & dfdata$Join_Count<=36]<-3
dfdata$Group[dfdata$Join_Count>36 & dfdata$Join_Count<=296]<-4
dfdata$Group[dfdata$Join_Count>296]<-5

# HEX Polokwane Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=1]<-1
dfdata$Group[dfdata$Join_Count>1 & dfdata$Join_Count<=10]<-2
dfdata$Group[dfdata$Join_Count>10 & dfdata$Join_Count<=100]<-3
dfdata$Group[dfdata$Join_Count>100 & dfdata$Join_Count<=1000]<-4
dfdata$Group[dfdata$Join_Count>1000]<-5

# HEX Polokwane Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=121]<-1
dfdata$Group[dfdata$Join_Count>121 & dfdata$Join_Count<=375]<-2
dfdata$Group[dfdata$Join_Count>375 & dfdata$Join_Count<=811]<-3
dfdata$Group[dfdata$Join_Count>811 & dfdata$Join_Count<=1563]<-4
dfdata$Group[dfdata$Join_Count>1563]<-5

# HEX Polokwane Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$Join_Count<=0]<-1
dfdata$Group[dfdata$Join_Count>0 & dfdata$Join_Count<=4]<-2
dfdata$Group[dfdata$Join_Count>4 & dfdata$Join_Count<=18]<-3
dfdata$Group[dfdata$Join_Count>18 & dfdata$Join_Count<=128]<-4
dfdata$Group[dfdata$Join_Count>128]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(Join_Count)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(Join_Count)) %>%
  mutate(Class_Diff = abs(Join_Count-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(Join_Count-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Small Area Layer; Polokwane Local Municipality

```
# Polokwane SAL Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=867.21]<-1
dfdata$Group[dfdata$POPDENS>867.21 & dfdata$POPDENS<=1004.37]<-2
dfdata$Group[dfdata$POPDENS>1004.37 & dfdata$POPDENS<=1871.32]<-3
dfdata$Group[dfdata$POPDENS>1871.32 & dfdata$POPDENS<=7351.29]<-4
dfdata$Group[dfdata$POPDENS>7351.29]<-5

# SAL Polokwane Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=1]<-1
dfdata$Group[dfdata$POPDENS>1 & dfdata$POPDENS<=10]<-2
dfdata$Group[dfdata$POPDENS>10 & dfdata$POPDENS<=100]<-3
dfdata$Group[dfdata$POPDENS>100 & dfdata$POPDENS<=1000]<-4
dfdata$Group[dfdata$POPDENS>1000 & dfdata$POPDENS<=10000]<-5
dfdata$Group[dfdata$POPDENS>10000]<-6

# SAL Polokwane Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=2076.41]<-1
dfdata$Group[dfdata$POPDENS>2076.41 & dfdata$POPDENS<=5202.72]<-2
dfdata$Group[dfdata$POPDENS>5202.72 & dfdata$POPDENS<=10417.54]<-3
dfdata$Group[dfdata$POPDENS>10417.54 & dfdata$POPDENS<=17924.33]<-4
dfdata$Group[dfdata$POPDENS>17924.33]<-5

# SAL Polokwane Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=711.6]<-1
dfdata$Group[dfdata$POPDENS>711.6 & dfdata$POPDENS<=1239.11]<-2
dfdata$Group[dfdata$POPDENS>1239.11 & dfdata$POPDENS<=2099.05]<-3
dfdata$Group[dfdata$POPDENS>2099.05 & dfdata$POPDENS<=4109.38]<-4
dfdata$Group[dfdata$POPDENS>4109.38]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(POPDENS)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(POPDENS)) %>%
  mutate(Class_Diff = abs(POPDENS-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(POPDENS-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```

GVF; Sub-Place; Polokwane Local Municipality

```
# Polokwane SP Geometric Interval Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=489.58]<-1
dfdata$Group[dfdata$POPDENS>489.58 & dfdata$POPDENS<=616.6]<-2
dfdata$Group[dfdata$POPDENS>616.6 & dfdata$POPDENS<=1106.18]<-3
dfdata$Group[dfdata$POPDENS>1106.18 & dfdata$POPDENS<=2993.13]<-4
dfdata$Group[dfdata$POPDENS>2993.13]<-5

# SP Polokwane Logarithmic Scale Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=0.01]<-1
dfdata$Group[dfdata$POPDENS>0.01 & dfdata$POPDENS<=1]<-2
dfdata$Group[dfdata$POPDENS>1 & dfdata$POPDENS<=100]<-3
dfdata$Group[dfdata$POPDENS>100 & dfdata$POPDENS<=10000]<-4
dfdata$Group[dfdata$POPDENS>10000]<-5

# SP Polokwane Natural Breaks Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=625.76]<-1
dfdata$Group[dfdata$POPDENS>625.76 & dfdata$POPDENS<=1455.76]<-2
dfdata$Group[dfdata$POPDENS>1455.76 & dfdata$POPDENS<=2821.55]<-3
dfdata$Group[dfdata$POPDENS>2821.55 & dfdata$POPDENS<=5502.26]<-4
dfdata$Group[dfdata$POPDENS>5502.26]<-5

# SP Polokwane Quantiles Group
dfdata$Group <- dfdata$Group[dfdata$POPDENS<=343.29]<-1
dfdata$Group[dfdata$POPDENS>343.29 & dfdata$POPDENS<=728.56]<-2
dfdata$Group[dfdata$POPDENS>728.56 & dfdata$POPDENS<=1094.37]<-3
dfdata$Group[dfdata$POPDENS>1094.37 & dfdata$POPDENS<=1621.32]<-4
dfdata$Group[dfdata$POPDENS>1621.32]<-5

dfdata

# Calculate Mean by group
M_Calc <- dfdata %>%
  mutate(Mean_Grp = mean(POPDENS)) %>%
  group_by(Group) %>%
  mutate(Mean_All = mean(POPDENS)) %>%
  mutate(Class_Diff = abs(POPDENS-Mean_Grp)^2) %>%
  mutate(All_Diff = abs(POPDENS-Mean_All)^2) %>%
  as.data.frame()

M_Calc

# Calculate Error - GVF
Summary <- data.frame(colSums(M_Calc))
Summary
Data_transp <- transpose(Summary)
(Data_transp$V6 - Data_transp$V7)/Data_transp$V6
```