Disaggregating employment data to building level: A multi-objective optimisation approach

by

Chantel Judith Ludick

Submitted in partial fulfilment of the requirements for

Magister Scientiae (Geoinformatics)

in the Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

August 2020

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

I, Chantel Judith Ludick declare that the dissertation, which I hereby submit for the degree MSc Geoinformatics 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.

Ludick

SIGNATURE: ....

DATE: ....31/07/2020....................

Abstract

Title: Disaggregating employment data to building level: A multi-objective optimisation approach

Student name: Ms Chantel Judith Ludick, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, South Africa

Supervisor: Dr Victoria Rautenbach, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, South Africa

Co-supervisor: Mr Quintin van Heerden, Smart Places, Council for Scientific and Industrial Research, South Africa

Degree: MSc Geoinformatics, Faculty of Natural and Agricultural Sciences, Department of Geography, Geoinformatics and Meteorology, University of Pretoria

The land use policies and development plans that are implemented in a city contribute to whether the city will be sustainable in the future. Therefore, when these policies are being established they should consider the potential impact on development. An analytical tool, such as land use change models, allow decision-makers to see the possible impact that these policies could have on development. Land use change models like UrbanSim make use of the relationship between households, buildings, and employment opportunities to model the decisions that people make on where to live and work. To be able to do this the model needs accurate data.

When there is a more accurate location for the employment opportunities in an area, the decisions made by individuals can be better modelled and therefore the projected results are expected to be better. Previous research indicated that the methods that are traditionally used to disaggregate employment data to a lower level in UrbanSim projects are not applicable in the South African context. This is because the traditional methods require a detailed employment dataset for the disaggregation and this detailed employment dataset is not available in South Africa.

The aim of this project was to develop a methodology for a metropolitan municipality in South Africa that could be used to disaggregate the employment data that is available at a higher level to a more detailed building level. To achieve this, the methodology consisted of two parts. The first part of the methodology was establishing a method that could be used to prepare a base dataset that is used for disaggregating the employment data. The second part of the methodology was using a multi-objective optimisation approach to allocate the number of employment opportunities within a municipality to building level. The algorithm was developed using the Distributed Evolutionary Algorithm in Python (DEAP) computational framework. DEAP is an open-source evolutionary algorithm framework that is developed in Python and enables users to rapidly create prototypes by allowing them to customise the algorithm to suit their needs

The evaluation showed that it is possible to make use of multi-objective optimisation to disaggregate employment data to building level. The results indicate that the employment allocation algorithm was successful in disaggregating employment data from municipal level to building level. All evolutionary algorithms come with some degree of uncertainty as one of the main features of evolutionary algorithms is that they find the most optimal solution, and so there are other solutions available as well. Thus, the results of the algorithm also come with that same level of uncertainty.

By enhancing the data used by land use change models, the performance of the overall model is improved. With this improved performance of the model, an improved view of the impact that land use policies could have on development can also be seen. This will allow decision-makers to draw the best possible conclusions and allow them the best possible opportunity to develop policies that will contribute to creating sustainable and lasting urban areas.

Abstrak

Titel: Disaggregering van indiensneming data na gebou vlak: 'n Multi-objektiewe optimerings benadering

Naam van student: Me Chantel Judith Ludick, Departement van Geografie, Geoinformatika en Meteorologie, Universiteit van Pretoria, Suid-Afrika

Promotor: Dr Victoria Rautenbach, Departement van Geografie, Geoinformatika en Meteorologie, Universiteit van Pretoria, Suid-Afrika

Mede-promotor: Mnr Quintin van Heerden, Smart Places, Wetenskaplike en Nywerheidnavorsingsraad, Suid-Afrika

Graad: MSc Geoinformatika, Fakulteit van Natuur- en Landbouwetenskappe, Departement van Geografie, Geoinformatika en Meteorologie, Universiteit van Pretoria, Suid-Afrika

Die beleid oor grondgebruik en ontwikkelingsplanne wat in 'n stad geïmplementeer word, dra by tot die vraag of die stad in die toekoms volhoubaar sal wees. Daarom, wanneer hierdie beleid opgestel word, moet hulle die moontlike impak op ontwikkeling oorweeg. 'n Analitiese hulpmiddel, soos modelle wat verandering in grondgebruik modelleer, stel besluitnemers in staat om die moontlike impak van hierdie beleid op ontwikkeling te sien. Modelle vir verandering van grondgebruik soos UrbanSim maak gebruik van die verhouding tussen huishoudings, geboue en werkgeleenthede om die besluite wat mense neem oor waar hulle moet woon en werk te modelleer. Om dit te kan doen, benodig die model akkurate data.

As daar 'n meer akkurate ligging vir die werksgeleenthede in 'n gebied is, kan die besluite wat deur individue geneem word, beter gemodelleer word, en die geprokekteerde resultate word dus verwag om te beter wees. Vorige navorsing het aangedui dat die metodes wat tradisioneel gebruik word om indiensnemingsdata op 'n laer vlak in UrbanSim-projekte te verdeel, nie in die Suid-Afrikaanse konteks van toepassing is nie. Dit is omdat die tradisionele metodes 'n gedetailleerde indiensnemingsdatas benodig vir die verdeling en hierdie gedetailleerde werkverskaffingsdataset is nie in Suid-Afrika beskikbaar nie.

Die doel van hierdie projek was om 'n metodologie vir 'n metropolitaanse munisipaliteit in Suid-Afrika te ontwikkel wat gebruik kan word om die indiensnemingsdata wat op 'n hoër vlak beskikbaar is, te verdeel tot 'n meer gedetailleerde gebouvlak. Om dit te bereik, het die metodologie uit twee dele bestaan. Die eerste deel van die metodologie was die formulering van 'n metode wat gebruik kan word om 'n basis-datastel op te stel wat gebruik word om die indiensnemingsdata te verdeel. Die tweede deel van die metodologie was om 'n multi-objektiewe optimaliseringsbenadering te gebruik om die aantal werkgeleenthede binne 'n munisipaliteit op die gebouvlak toe te ken. Die algoritme is ontwikkel met behulp van die ‘Distributed Evolutionary Algorithm in Python (DEAP)’ berekeningsraamwerk. DEAP is 'n oop bron evolusionêre algoritme raamwerk wat in Python ontwikkel is en gebruikers in staat stel om vinnig prototipes te skep deur hulle die opsie te gee om die algoritme aan te pas om aan hul behoeftes te voldoen

Die evaluering het getoon dat dit moontlik is om van multi-objektiewe optimalisering gebruik te maak om die indiensnemingsdata tot die gebouvlak te verdeel. Die resultate dui daarop dat die algoritme vir die toekenning van indiensneming suksesvol was met die verdeling van

indiensnemingsdata van munisipale vlak tot gebouvlak. Alle evolusionêre algoritmes het 'n mate van onsekerheid, want een van die kenmerke van evolusionêre algoritmes is dat dit die beste oplossing vind, en daar is ook ander oplossings beskikbaar. Die resultate van die algoritme het dus ook dieselfde vlak van onsekerheid.

Deur die gegewens te verbeter wat deur modelle vir verandering in grondgebruik gebruik word, word die prestasie van die algehele model verbeter. Met hierdie verbeterde werkverrigting van die model, kan 'n beter siening van die impak wat grondgebruikbeleid op ontwikkeling kan hê ook gesien word. Hiermee kan besluitnemers die beste moontlike gevolgtrekkings maak en hulle die beste moontlike geleentheid bied om beleid te ontwikkel wat sal bydra tot die skepping van volhoubare en blywende stedelike gebiede.

Acknowledgements

My journey through this dissertation and this degree has required more effort than I was able to deliver on my own. For this reason, I would like to acknowledge the many people who supported me throughout this process.

I would first like to thank my supervisors Mr Quintin van Heerden and Dr Victoria Rautenbach. The door to both their offices were always open whenever I ran into a trouble or had a question about my research or writing. They consistently allowed this paper to be my own work, but steered me in the right direction whenever they thought I needed it.

I must express my very profound gratitude to my mother and to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Finally I would like to give thanks to God for giving me the opportunities I have been afforded and for providing me with the strength and inspiration to complete this dissertation.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background

1.1.1 Overview

It is estimated that globally, 4.66 billion people lived in urban areas by 2019, which is around 55.7% of the world's population (Demographia, 2019). Population projections indicate that worldwide around 2.5 billion people will be added to the urban population by 2050. It is projected that most of the increases in the urban population will occur in Asia and Africa. The increase in urban population is not just accounted for by the overall growth of the population, but also the migration of people from rural areas to urban areas in search of new opportunities and improved lifestyles (UN DESA, 2018; World Bank, 2018; World Economic Forum, 2017).

In South Africa, the urban population increased from 60.62% in 2007 to 68.85% in 2017 (Statista, 2019). This growth has led to South Africa being one of the most urbanised countries in Africa (Chibba, 2016; Edmonds, 2013; World Bank, 2016). It is projected that the urban population in South Africa will increase from 38 million in 2018 to between 58 and 62 million by 2050. This means the urban population will account for about 77.5% of South Africa's population (Le Roux, et al., 2019).

An increase in the urban population raises questions as to what the impact will be on urban areas/cities and the sustainability of cities (Seto, et al., 2017; Brelsford, et al., 2017). Researchers have determined that urbanisation could be a key component to creating sustainable cities, but if not managed correctly, it could also be a threat to the sustainability of cities. Cities have often been places where the most innovation in terms of technology and sustainability has occurred, but at the same time, urbanisation has contributed to increases in pollution, environmental degradation, and inequality in cities. Therefore, it is important to be aware of the fact that urbanisation can have both positive and negative effects on cities. When development plans and land use plans for cities are created, both the positive and negative effects need to be kept in mind (Ramaswami, et al., 2016; Seto, et al., 2017; Brelsford, et al., 2017).

One of the major hurdles that policy and decision-makers face when developing land use policies for cities, is that they do not know what implications their land use policies will have on how a city develops (Le Roux & Augustijn, 2017; Guan, et al., 2011). Therefore, there is a need to use tools that can assist decision-makers to identify the implications that different policies could have on development. Land use change models are one of the tools that are currently available to assist in policymaking. Land use change models allow users to analyse both the cause and effect of land use change as well as provide a user with the capability to study how land use systems function (Le Roux & Augustijn, 2017).

The sections that follow will provide more background on where the research for the dissertation stems from.

1.1.2 Land use change models

Land use change models have been successfully used around the world to model the effects of different land use policies on the development of cities (Guan, et al., 2011; Le Roux & Augustijn, 2017). This is also true in South Africa. Le Roux and Augustjin (2017) simulated the impact of different land use policies on the development of the City of Johannesburg using the Dyna-Clue model, while Abutaleb et al. (2013) made use of the SLEUTH land use change model to map and analyse urban growth in Johannesburg over 20 years.

Land use change models are computer simulation tools that can be used to understand the causes, mechanisms and significance of urban growth (Chaudhuri & Clarke, 2013; Abutaleb, et al., 2013). Currently, there are many land use change models available, each making use of different methods to simulate growth. The three main classes that the models can be divided into are (1) empirical and statistical models, (2) dynamic models, and (3) integrated models. Empirical and statistical models make use of mathematically formulated processes to approximate reality and make predictions. The most well-known empirical and statistical model is the Markov chain model (Guan, et al., 2011; U.S. EPA, 2000).

Dynamic models use mathematical processes but also focus on time-dependent changes. Dynamic models have been implemented the most over the years and the ones that are most used are Cellular Automata (CA) models (e.g. SLEUTH) and agent-based models (e.g. UrbanSim). Integrated models, also known as hybrid models, make use of the fundamentals of other models and combine them to create a model that considers multiple disciplines when simulations are created. The various models that form part of the CLUE modelling framework, which is based on CA models with added components from other models, are the most frequently used integrated models (Verburg & Overmars, 2009; Britz, et al., 2011). Within these categories, models can be further subdivided based on whether they are more suited for modelling urban growth, urban change, or land cover change, to name a few (Guan, et al., 2011; Verburg, et al., 2013). As Dynamic land use change models have been implemented the most, this class of land use change model will be the focus of this dissertation.

1.1.3 Input data required by land use change models

Dynamic land use change models rely on a variety of datasets as input to perform simulations that produce data relating to projections on how an urban area will develop or change, as illustrated in Figure 1. These datasets range from information about households to employment, to transport, to land use, and development plans. Some of the information on households that are considered includes age, income, marital status, and children. The information of employment that is of importance includes employment status, where people work, the number of employment opportunities in an area, and the sector in which a person works.

Information on transport includes roads networks within the city, the different modes of transport used in a city, and the routes that vehicles use. Information on land use includes the actual land uses that are found in the city, the location of the buildings within the city, and the buildings types. All these factors are related and interconnected. Table 1 provides a summary of all the datasets and their accompanying attributes that are used as input data. When all these factors are combined, they make up all the components that influence how a city functions and in the end how a city develops. Many land use change models make use of some type of location choice model to determine which factors best describe the choices made by individuals or which factors influence the choices that individuals make about where they reside and work. For more detail on dynamic land use change models, see Section 2.3.

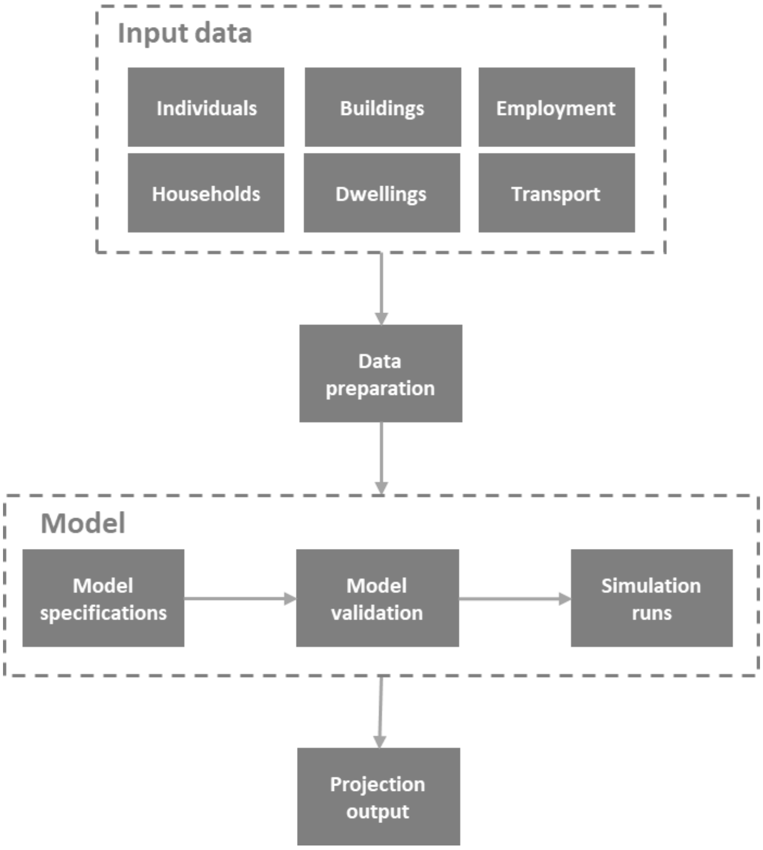


Figure 1: Structure of land use change models

Table 1: Datasets that are often used by land use change models

Datasets used	Attributes within datasets
Households	Age Income Marital status Children
Employment	Employment status Where people work Number of employment opportunities in an area Sector in which a person works
Transport	Roads networks Different modes of transport Routes that vehicles use
Land use	Actual land uses found in the city Location of the buildings within the city Buildings types
Development plans	Areas where development is planned Type of development planned

1.1.4 Location choice models

As mentioned in the previous section, land use change models often use some form of a location choice model or discrete choice model. Location choice models are used to model the changes that households or employment opportunities go through as development takes place. Some of the changes include the relocation of existing households or finding the best location for new households or employment opportunities. To do this, the models use the relationship that exists between households, buildings, and employment opportunities (Van Heerden & Waldeck, 2015; Waddell & Borning, 2004).

Figure 2 shows this multilateral relationship where a household is linked to both a building (i.e. residential building) and a job opportunity. At the same time, a job opportunity is also linked to a building (e.g. commercial building). The relationship between the three factors is established during the creation of a synthetic population dataset that is used as the main input dataset for most location choice models (Van Heerden & Waldeck, 2015).

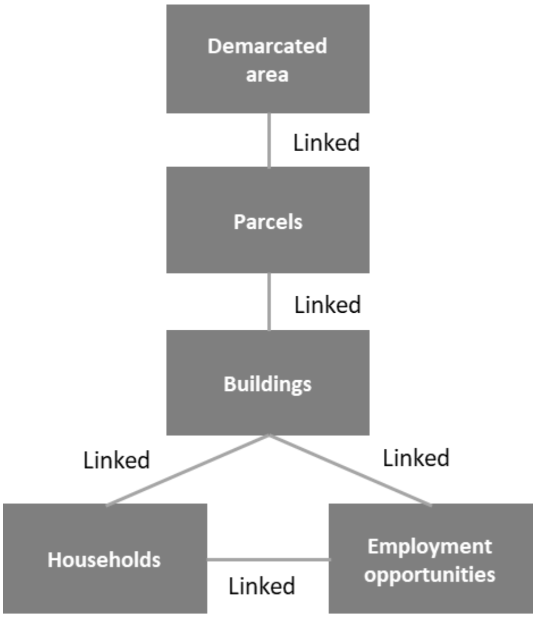


Figure 2: Relationship between buildings, households and employment opportunities.

1.1.5 Synthetic population

Various datasets are integrated and processed to create the synthetic population datasets. Figure 3 depicts some of the input datasets that are used in the development of the synthetic population, as well as where the synthetic population dataset fits into the overall modelling process. The synthetic population is a representation of the actual population in the study area. The population normally consists of individuals who form part of households. Synthetic populations are often created when individual records are not available because of various reasons such as privacy issues or lack of detailed data. Synthetic populations are also created when data with the level of detail required by the model is not available, for example, when the data are only available at the sub-place level (Census demarcation) but the model requires it to be at building level.

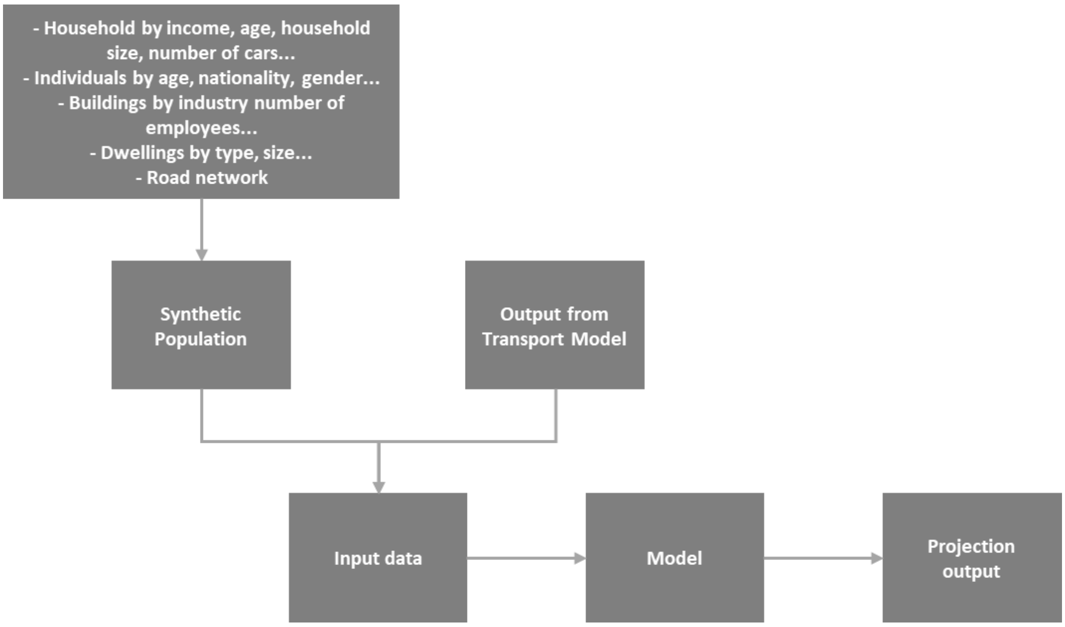

- Household by income, age, household size, number of cars...
- Individuals by age, nationality, gender...
- Buildings by industry number of employees...
- Dwellings by type, size...
- Road network
Synthetic Population
Output from Transport Model
Input data
Model
Projection output

Figure 3: Synthetic population input datasets

The idea behind traditional synthetic population datasets is to use samples of the detailed population, employment, and household data as well as different control attributes and control totals to develop an accurate representation of the population. Figure 4 gives a visual explanation of the concept. For more detail on land use change models, see Section 2.4.

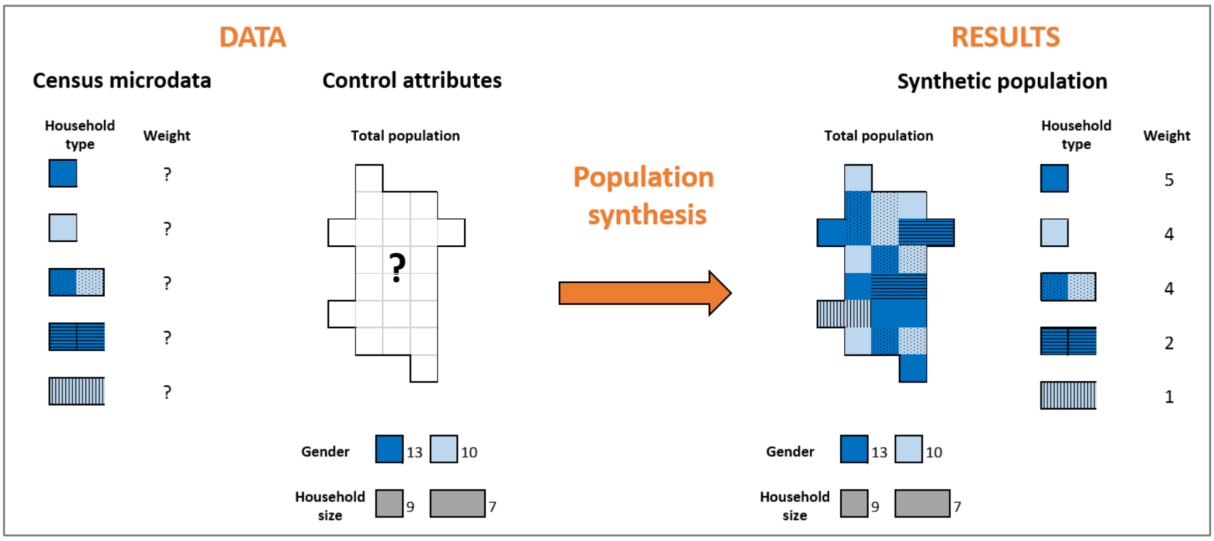


Figure 4: Diagram explaining the creation of a synthetic population (Technical University Munich, 2016)

Currently, in South Africa, there is a lack of detailed data available on where exactly people work and how these employment opportunities are linked to where individuals live. The lowest level at which employment data are provided in South Africa is at the sub-place level and many models require datasets to be at a much lower and more detailed level (e.g. agent-based models require data to be at building level).

1.2 Problem statement

There are certain factors and trends that influence where individuals work and live. Land use change models use these factors and trends to project future changes in urban areas. The factors that influence why individuals make certain decisions can only be identified when the relationship between households, employment opportunities, and buildings is established. Once the factors are identified, the larger decision trends that are followed within a city can also be identified.

Typically there are no datasets available that already have the link between households, employment opportunities, and buildings established. Therefore, a process is needed to generate the linkages between the three components. There are a few existing methods that can link households or employment opportunities to buildings by using very detailed or low-level data from Census, but detailed datasets are not always readily available in all countries. This is especially the case in terms of employment data in South Africa.

This creates the problem of having to disaggregate or allocate the available employment data (that is at a higher level) to a more detailed level. The detailed level data will then make it possible to establish the relationship between employment opportunities and buildings. Once both the households and employment opportunities are allocated to building level, the relationship between these two components can be established. Figure 5 demonstrates the research problem.

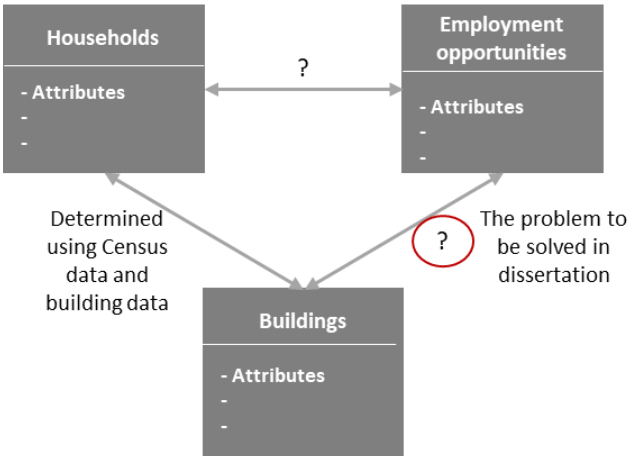


Figure 5: Diagram explaining the research problem

1.3 Research aim and objectives

1.3.1 The aim

The aim of this project is to develop a methodology for a metropolitan municipality (Metro) in South Africa that could be used to disaggregate the employment data that is available at a higher level to a more detailed building level. The methodology consists of an algorithm that can be used to disaggregate employment data and a process to create a base dataset that the algorithm uses.

1.3.2 The objectives

The objectives that will contribute to accomplishing the aim of the dissertation are:

1. Perform a literature review of existing theory and related work on disaggregation methods used in land use change modelling.

2. Create an employment dataset that can be used in the development of the employment allocation algorithm. The development of the dataset consists of the following processes:
    a. Using the literature reviewed as part of Objective 1, identify datasets that can be used for disaggregation.
    b. Develop a procedure that can be used to prepare a dataset (i.e. employment capacity, a figure the provide the amount of employment opportunities that a building can accommodate) for a metro in South Africa.
3. Design and implement the employment allocation algorithm for a metro in South Africa. The design and implementation would consist of the following sections:
    a. Develop the algorithm to allocate employment opportunities to buildings.
    b. Validate the performance of the algorithm in allocating employment opportunities.
4. Analyse and discuss the results of the employment allocation algorithm and assess the use of the algorithm in two other metros in South Africa.

1.4 Significance of the research

The main concept of the research is to create a methodology that can be used to disaggregate the level of data that is currently available for employment opportunities. This means that the methodology should be adjustable for any area in South Africa and not just for the study area used in the research. This will provide the opportunity to have more detailed employment data across South Africa. The methodology could provide concepts or methods for other countries with similar data challenges on how higher-level employment data can be disaggregated. If these countries have similar data structures to South Africa, the methodology could even be adapted for their needs.

The main concepts in the methodology could be used for solving similar data challenges with other types of data. For example, the concepts could be adapted for disaggregating high-level household data to residential buildings to have detailed information on where people live. Although the idea behind the disaggregated dataset for this project is to use the data as part of an input for a land use change model, such as UrbanSim, the final results can be used in other contexts as well. Detailed employment data can help provide insight into solving many problems including identification of detailed employment hubs in a city, identifying areas with a shortage of employment opportunities that can then provide areas where investment is required, and it can provide detailed information on the different types of economic focus areas within a city.

1.5 Overview of the chapters

The structure of the dissertation is as follows (Figure 6 provides an overview of the document structure):

Chapter 2 - Literature Review. The literature review includes a discussion on urban growth and land use policies, the UrbanSim land use change model, synthetic population, existing disaggregation methods, and Evolutionary algorithms. [Objective 1 and Objective 2]

Chapter 3 – Methods. In the methods chapter, the research design used for the dissertation is provided. It provides an overview of the study area for which the disaggregation methodology is developed. It provides the methodology for the dissertation as well as then high-level steps that were followed in the development of the disaggregation methodology. Finally, it discusses the limitations and assumption of the research.

Chapter 4 - Data preparation. The data preparation chapter discusses the various datasets that were used in the development of a final base dataset. The different steps for creating the final base dataset is also presented. [Objective 2]

Chapter 5 - Development of the employment allocation algorithm. In this chapter, a short discussion on the library that was used in the design of the algorithm is provided. After this, the process of developing the algorithm is discussed. The final part of the chapter is the validation of the algorithm results and a discussion of the results. [Objective 3 and Objective 4]

Chapter 6 – Testing the algorithm in other case study areas. In this chapter, the algorithm developed in the previous chapter is applied to two other case study areas. As part of this chapter and overview of the two case study areas are given. After this, the algorithm implementation is briefly discussed. Finally, the results of the algorithm are validated and discussed for each study area. [Objective 4]

Chapter 7 – Conclusion. In this chapter, the most noteworthy results are discussed, along with recommendations on further research that can be performed.

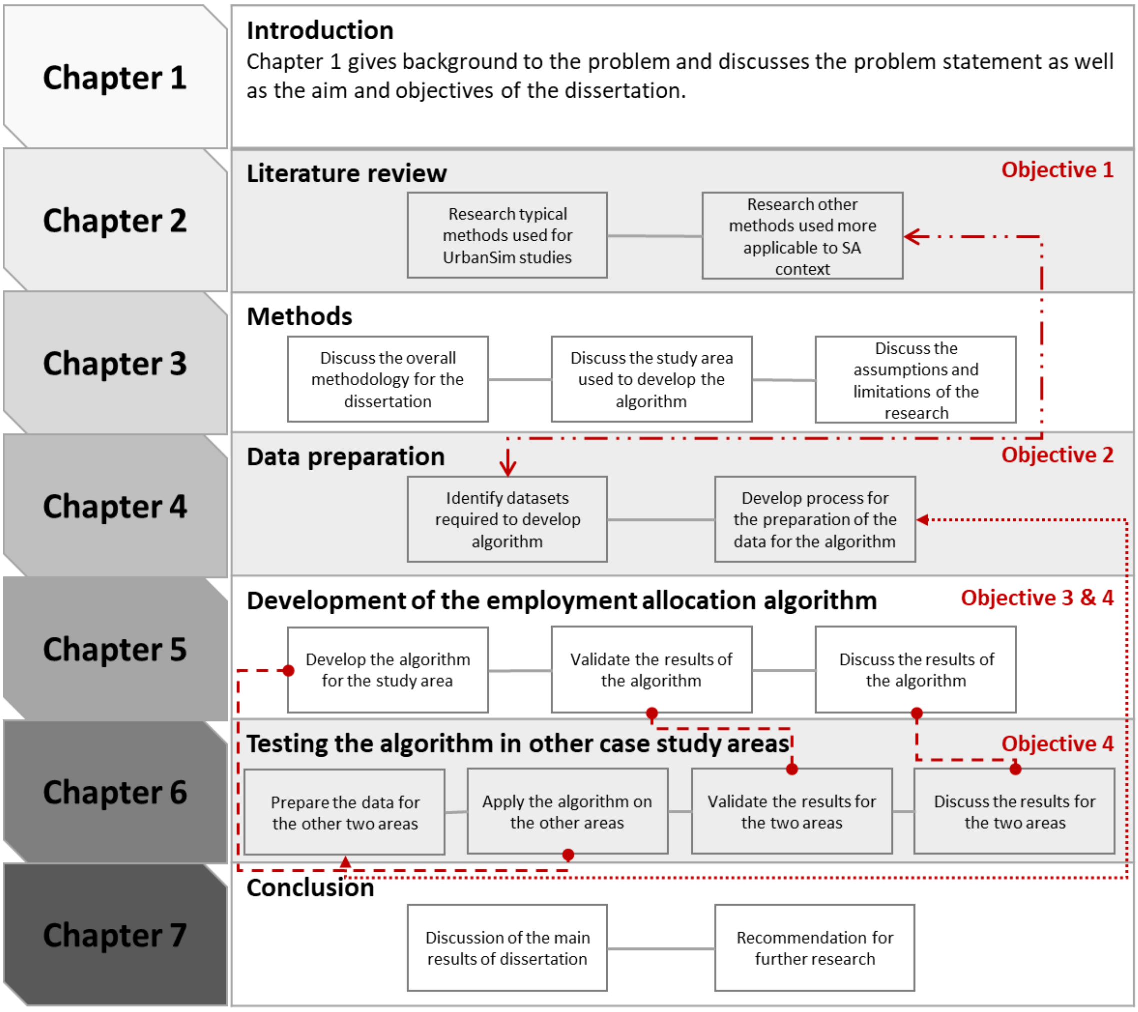

Chapter 1
Introduction
Chapter 1 gives background to the problem and discusses the problem statement as well as the aim and objectives of the dissertation.
Chapter 2
Literature review
Objective 1
Research typical methods used for UrbanSim studies
Research other methods used more applicable to SA context
Chapter 3
Methods
Discuss the overall methodology for the dissertation
Discuss the study area used to develop the algorithm
Discuss the assumptions and limitations of the research
Chapter 4
Data preparation
Objective 2
Identify datasets required to develop algorithm
Develop process for the preparation of the data for the algorithm
Chapter 5
Development of the employment allocation algorithm
Objective 3 & 4
Develop the algorithm for the study area
Validate the results of the algorithm
Discuss the results of the algorithm
Chapter 6
Testing the algorithm in other case study areas
Objective 4
Prepare the data for the other two areas
Apply the algorithm on the other areas
Validate the results for the two areas
Discuss the results for the two areas
Chapter 7
Conclusion
Discussion of the main results of dissertation
Recommendation for further research

Figure 6: Document structure

Chapter 2: Literature Review

2.1 Introduction

This chapter discusses urban growth and land use policies, as well as the UrbanSim land use change model in more detail. The literature review also includes the notion of synthetic populations, which is a sub-component of many land use change models (including Urbansim) and the part of the model where the employment allocation data is used. Different literature that is related to the problem identified for this research is also discussed. The last parts of the literature review provides a discussion on algorithms.

2.2 Urban growth and land use policies

Urban growth and land use policies are intricately linked as both phenomena are very dependent on each other. How urban growth takes place in a city depends very much on the land use policies that are implemented in that city, but this dependency is reciprocal since land use policies also depend on urban growth that has taken place in the past. This past growth influences what type of policies need to be implemented in order to reduce the negative effects caused by certain types of urban growth (Black & Henderson, 1999).

Two of the main types of urban growth that are experienced in most urban areas are urban sprawl and urban densification.

Urban sprawl

Urban sprawl is defined as excessive spatial growth or uncontrolled expansion of urban areas (Brueckner, 2000; Oxford Dictionary, 2018). In South Africa one of the major contributors to urban sprawl is the development of informal settlements on the outskirts of cities where there is a higher availability of land (Dovey & King, 2011). Many negative effects are associated with urban sprawl, such the inefficient use of resources - especially land. The widespread development of residential areas and retail shopping malls has resulted in a lot of land being wasted and left unused. It has also led to many biodiversity areas unnecessarily being lost (Brueckner, 2000; Ioannides & Rossi-Hansberg, 2010).

Another big problem associated with built-up areas being so widely spread, is traffic congestion. The longer commutes that residents have to take, results in extra cost in terms of money spent on petrol as well as the extra time that is spent travelling between work and home. The low density of areas also led to high social inequality in many areas within cities. Not all residential areas have equal access to all services such as running water, electricity, proper sanitation, refuse removal. This is because many residential areas are not located close to the main part of the city so it takes a lot more effort, planning and money to provide these areas with the necessary services (Ioannides & Rossi-Hansberg, 2010; Frumkin, 2016).

Originally, urban sprawl was encouraged during the development of many major cities, but over the last decade or so, urban densification has been encouraged as it is believed to combat the negative effects of urban sprawl (Edmonds, 2013).

Urban densification

Urban densification is defined as the process of creating a city that is more compact as well as efficient in using resources and its infrastructure (Haaland & van den Bosch, 2015). One of the focus points of urban densification is also to create cities that are sustainable in terms of, for example, electricity usage, the usage of space, and the protection of biodiversity areas (Edmonds, 2013).

Urban densification neutralises urban sprawl by encouraging any type of development to take place in areas where there are existing structures as well as an established infrastructure. New residential areas should be developed closer to one another and within these residential areas; the structure should be more compact. Mixed land use within the city is encouraged and the improvement and usage of public transport are encouraged. Although many researchers believe that urban densification will solve the problems caused by urban sprawl, there are also a few researchers that have identified several negative effects associated with it (Haaland & van den Bosch, 2015).

One of these negative effects is crowding and the effect that it has on the living quality of the urban residents. Another major negative effect is the increase in various types of pollution, including noise and air pollution. Since the built-up areas are a lot denser, pollution is higher and more concentrated. It also affects more people than it would with more widespread built-up areas. The higher density housing and mixed land use are also believed to lower the living quality of urban residents. A trend of decrease in urban green spaces in many major cities in Australia, Vietnam, Bangladesh, and many other Asian countries has also become a cause of concern and is believed to be caused by the increased densification. The decrease in urban green space may have many unknown effects on the health of residents as well as on air quality in a city (Mohajeri, et al., 2015; Lin, et al., 2015).

Since urban densification is being encouraged in many cities, it is important that the effect that it might have on a city be studied further. This can be done by making use of tools like the UrbanSim land use change model to project future growth and using the results to determine possible effects of urban densification.

2.3 UrbanSim

To create an effective employment dataset, it is important to understand how the data is used by a land use change model. Therefore, an in-depth study into the structure of the UrbanSim land use change model was conducted.

2.3.1 History

The UrbanSim system was originally designed by Paul Waddell from the University of California, Berkley and was first implemented in 1996 using Java. The idea behind UrbanSim was to provide aid in developing planning policies in Honolulu, Hawaii and in the early stages of development it was also applied to Springfield, Oregon in 1998. After this, the first official version of UrbanSim was released in 2000 (Waddell & Borning, 2004; UrbanSim, n.d.).

Over the subsequent 15 years, UrbanSim has continually been improved and refined by multiple contributors. In 2005 UrbanSim was re-implemented in Python and at the same time, the Open Platform for Urban Simulation (OPUS) software was also implemented. UrbanSim was re-implemented in Python because of the useful libraries that are provided. At the time, one of the biggest draws and most useful components of Python was the Numpy numerical library. After OPUS was implemented, the code for UrbanSim was re-engineerd in order to reduce the complexity of using UrbanSim. At this point, the UrbanSim library started making extensive use of the Pandas library, IPython and statsmodels and it relied heavily on the PyData community (Python, n.d.; UrbanSim, n.d.).

In 2012, the focus was shifted to developing a better user interface. To do this, time and resources were invested in creating a 3D urban visualisation, a trend that was widely implemented at the time. It was believed that the 3D visualisations would allow UrbanSim to produce results that can be better understood and interpreted by policymakers and the public (UrbanSim, n.d.). When visualisations were improved through the usage of 3D visualisations, there was a movement in 2016 to establishing a more scalable and user-friendly system by making use of another emerging trend at the time, known as the *cloud*. By early 2017, the very first cloud-based urban simulation platform was launched and a commitment was made to developing this into a useful tool that can turn planning into a democratic endeavour. Currently, the newest UrbanSim implementation is hosted on the GitHub site and it is maintained by multiple contributors including the main contributor UrbanSim Inc. (Python, n.d.; UrbanSim, n.d.).

2.3.2 Overview

UrbanSim is an open-source, agent-based simulation system that is used all over the world to support the process of creating development plans and policies for urban areas. It does this by modelling urban development over time and space. The model is calibrated using historical trends and development theories, which allows it to make projections on land use change and urban growth (Rasmussen, et al., 2015; UrbanSim, n.d.; Waddell, 1998).

The main focus points of UrbanSim are land use, transportation, public policy, and the environment. The interactions between these factors are determined during the simulation of growth. These factors were identified as having the biggest influence on how an urban area will develop. UrbanSim uses the representation of an urban area's real estate market along with how that interacts with the area's transport market. With this, it uses representations of the choices that are made by households, businesses and real estate developers along with how each of these things are affected by government policies and investments (Waddell, 1998; U.S. EPA, 2000).

UrbanSim consists of multiple sub-models or internal models and it provides tools to calibrate the different models that allow them to fit the specific study area. Tools to search for the model specifications that are best suited for the study area to which it is being applied is also provided. Once the models are calibrated and adapted to work for the specific study area, different scenarios can be applied to simulate the effect of the different development plans and policies (UrbanSim, n.d.). Simulations for three different geographic levels or representations are allowed. The three levels from the most detailed level to the least detailed level are a parcel-level model, census block-level model and a zone-level model. The zone-level model is the fastest way to implement an UrbanSim model for an area. These models differ in terms of what input data is used for the model as well as the level of detail that the input data is required to have (UrbanSim, n.d.; U.S. EPA, 2000).

The input data that is required to run UrbanSim are control totals for household data with information on for example income, age, and children. Control totals for employment data are also required and should be divided by Standard Industry Classification. A synthetic population dataset is required that consists of samples that are usually generated from census data and control totals. Datasets on land use and buildings are also required. This data should contain the land use classes, property boundaries, the type of buildings and the market value of those buildings. Another major dataset that is required is the transportation network in the area. This includes the road network, railway lines, bus routes and taxi routes that cover the area. The road network data is used by a transport model to determine a cost for the various routes. This costs-dataset is then used as an input dataset for UrbanSim. Lastly, other datasets like the study area boundary and important environmental information that can be used to set the development constraints, should also be used (De Palma, et al., 2014; U.S. EPA, 2000; Patterson, et al., 2010; Waldeck & Van Heerden, 2017).

2.3.3 Model structure

As mentioned earlier, UrbanSim is made up of various sub-models that each contribute differently to projecting the future growth of an area. The models are built using the data for the study area and the parameters used by each of the models are determined using advanced statistical methods. This ensures that the model accurately reflects the local conditions. The core models are the two relocation models, the two transitional models, the two location choice models, the price and rent models and the development model (Patterson, et al., 2010; Gallay, 2010).

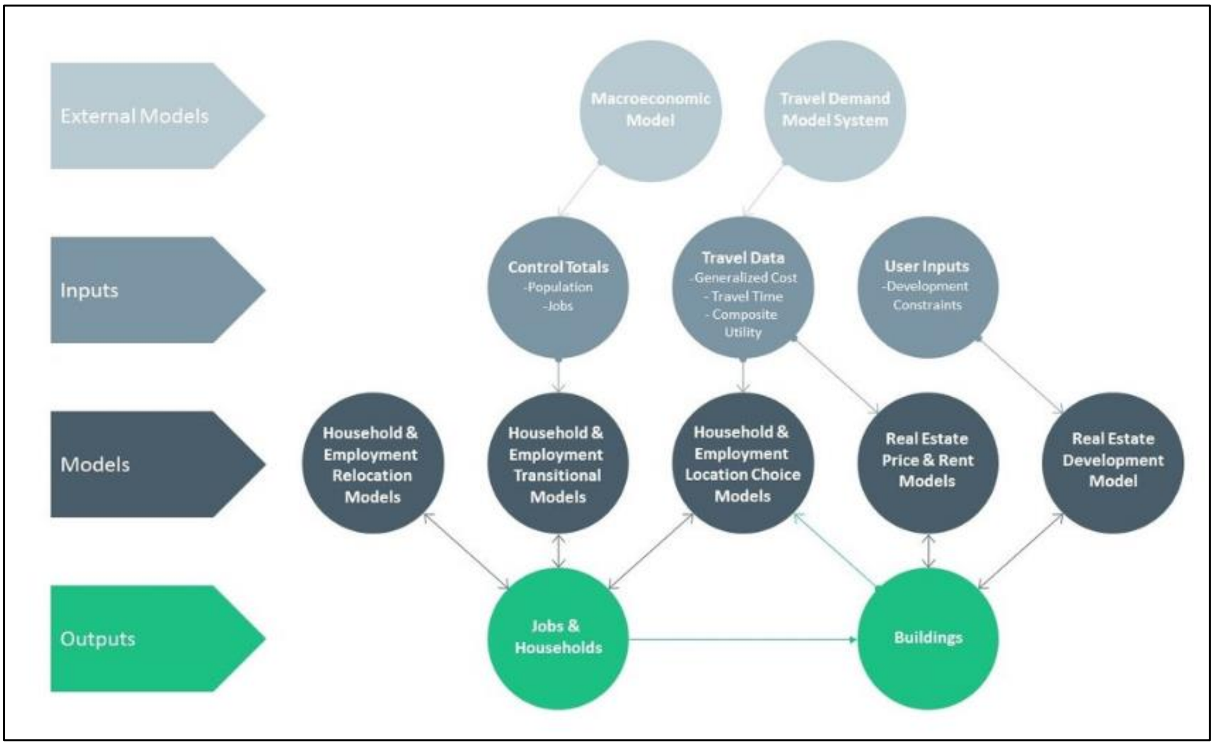

External Models
Inputs
Models
Outputs
Macroeconomic Model
Travel Demand Model System
Control Totals -Population -Jobs
Travel Data -Generalized Cost - Travel Time - Composite Utility
User Inputs -Development Constraints
Household & Employment Relocation Models
Household & Employment Transitional Models
Household & Employment Location Choice Models
Real Estate Price & Rent Models
Real Estate Development Model
Jobs & Households
Buildings

Figure 7: UrbanSim process (UrbanSim, n.d.).

The relocation models are the household relocation model and the employment relocation model. The household relocation model keeps track of the movement of households by generating a list of households that are not linked to a location and then moving these households to an area where vacant land is available. The probability that a household will relocate is established by user-defined external relocation probabilities. The employment relocation model is used to determine which of the individuals in the households will change their employment in each of the projected years. Once again, the probability of movement is defined by external relocation probabilities. The relocation probabilities are defined with the initial data inputs. The output from the relocation models is used by the location choice models (Van Heerden & Waldeck, 2015; Waddell, 2002; De Palma, et al., 2014).

The two transitional models are the household transitional model and the employment transitional model. The household transitional model generates a list of households that are required to be subtracted or added to the existing households. The employment transitional model handles the newly created employment opportunities as well as the removed employment opportunities in each employment sector. The growth and decline of employment opportunities in each sector that is retrieved from economic forecasts are used to determine the change in employment opportunities. The years used for these forecast will change from project to project. There will usually be a base year (Year that has already passed and data is available for) and then a year in the future (example the year 2050) until which the projections will run. Employment data for each year between the base year and future year will be provided. The output from the transitional models is also used by the location choice models (Waddell, 2002; Van Heerden & Waldeck, 2015; Gallay, 2010).

The real estate price and rent model simulates the change in land prices over time. It makes use of regression to determine the price of land by taking attributes such as land use types and neighbourhood characteristics into consideration. Many assumptions are made during this section of the simulation. One of the main assumptions that are made is that developers, businesses and households are all price takers meaning that they do not affect the market price. Another is that the annual price adjustments are used to match aggregate supply and demand over time. The output from the real estate price and rent model is used by both the location choice model and the development model (Van Heerden & Waldeck, 2015; Waddell, 2002).

The location choice models consist of the household location choice model and the employment location choice model. The location choice models use the results from the transitional models and the relocation models to perform further analysis to determine how households decide where to live. The household location choice model makes use of the Multinomial Logit (MNL) model to assign the unplaced households. A set of explanatory attributes are used here to determine the location where the households should be placed or assigned. The employment location choice model also makes use of the MNL model to assign unplaced employment opportunities (Waddell, 2002; UrbanSim, n.d.; Van Heerden & Waldeck, 2015; Jjin & Lee, 2017).

The real estate development model is used to specify the choice made by developers in terms of where development should take place and in which areas redevelopment should take place. The model creates a list of alternative developments that can take place. A multinomial model is once again used. In this situation, the model is used to determine the probability of each alternative of development taking place (Kakaraparthi & Kockelman, 2011; Van Heerden & Waldeck, 2015).

2.4 Synthetic Population

One of the most important and complex datasets that are required by many land use change models, including the UrbanSim model, is the synthetic population.

2.4.1 Overview

All types of agent-based models attempt to project the future state of a system by finding patterns in the behaviour of individuals over time (Axhausen, et al., 2010). Often the execution of these models takes place in two phases. The first phase is the creation of the original set of individuals and setting up the initial state of the system. The second phase is where the projection part comes in, here the system's state and the individual's state are advanced using different timesteps (Axhausen, et al., 2010; Pritchard & Miller, 2012).

The initial set of individuals that is used by these models are often in the form of synthetic population datasets. A synthetic population is created when there is a need for more detailed information than the original dataset provides (Pritchard & Miller, 2012; Technical University Munich, 2016). The detailed information that is not available could be that there is no detailed information on what the relationship between different individuals are. It could also be that the data is not detailed enough in terms of the geographic level or scale that the current data is provided at (Simpson & Tranmer, 2005). The Technical University of Munich (2016) describes a synthetic population as a microscopic representation of the actual population that only matches the statistical distribution of the actual population rather than being identical to it. Geard et al. (2013) echo this statement by describing a synthetic population as a population that matches the structure and dynamics of the actual population. The synthetic population is mainly used to represent the individual actors whose decisions influence how the development will take place (Moeckel, et al., 2003).

A situation in which population synthesis is often applied is with census data (Axhausen, et al., 2010). With most census data, a lot of the spatial detail is removed to protect the privacy of individuals (Moeckel, et al., 2003). The linkages between individuals, households, and employment opportunities are also removed for privacy reasons or the data is not available. In some countries, data on an individual level is provided, but this data is only a sample of the original data. In both these situations, a process like population synthesis is required to build up the dataset to either provide less aggregated data or data that is a representation of 100 percent of the population (Lenormand & Deffuant, n.d.; Pritchard & Miller, 2012).

How the synthetic population is generated is often dependent on the purpose that the data is going to be used for (Geard, et al., 2013). Therefore, there are several methods currently available for creating synthetic populations, each with their strengths and weaknesses.

2.4.2 Existing population synthesis methods

The most well-known population synthesis method is Iterative Proportional Fitting (IPF) that was first proposed by Beckman et al. (1996) (Antoni, et al., 2017; Moeckel, et al., 2003). This method makes use of iterative steps and contingency tables to fit agents to a set of constraints (Založnik, 2011). The main aim of IPF is to convert one-dimensional data into multi-dimensional data. This is done by altering data represented as a matrix so that the total of the rows and columns are equal to the values of the original one-dimensional data. If these values are equal the process attempts to minimise the deviation. The original values that the totals of the rows and columns are compared to are usually control totals that are predetermined (Choupani & Mamdoohi, 2016; Moeckel, et al., 2003; Konduri, et al., 2016). When IPF is used it is important to note that for the results to be reliable the sample used in the process needs to be highly representative of the population (Namazi-Rad, et al., 2014).

One of the drawbacks of IPF that was identified by Simpson and Tranmer (2005) is that it is not able to handle different sized tables and it is not a multipurpose routine. This means that a user is required to edit the algorithm to suit these needs (Simpson & Tranmer, 2005). IPF is also not able to work with values of 0, which can lead to some errors in the results as the values have to be set to 0.1 or 0.01 which could influence any probabilities that are generated. Some researchers make use of Monte-Carlo sampling with IPF to "cancel out" some of the drawbacks of IPF (Choupani & Mamdoohi, 2016; Kim & Lee, 2015; Moeckel, et al., 2003).

The Iterative Proportional Updating (IPU) method was created as a solution to the fact that IPF only has one level of hierarchy and is thus unable to create a synthetic population that is accurate at both individual and household level, something that is required by most agent-based models. IPU allows the generation of households while improving the individual level distribution at the same time. This is done by controlling the multiple hierarchy levels while at the same time adjusting the weights to ensure that the distribution of the different levels matches as closely as possible. IPU is often used with IPF to provide the best possible results (Kim & Lee, 2015; Namazi-Rad, et al., 2014; Konduri, et al., 2016).

Synthetic reconstruction (SR) is another well-known population synthesis method. This method makes use of a deterministic algorithm to construct the population. Both disaggregated and aggregated data are used during the process of creating the synthetic population. The disaggregated data is used as the base dataset or seed dataset that is a sample that represents the target population. A weighting technique is then used along with the attributes of the seed dataset to populate the study area. During this step, the aggregated data is used as a reference for how the data should be distributed. Within this step, different techniques can be used for the weighting, depending on the conditions of the data. Some of the techniques that are often used are deterministic re-weighting algorithms or Monte Carlo sampling (Tanton, 2014; Huang & Williamson, 2001; Namazi-Rad, et al., 2014).

A problem that has been noted with this approach is that the efficiency decreases as the number of attributes increase. One of the solutions that have been identified is to approximate a joint distribution based on the aggregated and disaggregated data. Once this is done the units from the disaggregated data is used as this is the data that now represents the target population. This technique is similar to the IPF technique (Namazi-Rad, et al., 2014).

Combinatorial Optimization (CO), otherwise known as the reweighting approach, is often used as an alternative method to IPF and is also seen as a simpler technique than IPF. Once again, a seed dataset is created from disaggregated sample data. CO aims to create a list of values whose aggregate values match predefined target values. This is done by iteratively going through the list of values that is divided into zones and adding, subtracting or swapping values to reach a level of fit that is deemed appropriate by the user (Abraham, et al., 2012; Huang & Williamson, 2001; Pritchard & Miller, 2012).

This method is not only used to create a new synthetic population but is often also used to refine an existing population or to modify a population where circumstances have changed that influences the weighting (Abraham, et al., 2012; Pritchard & Miller, 2012). One major difference between CO and IPF is the fact that IPF works with contingency tables and CO makes use of lists in the place of the contingency tables (Abraham, et al., 2012; Namazi-Rad, et al., 2014).

Hill Climbing (HC) is a population synthesis method that makes use of stochastic data-driven procedures to construct the synthetic population. HC is a mathematical optimisation technique that makes use of random search, which is a technique that is greatly used in computer science. HC starts by creating a random solution from the seed data and builds on that solution until it finds a solution that maximises the objective function. This carries on until the solution is as close as possible to the distribution of the real population (Namazi-Rad, et al., 2014; Tanton, 2014; Konduri, et al., 2016).

Since the background on how UrbanSim was developed, how it works, as well as where the outputs for UrbanSim can be used is now understood, the next step was to research specific examples of how the disaggregation of data for UrbanSim has previously been done.

2.5 Existing disaggregation methods

The review of existing disaggregation methods was done by broadly examining disaggregation methods that have been used in UrbanSim research and examining disaggregation methods that are more applicable to the South African context.

2.5.1 Requirement for further research

From the previous sections, it was identified that with most urban growth modelling processes, traditional synthetic population methods, like iterative proportional fitting, are used to disaggregate the data. The disaggregation is done by making use of control totals that are usually available at different geographic levels (i.e. defined by census) and a sample of detailed data that is available on a disaggregated level (e.g. on person level). The sample of data is used as seed data and using different synthesis methods, this seed data is fused with the control totals to provide a disaggregated dataset.

Traditional methods for creating synthetic populations cannot be used to disaggregate employment data in the South African context, as the detailed sample data that is required is not available. The census questionnaire does not require individuals to specify where they work. This means it is not known in what location an individual works, but only how many working individuals there are at an aggregate geographic level (e.g. provinces, district municipality, local municipality, main place, and sub-place). Figure 8 and Figure 9 show the different geographic levels that are demarcated by Statistics South Africa (StatsSA). The data used in the maps were the official StatsSA boundaries provide by StatsSA. Figure 8 illustrates the change from the highest demarcation level (i.e. province) to the lowest and most detailed demarcation (i.e. sub-place). Figure 9 provides a more thorough look at the change in detail from local municipal boundary, to main place boundary, to sub-place boundary.

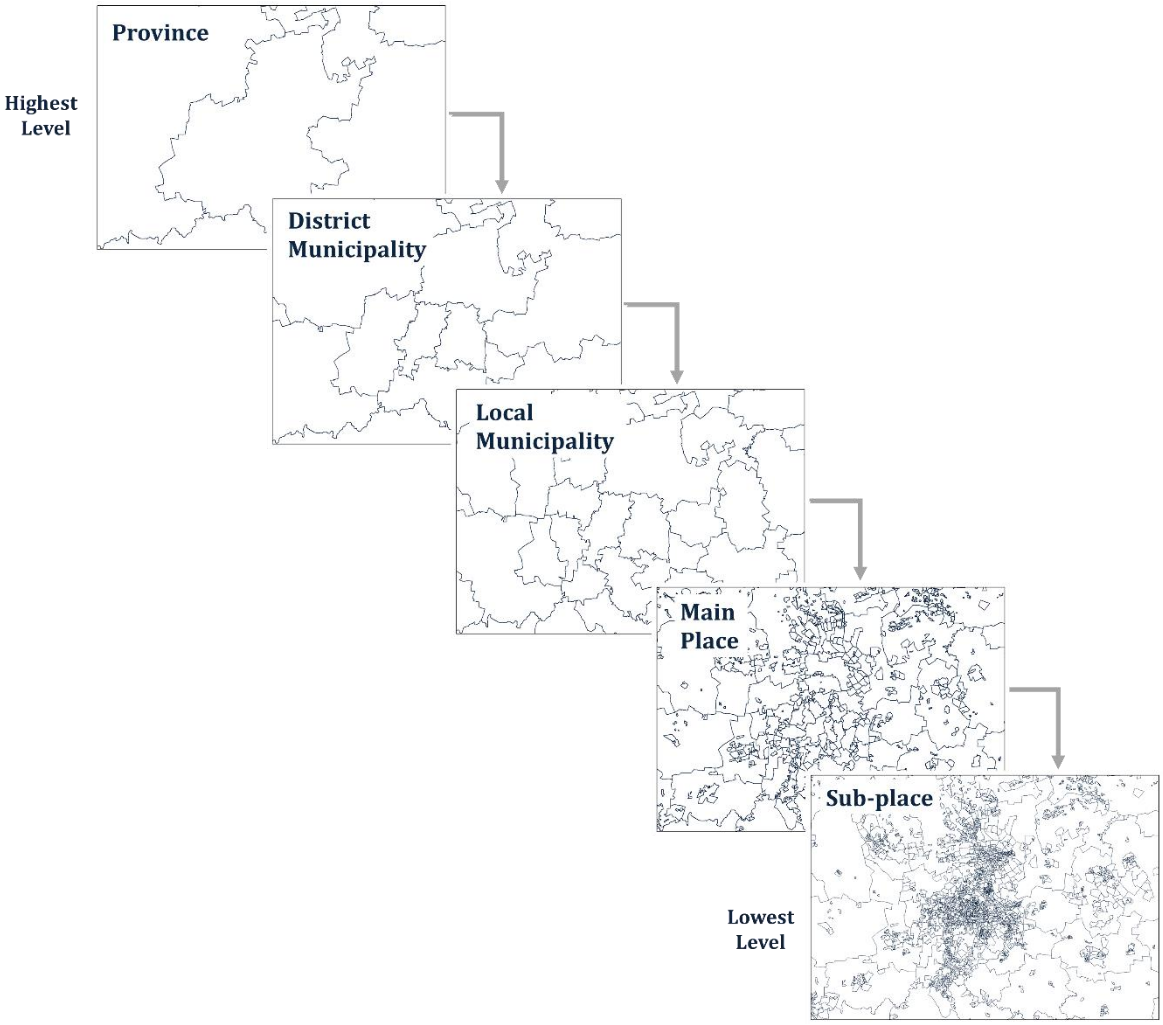


Figure 8: Different levels of Census demarcations

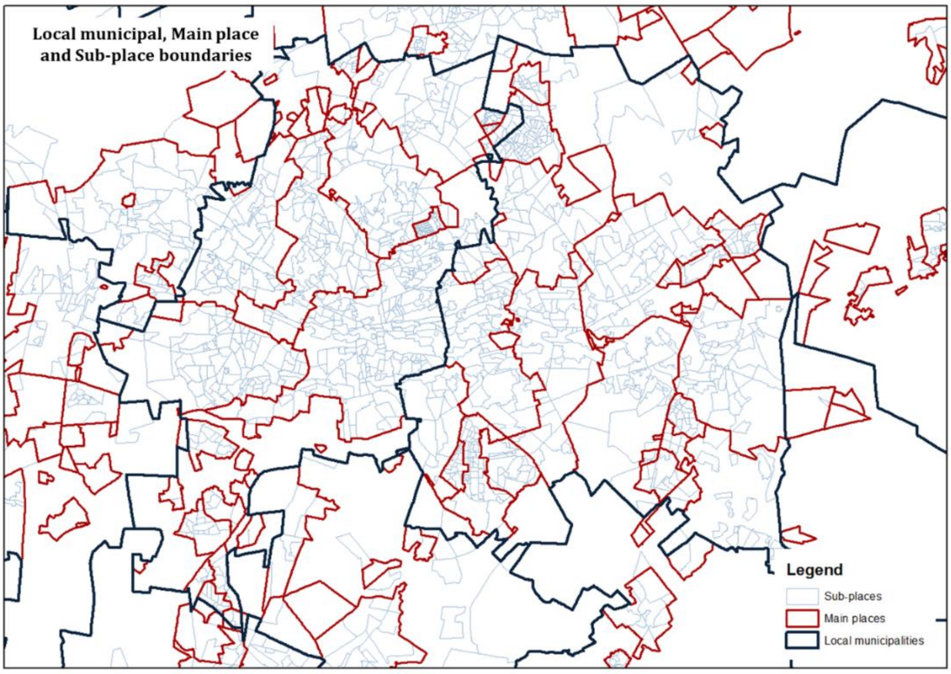


Figure 9: Detailed view of the final three demarcations

2.5.2 Literature reviewed

Five different methods for disaggregating data was examined:

1. Three Methods for Synthesizing Baseyear Built Form for use in Integrated Land Use-Transport Models. Abraham et al. (2005) identified a problem in the accuracy of data on floor space inventory in three land use transportation modelling projects that were performed in the United States. The floor space data was incomplete and it showed inconsistencies with both population and employment data. The research looked at three different methods that could be used to synthesize a built form input dataset at a micro level that could be used by land use transportation models. An algorithm was created for each of the three land use transportation projects (Abraham, et al., 2005).

The first algorithm was the Oregon approach that had three main stages for assigning the data to individual grid cells. The first two stages included an analysis of the demand and supply floor space on both the macro and micro level. The last stage used the developed algorithm to assign the development types as well as the quantity of space. The second algorithm was the Sacramento approach which consisted of six steps that resulted in the assignment of the total inventory of floor space to suitable parcels. The last algorithm was the Oahu approach that consisted of three main steps, each with sub-steps. In this approach, the results were jobs assigned to certain grid cells as well as synthesised floor space in grid cells (Abraham, et al., 2005). This approach shows the advantage to using floor space when disaggregating data.

2. Disaggregate models with aggregate data: Two UrbanSim applications. Patterson et al. (2010) looked at two methods for using UrbanSim when there is only aggregated data available for the study area. Two case studies were used to test different methods using Brussels and Lyon as study areas. The Brussels case had a problem with a lack of individual building data. The Lyon case only had employment data that was not detailed and aggregated. As with the Brussels case, there was also no building data available.

In the Brussels case study, the data were disaggregated from zonal level to grid cell level, using various steps. During this process, fictional buildings were created after the data was assigned to the grid cells. Throughout the process, many problems occurred which led to errors in the data that had to be fixed. This led to the process not being efficient (Patterson, et al., 2010).

For the Lyon case study, similar steps to the Brussels case was performed, but the grid cells were created to be the same number as the zones in which the data was originally in. This change in the method saved time in the process of disaggregation as many of the errors from the Brussels case were avoided. In the end, it was concluded that UrbanSim could be applied using aggregate data, but applying it in analysis was not advised. Since the disaggregation was performed using manual steps instead of an algorithm, the disaggregation method used by Patterson et al. (2010) was not as efficient as the studies done by other researchers that were investigated.

3. A heuristic combinatorial optimisation approach to synthesising a population for agent-based modelling purposes. Huynh et al. (2016) developed a heuristic approach to synthesise a population for New South Wales in Australia by using a sample free method. A two-stage method is used to allocate individuals to households by making use of the demographic data of both the individual and household data. In the first stage, a set of constraints is used to do a heuristic allocation that restricts which individual types can be linked to which household types. The second stage makes use of a combinatorial approach to allocate the individuals that remain to households based on a range of demographic attributes. In this stage, there are also added constraints that filter the allocation. For example, the minimum and maximum age gap that is possible between the mother and child that is assigned to a household. The approach does not look specifically at employment data but some of the techniques could be used in disaggregating employment data (Huynh, et al., 2016).

4. Generating a located synthetic population of individuals, households, and dwellings. Antoni et al. (2017) made use of the MobiSim Population Synthesiser to synthesise the population, as well as assign the population to buildings for three cities in France. The method is divided into two steps. The first being the generation of the population that is linked to households. The second step is linking the population generated in the first step to the buildings. To create the household dataset the adult agents are first assigned to a household based on various attributes such as marital status. After this, the children are assigned to households based on various attributes such as whether the household is a family household or a single household. For the synthesis, no micro samples are used as this type of data is not always available in France and other European countries (Antoni, et al., 2017).

When a household is linked to a building, the capacity of the building along with the type of building is used to randomly link a building to a household. The result is a final dataset where each individual is linked to a household and each household is linked to buildings and each building is linked to a specific geographic area within the study area. Once again, the approach does not specifically look at employment data but the idea is very similar to the objectives of this study, which is linking an employment opportunity to a building which is then linked to a specific geographic area (Antoni, et al., 2017).

5. Determining the Place of Work for Urban Growth Simulation. Waldeck and Holloway (2016) also identified the problem that there are no data available in South Africa about where people work. Therefore, they developed a method that could be used to create a dataset of where individuals are employed in South Africa. This method made use of building valuations to serve as a proxy for the number of jobs that can be located in a building. For this method, they determined a correction factor for each building type that was multiplied by the valuation of the building. They then disaggregated the number of jobs proportionally by using the valuation of the building divided by the total valuation of the buildings in the city that are not used for residential purposes.

The total number of jobs per building type was then compared to the total number of jobs in the corresponding employment sector. If there was a difference in the values, the correction factor was adjusted and the process was repeated until an acceptable difference was reached. Since this is a South African example, it is the most closely related approach. The validation process that is used could possibly be implemented for this study. The factors considered in the disaggregation process could also be considered for this study (Waldeck & Holloway, 2016).

2.5.3 Summary of methods

Out of the various methods that were reviewed, a component in each could contribute to developing the methodology for this dissertation. Patterson et al. (2010) research indicated clearly that manually disaggregating data was not that efficient and that the use of an algorithm or an automated process would increase the efficiency and repeatability of the disaggregation. The research performed by Abraham et al. (2005) also showed that using algorithms for the disaggregation is more effective. One of the other factors that also affected the research done by Patterson et al. (2010) was the lack of building data. Therefore, it is important to use actual building data, if possible, instead of creating fictional or proxy buildings. Doing this could increase the accuracy of where employment opportunities are located in the study area.

Abraham et al. (2005) research revealed that the floor space of a building is an important dataset when trying to determine the capacity of a building. Huynh et al. (2010) and Antoni et al. (2017) research showed that it is possible to use the various attributes that are available with the employment data to allocate the employment opportunities. Possible attributes that could be used are the economic sectors that employment opportunities fall in. These economic sectors could be linked to the type of building or to the type of land use that is implemented where a building is located. These types of attributes could be used to restrict what types of employment opportunities could be linked to a building.

Another factor that could be used for the allocation is the capacity of the buildings to assign a certain number of employment opportunities, as Antoni et al. (2017) did with allocating the households. This is an important factor to consider when identifying a building dataset to use for the disaggregation of the employment data. Finding data that has the capacity of the buildings or data that could be used to calculate the capacity of the buildings could greatly assist in disaggregating the employment data as accurately as possible.

If data about the capacity of buildings is not available, a similar approach to Waldeck and Holloway (2016) could be followed. This would mean using some kind of proxy to determine how many employment opportunities could be allocated to a building. The proxy that could be used for this would depend on what data is available for the buildings.

2.6 Evolutionary algorithms

Evolutionary computation is a collection of algorithms that make use of theories found in biological evolution, such as natural selection, to identify optimal solutions for a problem. What makes evolutionary computation techniques so popular, is the fact that they can be applied to solve a wide range of problems that occur in various fields of study. Some of the most well-known subfields of evolutionary computation include Genetic Algorithms, Evolutionary Algorithms, Differential Evolution, Swarm Intelligence, and Cultural Algorithms to name a few (Eiben & Smith, 2015). Out of these subfields, the evolutionary algorithm was chosen to use as the basis for the development of the allocation algorithm.

Evolutionary Algorithms (EA) use heuristic approaches to solve complex problems by using the basic principles of natural selection to find the most optimal solution to a problem. An EA initially begins with a population that consist of a certain number of individuals that each represent a tentative solution. Each of the individuals are then measured against an objective function and assigned a fitness value that indicates how suitable an individual is for solving the problem.

The objective function is customised to each specific problem and can be based on any measure that can be representative of the quality or accuracy of the solution. Those individuals that have the best fitness are selected to form parents. The parents are then reproduced using variation operators (e.g.

crossover and mutation) to generate new offspring. The offspring then replace some of the individuals from the original population and a new generation is created (Eiben & Smith, 2015). Figure 10 illustrates the process that most EAs follow.

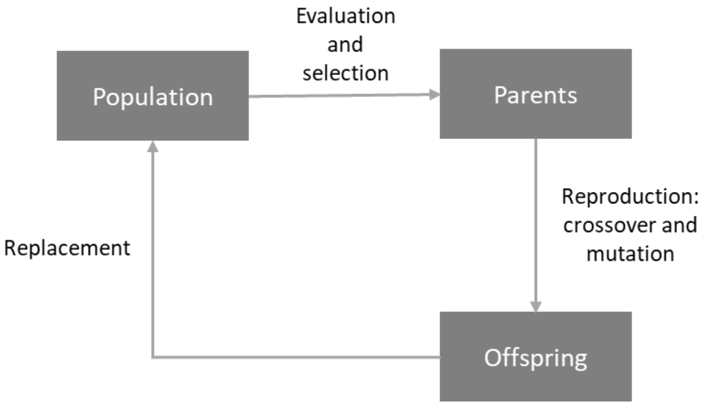
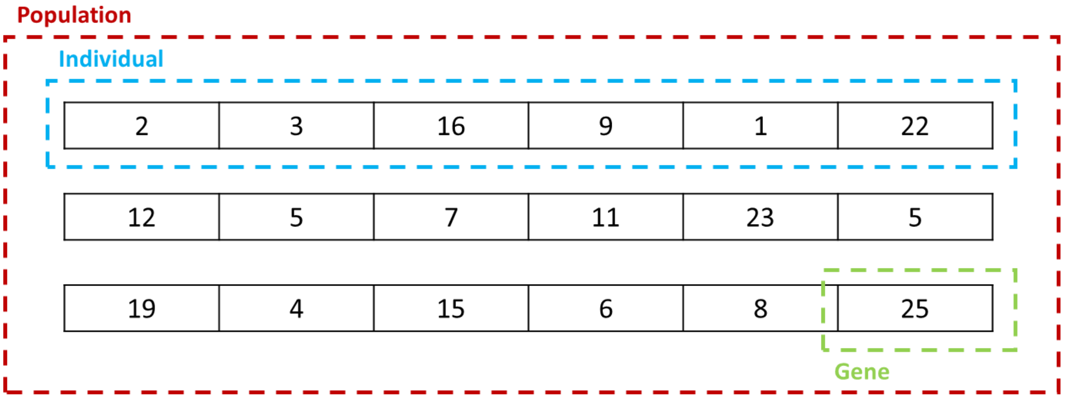


Figure 10: A generation in evolutionary algorithms (Talbi, E.G., 2009)

Figure 11 gives a visual representation of the components that a population consist of. The figure shows that a population consist of multiple individuals and an individual consist of multiple genes.



2	3	16	9	1	22
12	5	7	11	23	5
19	4	15	6	8	25

Figure 11: Diagram depicting the population, individual, and gene

The basic components that are required to perform the overall process that most EAs follow are:

1. Initialisation,
2. Selection,
3. Generation, and
4. Termination.

Initialisation

The initialisation component is the part where the initial population or the first generation is created. An important factor to consider when creating the initial population is to ensure that there is enough diversity in the population. If there is not enough diversity in the population, it could lead to premature convergence. Premature convergence is a condition in evolutionary algorithms where the algorithm converges to a solution before the global optimum solution is reached. This could lead to the algorithm not providing the most optimal solution, as it is trapped in the local minima because it did not have enough variety between the individuals.

There are two primary methods that are mostly used to create the initial population; these methods are random initialisation and heuristic initialisation. Random initialisation is when the initial population is created by using randomly generated individuals. Heuristic initialisation is when the individuals are created based on a known heuristic for the problem. The second component of EA is the selection process (Eiben & Smith, 2015).

Selection

In the selection process, the individuals in the population are evaluated against the fitness function to determine their fitness to be a viable solution. The fitness function provides a measure of how close an individual is to achieving the objectives of the algorithm by examining certain characteristics of an individual. There are two types of fitness functions that can be used, namely single-objective functions or multi-objective functions. A single-objective function is when there is only one measure of fitness that determines the suitability of a solution. Therefore, each individual is only assigned one fitness score.

A multi-objective function allows the use of multiple measures of fitness to find the most optimal solution. Therefore, each individual will be assigned multiple fitness scores, depending on the number of fitness measures used. When a multi-objective function is used, it is known as multi-objective optimisation. Once all the individuals of the population have been measured against the fitness function to determine their fitness, a certain number of individuals with the best fitness is selected. The next step is using these selected individuals to create a new generation by making use of variation operators (Eiben & Smith, 2015; Eiben & Smith, 2015).

Generation

The variation operators consist of two processes, namely crossover and mutation. Crossover is the process of combining two parents (individuals) to create one or two offspring that have genes from both parents. Some of the offspring will have improved characteristics as they have the best parts of both parents, but it is possible that the parts taken from the parents are the worst parts and will result in an undesirable combination. The percentage of each parent that is provided to each offspring is decided by random selection. Crossover is required to ensure that the offspring is not identical to the parent and can therefore possibly provide a better solution to the problem than the parents did. Figure 12 gives an example of the crossover process. The amount of crossover that occurs is determined by a specified probability of crossover taking place. Therefore, not all offspring are created by crossover (Eiben & Smith, 2015).

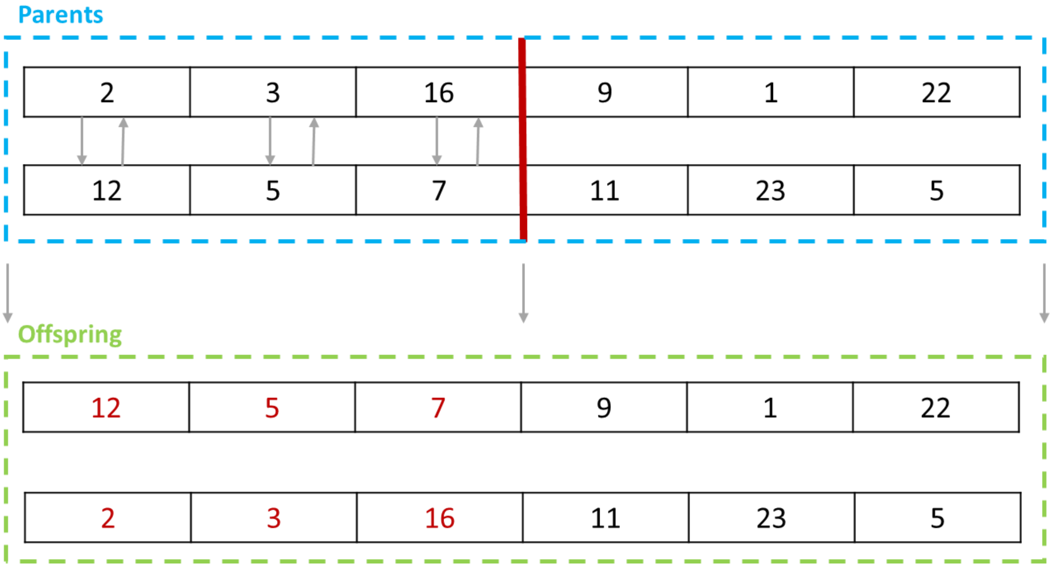


Figure 12: Diagram depicting crossover

While crossover takes place between two individuals, mutation is a unary variation operator that takes place on a single individual. With mutation, changes are made to the individual to ensure the offspring is not identical to the parent. There are various forms of mutation, some of with include swapping values or genes within an individual and adding or subtracting a specified value to or from the genes. Which gene within the parent is changed can be decided by making use of random selection. A probability distribution is specified to determine how many offspring should undergo mutation. A probability is also specified to determine how severe the mutation should be. Mutation is an important step to ensure that the algorithm does not become stuck in local extrema. Figure 13 gives an example of the mutation process.

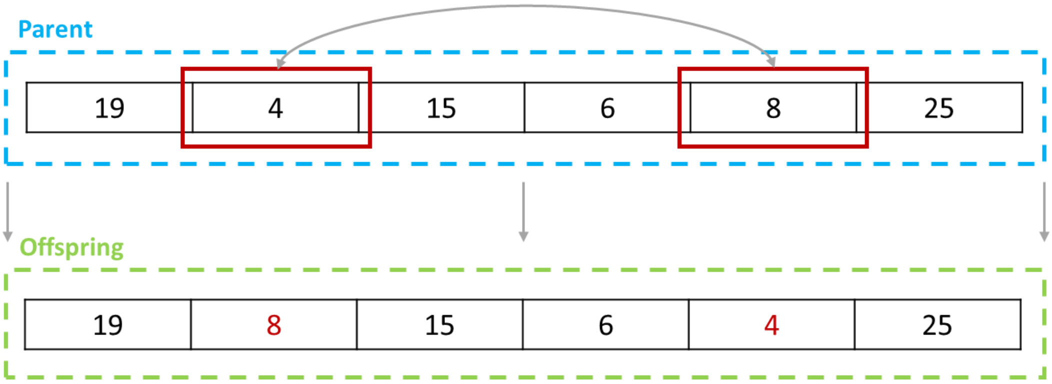


Figure 13: Diagram depicting mutation

Termination

The final component is the termination of the algorithm. The algorithm can be stopped using various methods. Some of the options include specifying a certain number of generations to be created, setting a threshold value for the improvement in fitness, setting a maximum CPU time, and setting a population diversity threshold. When termination occurs in single-objective functions, the most optimal solution is selected as the final solution (Eiben & Smith, 2015). This is not the case when multi-objective functions are used.

Often with the use of multi-objective optimisation, the objectives used to measure the fitness can be conflicting. Therefore, the optimal solution is not a single solution but a set of solutions that is known as Pareto optimal solutions. A solution can be classified as Pareto optimal when one objective cannot be improved without worsening another objective. The solutions that cannot be improved upon are known as non-dominated solutions or individuals. All the other solutions or individuals that are created during the algorithm's run, that have been improved upon, is known as dominated solutions.

In order to select the best solution when using multi-objective optimisation, a compromise needs to be reached between satisfying the conflicting objectives. This compromise is known as the Pareto front. The Pareto front consist of all the non-dominated individuals (Pareto optimal solutions). Figure 14 illustrates the typical shape of the Pareto front and shows the trade-off that exists between two objectives, in this case both minimising objectives.

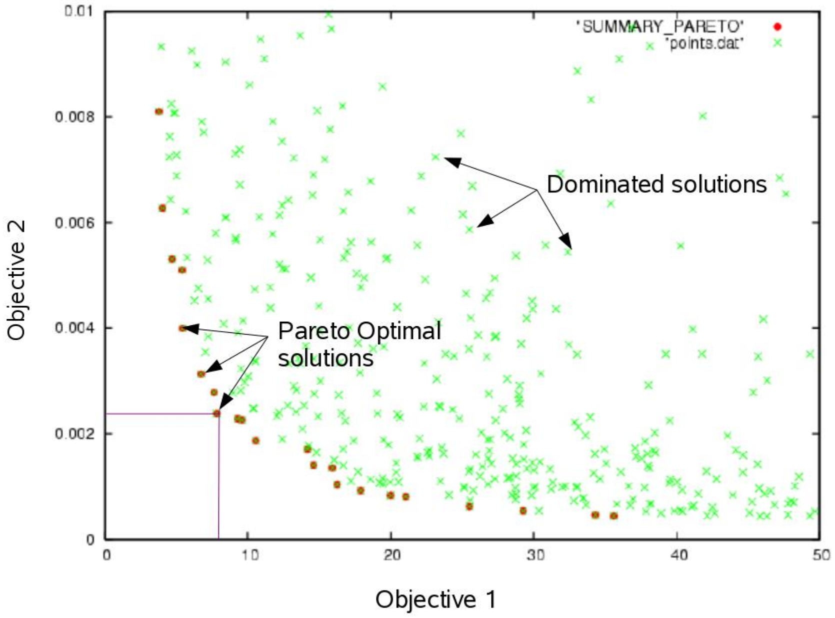


Figure 14: Shape of typical Pareto front (Cenaero, 2019)

Once the Pareto front has been established, it falls on the user to decide which of the Pareto front solutions should be the final solution. There are no commonly used methods for finding the global optimal solution, therefore the decision maker decides which objective is more important or carries more weight and from this the final solution is chosen.

Chapter 3: Methods

3.1 Introduction

This chapter discusses the main components of the methodology used for the research. The main components discussed in this chapter is the research design, the study area that was used as the test case, the main steps that make up the methodology, and the limitations and assumptions that were made.

3.2 Research design

The research design consists of both an empirical research method and a creative research method. Empirical research is described as methods that depend on observation. This method usually consists of, for example, surveys, experiments, and case studies. The empirical research is the secondary method of the research and is the literature study that is performed to learn about methods that are currently available for disaggregating data (Olivier, 2009). Creative research consists of methods that are used to develop new ideas or devices that can be used in computing. This method usually includes models, algorithms, and prototypes, for example. Creative research is the primary method of the research and is the development of the algorithm for the disaggregation of the data (Olivier, 2009). The strength of using both empirical methods as well as creative methods is that the literature survey allows a user to learn as much as possible about previous research on the subject which could provide some base to work from in the development of the algorithm.

3.3 Study Area

The study area is the City of Ekurhuleni (CoE), which is a metropolitan municipality located in Gauteng, South Africa. It is one of five districts that make up the Gauteng Province. The CoE is also one of only eight metropolitan municipalities in South Africa. Nine cities that were originally located in the former East Rand were integrated to form the municipality. The municipality covers a geographical area of 1 975km^2 and consists of 101 wards (Municipalities of South Africa, 2019). Figure 15 illustrates the location of the CoE within Gauteng.

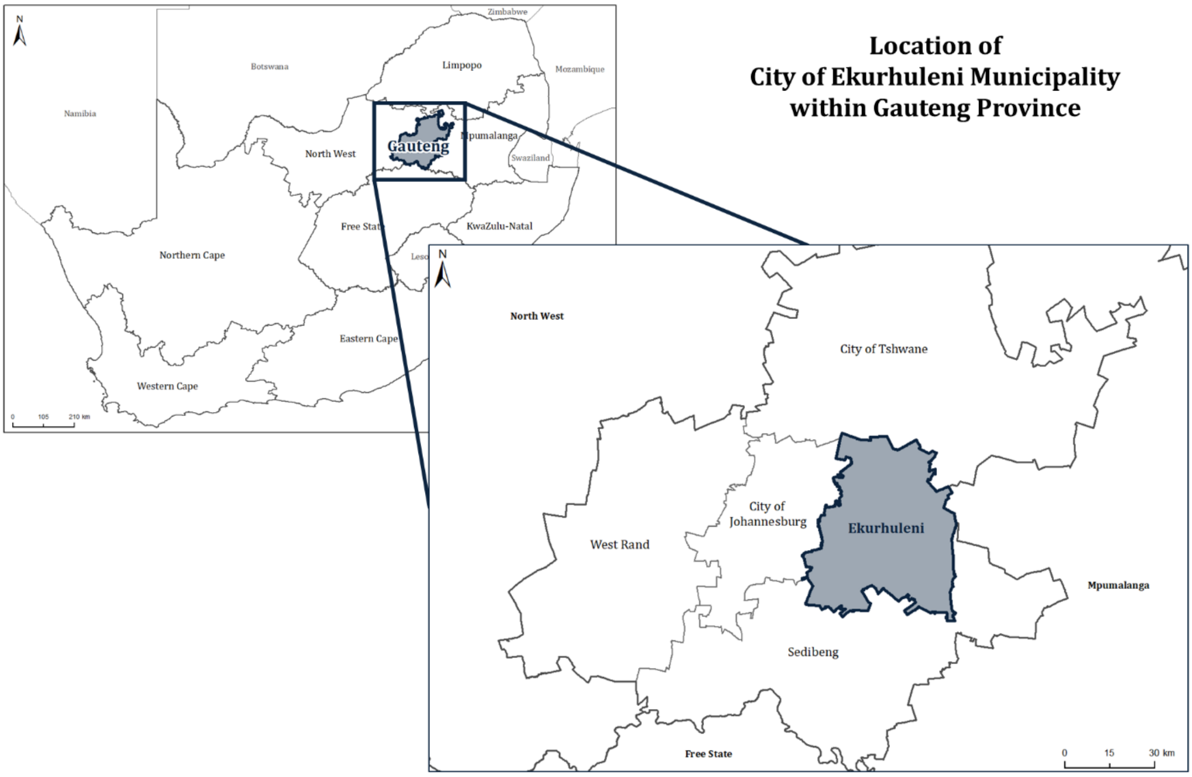


Figure 15: Location of City of Ekurhuleni within Gauteng Province

The CoE had an estimated population of 3 687 115 people in 2017, with a population density of 1 867 people per square kilometre. An estimated 2 553 382 of the population in 2017 were of working age. An estimated 1 326 663 people were employed in 2017, which is an estimated unemployment rate of 27.23% in 2017 (Quantec, 2019). The CoE is projected to have an increase of 1.7 million people by 2050, which is an estimated population growth of 60%. It is also projected that the CoE will be the 3rd fastest growing municipality in Gauteng (Le Roux, et al., 2019).

The CoE has a very diverse economy that contributes almost a quarter of Gauteng's economy. It is known as Africa's Workshop since many of the country's goods and commodity factories are located here. The CoE is also known as the transportation hub of South Africa as it is home to O.R. Tambo International Airport, which is the busiest airport in Africa. The CoE also has the largest railway hub in South Africa. Therefore, the CoE is not only a major employment hub in Gauteng but also in South Africa, which is one of the reasons why it was chosen as the study area (Municipalities of South Africa, 2019).

The many job opportunities make it the perfect area to develop and test the methodology for disaggregating employment data. Another reason for choosing CoE as the study area is the fact that it is a metropolitan municipality and therefore more resources are available for the municipality as more research has been done using metropolitan municipalities rather than smaller municipalities. This allows for the availability of a range of datasets that can be used in the development of the methodology. Figure 15 depicts the location of the CoE within Gauteng. Figure 16 and 17 are maps that show the various metro regions and the main cities located in the CoE.

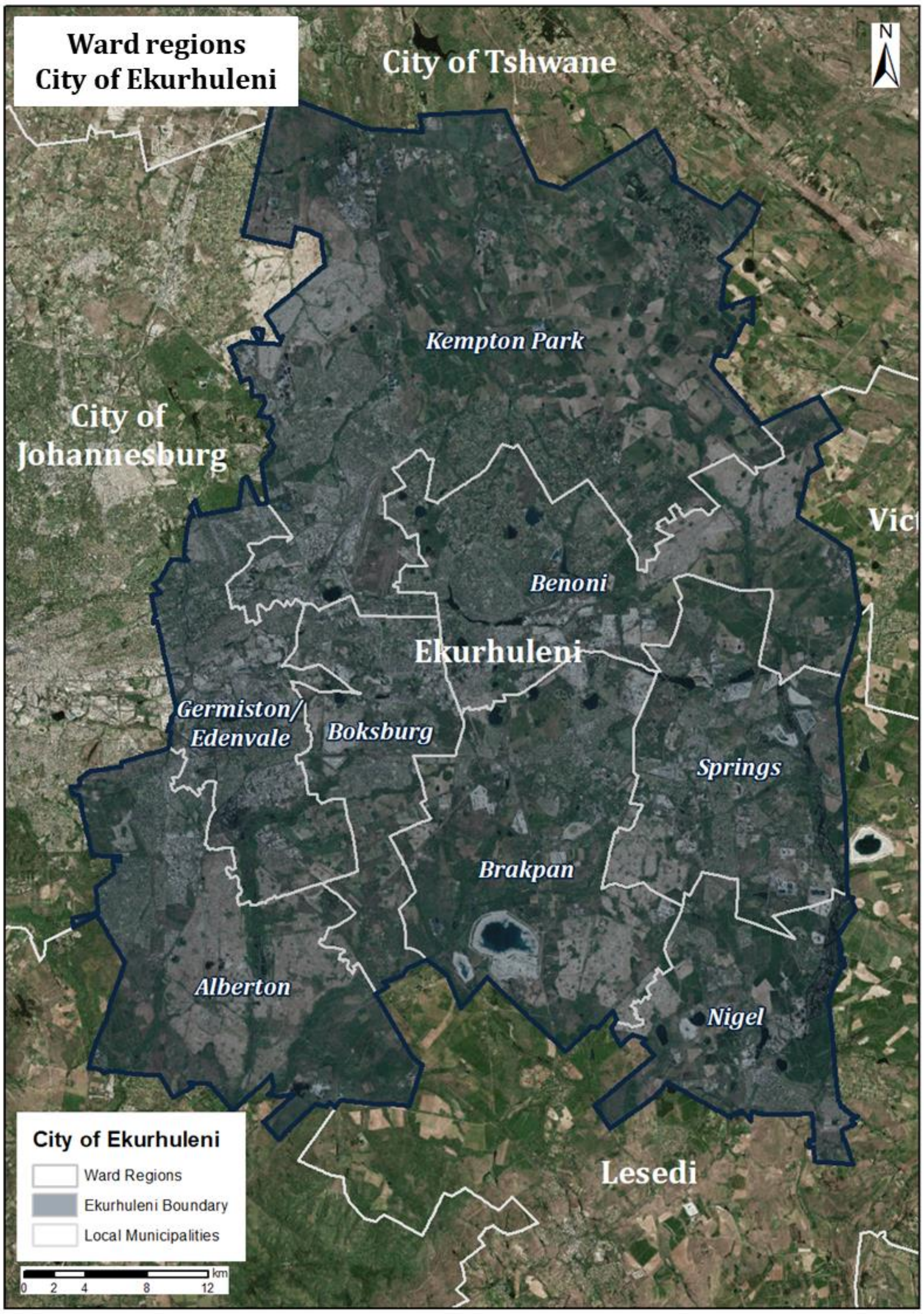


Figure 16: Metro regions in the City of Ekurhuleni

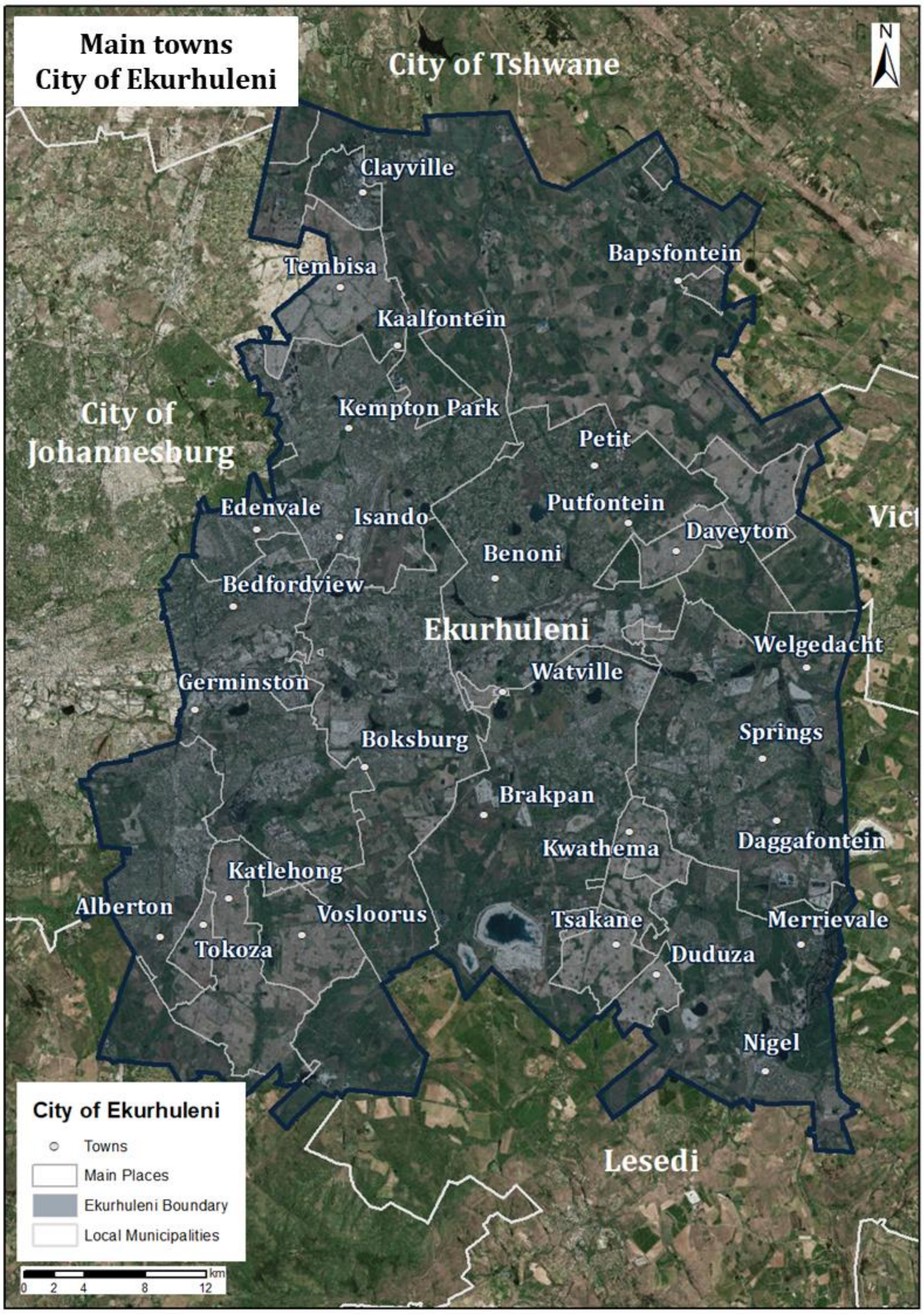


Figure 17: Main towns in the City of Ekurhuleni

3.4 Methodology

The methodology consists of six main components. Each of these components translates to a step in the final methodology. The main steps that were performed to develop the methodology include the following:

1. The first step was finding any existing disaggregation methods that were relevant to see if they had components that could be used to develop the methodology.
2. The second step was identifying the datasets that were required for allocating the employment data.
3. The third step was preparing the data. During the data preparation, a base dataset was created that was used to develop the employment allocation algorithm. This step included integrating all the datasets that were identified in the second step.
4. The fourth step was the development of the employment allocation algorithm that makes use of the base dataset to allocate the available job opportunities to the different buildings. This step included the validation of the results of the allocation algorithm to determine its accuracy.
5. The fifth step focused on discussing the results of the algorithm and discussing what the benefits of more detailed employment data are.
6. The final step was to run the algorithm for two other metropolitan municipalities to see the compatibility of the algorithm with other study areas.

Figure 18 shows the six major steps and which objectives for part of each step.

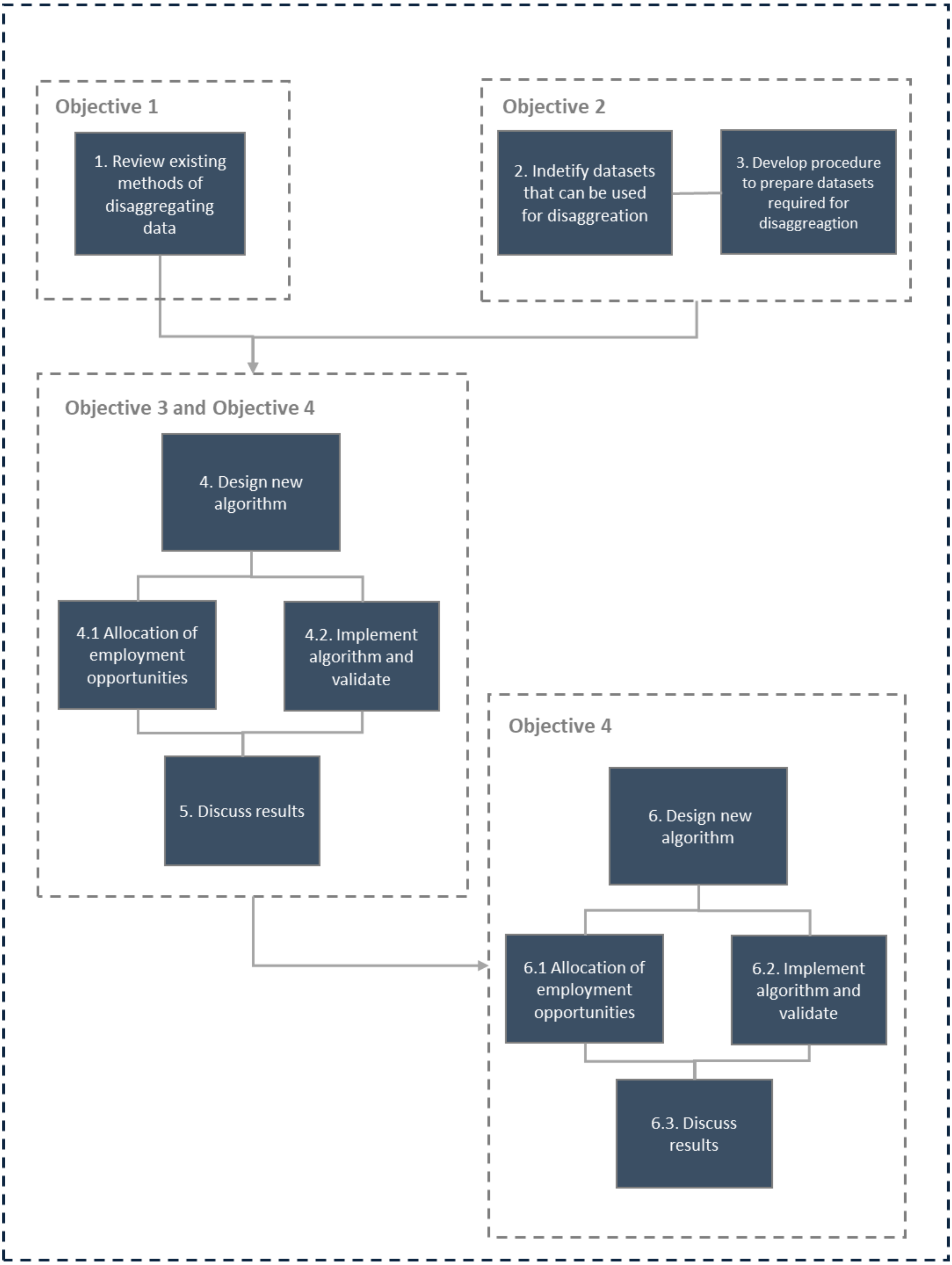

Objective 1
1. Review existing methods of disaggregating data
Objective 2
2. Indetify datasets that can be used for disaggreation
3. Develop procedure to prepare datasets required for disaggreagtion
Objective 3 and Objective 4
4. Design new algorithm
4.1 Allocation of employment opportunities
4.2. Implement algorithm and validate
5. Discuss results
Objective 4
6. Design new algorithm
6.1 Allocation of employment opportunities
6.2. Implement algorithm and validate
6.3. Discuss results

Figure 18: Methodology

Chapter 4: Data preparation

4.1 Introduction

This chapter discusses the various datasets that were used in developing the methodology for the disaggregation of the employment opportunities.

4.2 Overview of data used

The first step to developing the algorithm that can allocate employment opportunities is to prepare the data that is used for the allocation. During the data preparation, it was required to develop a base dataset that could be used to develop the algorithm. To create the base dataset, all the datasets that are required to be able to calculate the total number of employment opportunities that a building could accommodate (i.e. employment capacity per building) needed to be identified. Many of these datasets were identified based on the literature study that was done in Section 2.5. Table 2 provides a summary of the gist of each study.

Table 2: Summary of existing methods

Authors	Research topic	Relevant takeaway	Data used
Abraham et al. (2005)	Three Methods for Synthesizing Baseyear Built Form for use in Integrated Land Use-Transport Models	Using algorithm improves efficiency. Floor space is important factor when calculating capacity of building.	Floor space Population totals Employment totals
Patterson et al. (2010)	Disaggregate models with aggregate data: Two UrbanSim applications.	Manual disaggregation is not efficient. It is better to use actual building data if available.	Building proxy's Household totals Employment totals
Huynh et al. (2016)	A heuristic combinatorial optimisation approach to synthesising a population for agent-based modelling purposes.	The attributes of the data can be used to assist in allocation.	Population demographic data Household demographic data
Antoni et al. (2017)	Generating a located synthetic population of individuals, households, and dwellings	Building capacity is good indicator to use for allocation. The attributes of the data can be used to assist in allocation.	Population demographic data Household demographic data Building data
Waldeck and Holloway (2016)	Determining the Place of Work for Urban Growth Simulation.	Proxy data can be used to determine number of employment opportunities per building.	Building data Building valuation Employment totals

Patterson et al. (2010) research indicated that the use of proxy buildings was not effective and caused a lot of errors in the data preparation stage. Therefore, the first requirement of the base dataset was acquiring a buildings dataset. The buildings dataset consisted of all the residential and non-residential buildings that were located in the CoE in 2012. The use of buildings from 2012 is appropriate as the last Census in South Africa was performed in 2011. Any employment data that is available would thus also be from 2011.

Antoni et al. (2017) indicated that the use of building capacity proved to be very beneficial in the disaggregation process. Therefore, a dataset of building footprints was acquired from which the capacity of the buildings could be determined. Waldeck and Holloway (2016) revealed that the use of proxy data to determine the employment capacity of buildings could also be effective. Therefore, the various building classes were inspected to ascertain whether proxy data was available for any of the classes to aid in determining the employment capacity. During this inspection, it was identified that there was auxiliary data on schools and police stations that could provide an accurate estimation of employment capacity.

The final dataset that was used was the number of employment opportunities per economic sector. This was in line with the research done by Huynh et al. (2010) and Antoni et al. (2017), in which certain attributes were used to perform the required linkages. Each of the buildings in the buildings dataset had a land use associated with it, which can be associated with a certain economic sector and this information could then be used in determining the employment capacity of a building.

The various datasets that were identified, were combined to create the final datasets. This dataset consisted of employment capacities for each building and these capacities were used to allocate employment opportunities. Table 3 shows the datasets that were used to develop the base dataset. The sections that follow discuss the process of creating the base dataset.

Table 3: Datasets used for the City of Ekurhuleni

	Dataset	**Data custodian**
1	City of Ekurhuleni residential building	Council for Scientific and Industrial Research (CSIR)
2	City of Ekurhuleni non-residential building	CSIR
3	City of Ekurhuleni building footprints	CSIR
4	Sub place boundaries for the City of Ekurhuleni	StatsSA
5	Main place boundaries for the City of Ekurhuleni	StatsSA
6	Metro regions in the City of Ekurhuleni	StatsSA
7	Local municipal boundary for the City of Ekurhuleni	StatsSA
8	Primary and High schools in the City of Ekurhuleni	Gauteng Department of Education (GDE)
9	Police stations in City of Ekurhuleni	South African Police Service
10	Employment per economic sector for the City of Ekurhuleni	Quantec

4.3 Base dataset for employment allocation

The base dataset was developed as the starting point for the employment allocation. As mentioned in the previous section, various datasets were integrated to create the base dataset. From the literature study in Section 2.5, it was identified that certain information was required to be able to allocate employment opportunities to a specific building. The following four requirements were identified:

1. Location and use of the buildings,
2. Size (i.e. area) of the buildings,
3. The economic sector that the employment opportunities are linked to, and
4. Employment capacity of the buildings.

Figure 19 illustrates the four requirements and in which subsection of the chapter the requirement is being discussed.

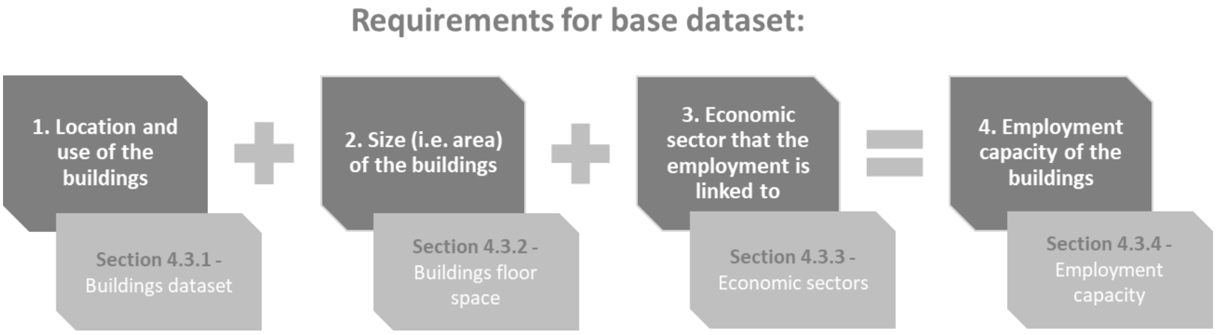

Requirements for base dataset:
1. Location and use of the buildings
Section 4.3.1 - Buildings dataset
2. Size (i.e. area) of the buildings
Section 4.3.2 - Buildings floor space
3. Economic sector that the employment is linked to
Section 4.3.3 - Economic sectors
4. Employment capacity of the buildings
Section 4.3.4 - Employment capacity

Figure 19: Summary of the requirements for the base dataset

4.3.1 Buildings dataset

Requirements for base dataset:
1. Location and use of the buildings
2. Size (i.e. area) of the buildings
3. Economic sector that the employment is linked to
4. Employment capacity of the buildings

Figure 20: First requirement of base dataset

The first requirement for the base dataset is illustrated using the darker shade in Figure 20. The fundamental dataset that was required to be able to disaggregate the employment data to building level, was a buildings dataset that had the location of all the buildings in the CoE that have employment opportunities available. Accordingly, a dataset that consisted of all the buildings located in CoE in 2012 was used. As mentioned in the previous section, the existing buildings from 2012 was used as this is the closest data that could be used to link the available employment data, which is data from the 2011 Census. The buildings dataset consisted of around 983 000 buildings for the CoE.

The buildings dataset consisted of the location of the buildings as well as the underlying land use. Each building has a primary land use class that broadly indicated what a building is used for and a secondary land use class that provides more detail. The dataset had 11 primary land use classes and these classes were further sub-divided into 47 secondary land use classes. These classes are shown in Table 4.

Table 4: Different building classes

	Primary building classes	Secondary building classes
1	Mining	Mine related buildings
2	Transport	Parking garages Vehicle storage areas Toll plaza Bus, taxi, other terminals Railway stations Airports
3	Utilities and infrastructure	Water storage and sewerage treatment plants Energy production and distribution Refuse disposal (landfills) Posts and telecommunications Access control
4	Health care facilities	Hospitals and clinics Health care - other facilities Animal clinics and sanctuaries
5	Education	Pre-school Primary school Secondary school Tertiary education institutions Other schools Other education institutions
6	Commercial	Shopping mall Shopping centre Commerce Petrol station and service station Office parks Informal trading
7	Industrial	Light industries and warehousing/distribution Heavy industries Fuel depots
8	Recreation and leisure	Amusement and show places Sports Facilities
9	Tourism	Holiday resorts and associated buildings Camping sites, caravan parks and associated buildings Hotels, guesthouses, and bed and breakfasts
10	Institutions	Research institutions Defense Police Emergency services Correctional services Government Foreign government
11	Residential	Free hold formal houses Informal structures Cluster/complexes Small holdings / agriculture Security estates

The buildings for five secondary land use classes were removed and replaced by ancillary datasets that provided more information than the buildings dataset. This allowed for a more accurate calculation of the employment capacity for those classes. These five secondary classes included the police buildings from the *INSTITUTION* primary class and four classes from the *EDUCATION* primary class, namely:

1. Pre-school,
2. Primary schools,
3. Secondary school, and
4. Other schools,

Police stations

The capacity for the police buildings was calculated by using an ancillary dataset that consisted of the location of each of the police stations within the CoE and the size of the population that the stations serve. As was the case with the *EDUCATION* class, the police buildings were replaced with the ancillary dataset and each station was now just represented by one building. This is because some of the buildings in a police station are used for storage and other activities and do not necessarily have any employment opportunities linked to them.

Education

A dataset with the location of all the GDE schools in the study area was used. The dataset included the location of the schools, the number of classrooms that each of the schools has, and the number of learners that each school has. This dataset was used to replace all the buildings that were classified with the four secondary classes mentioned above. Replacing the multiple buildings on school grounds with just one point representing all the buildings on school grounds could be done as the information was not needed for each building on the school grounds, but just for the entire school. This was once again appropriate as many of the buildings located at a school is used for activities like storing, sports activities, school halls and do not necessarily have any employment opportunities linked to them.

The final processing that was then required for the buildings dataset was just to join the buildings to any other relevant datasets, that included the sub-place, main place, and metro region that each building falls in, as well as the local municipality they fall in. A building ID was also created for each of the buildings to be able to link the allocated employment opportunities back to the spatial data in the end. Figure 21 shows a sample of the building dataset, classified according to the primary land use class.

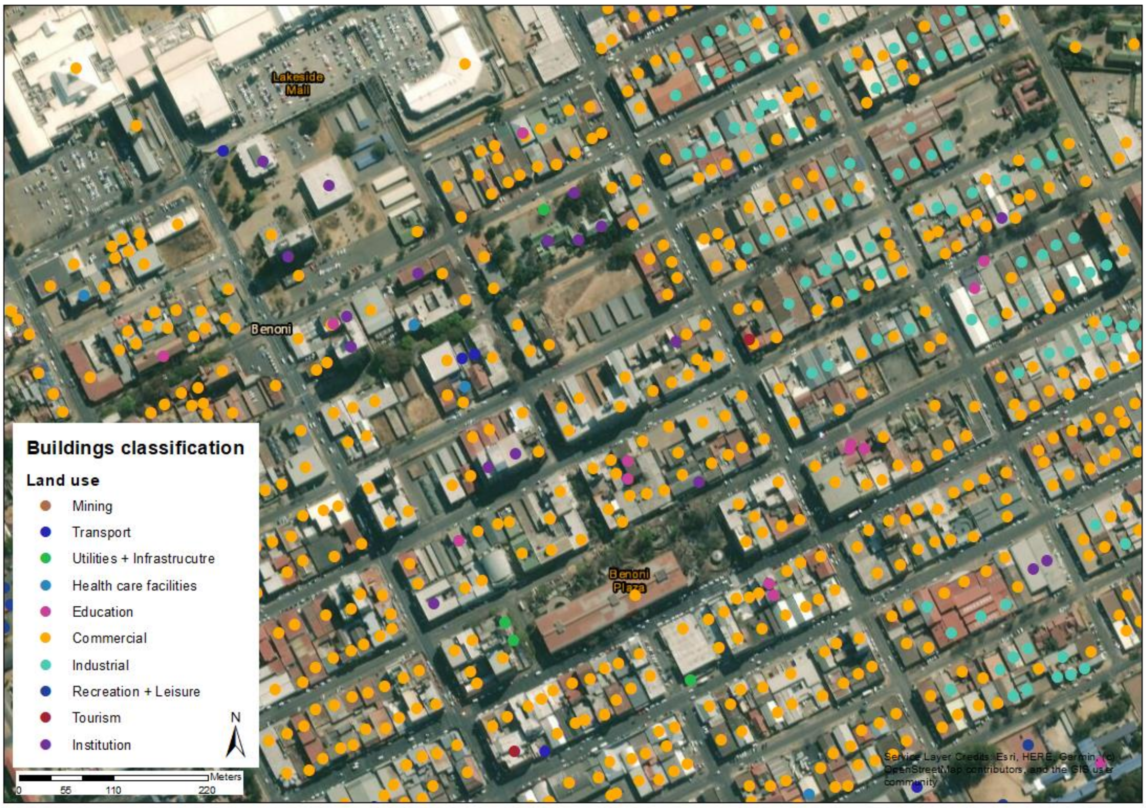


Figure 21: Sample of buildings dataset

Figure 22 shows the distribution of residential buildings across CoE. The darker shade of blue is indicative of areas with a higher density of buildings. Figure 23 shows the distribution of the non-residential buildings across CoE. The darker shade of red is indicative of areas with a higher density of buildings. The distribution of the non-residential buildings is of more importance as most employment opportunities are in non-residential buildings. However, the residential buildings are also relevant as there are also some employment opportunities in residential buildings, such as home-based jobs or jobs for domestic workers.

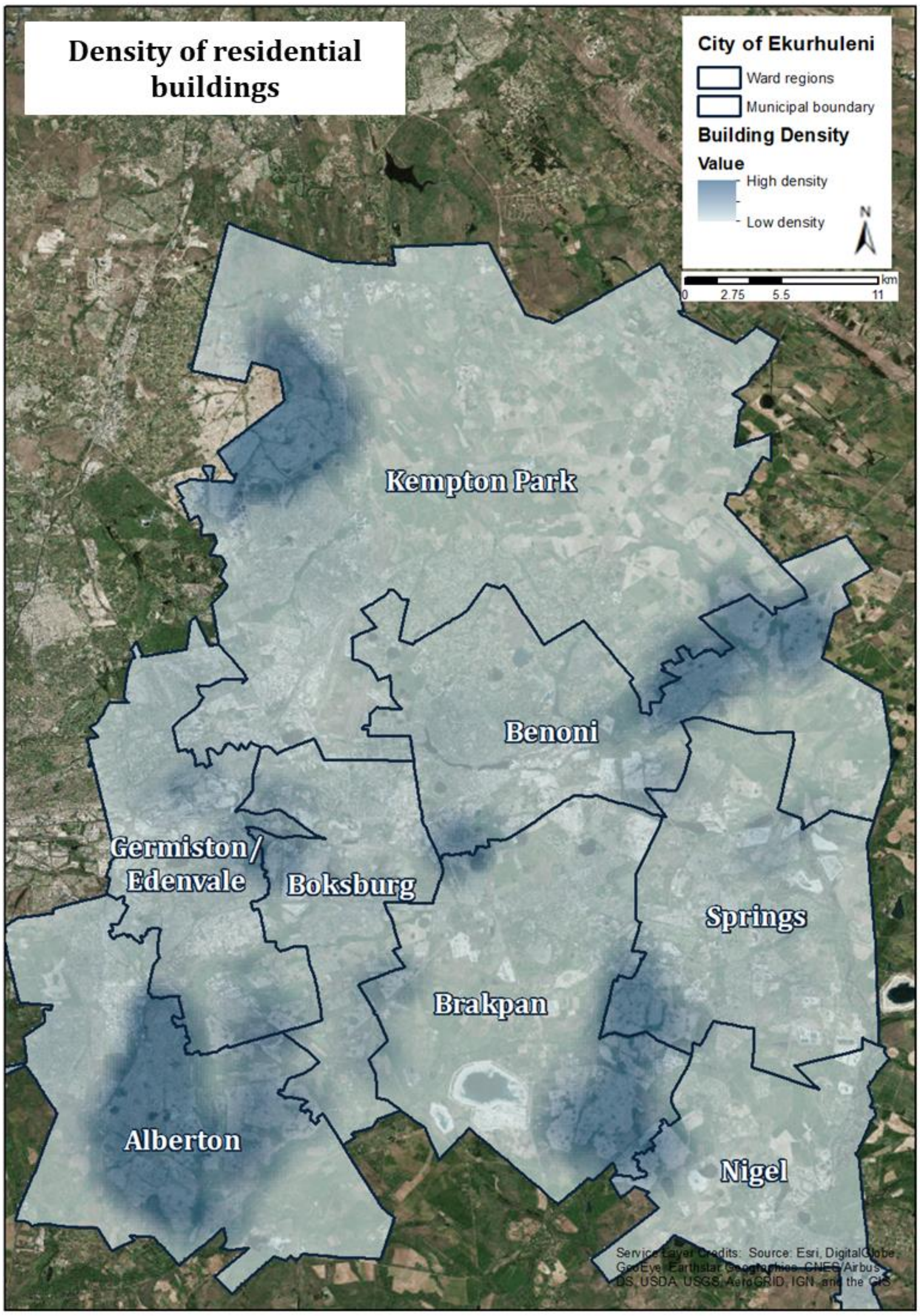


Figure 22: Map showing areas with a high density of residential buildings

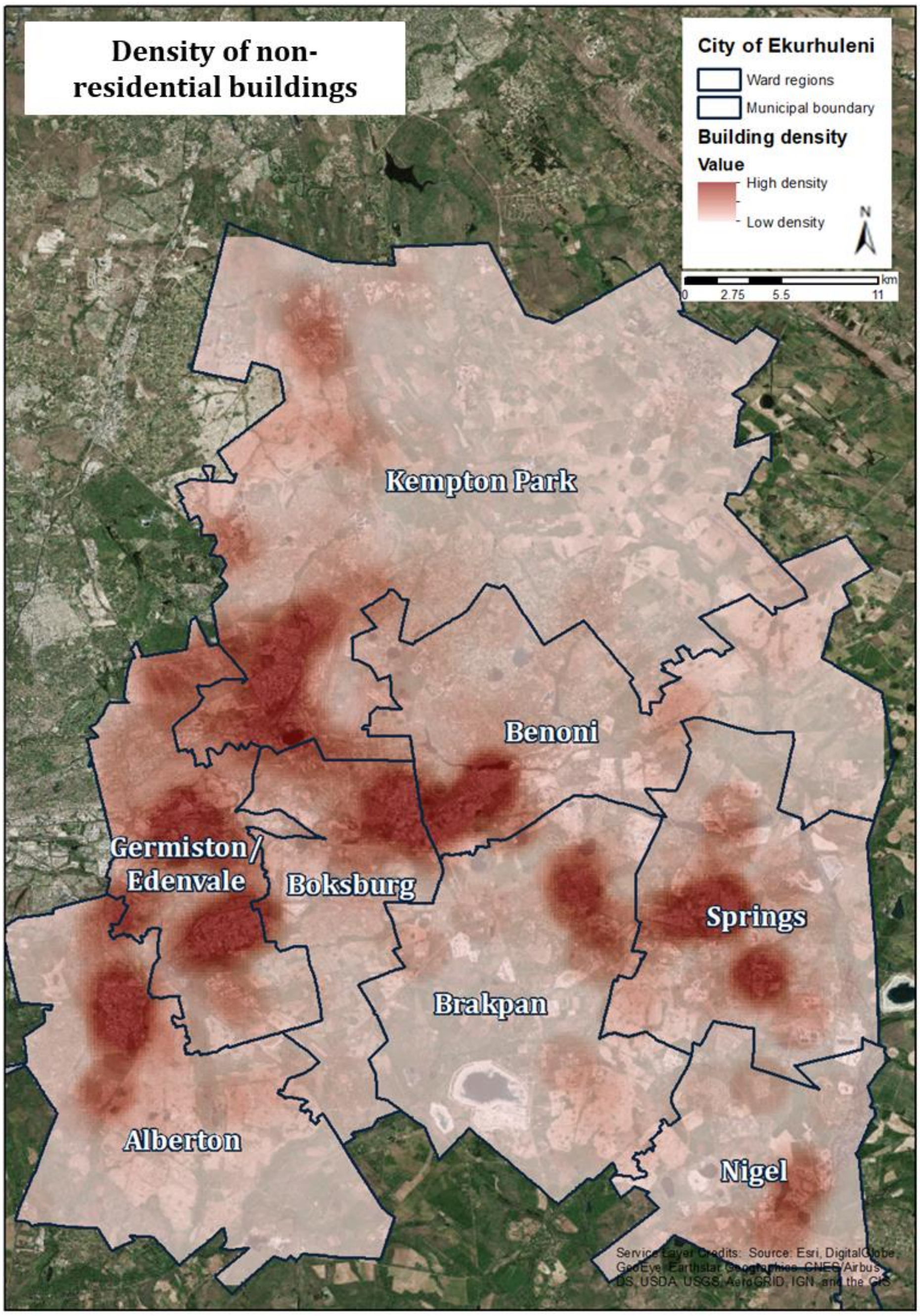
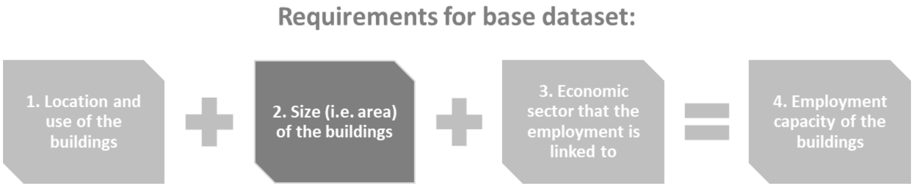


Figure 23: Map showing areas with a high density of non-residential buildings

From Figure 22 it can be seen that the areas with the highest residential building density are within the Alberton metro region. Other areas with high density can be found in the Benoni, Brakpan, and Kempton Park metro regions. From Figure 23 it can be seen that the areas with higher non-residential building density are located more to the centre of the municipality. The Germiston/Edenvale metro region shows the highest non-residential density.

Unlike with the residential density, all the metro regions have areas of dense non-residential buildings. This is because in all areas there tend to be areas with higher density non-residential buildings, no matter the type of development that occurs in the area. With the residential buildings, there are areas where the type of development that takes place is less dense. For example, in metro regions that consist mostly of suburban areas, the density would be lower as this is the type of development associated with suburban areas. Metro regions that consist more of informal housing areas will have denser residential buildings as this is the type of development associated with development that is more informal.

4.3.2 Building floor space



Requirements for base dataset:

1. Location and use of the buildings + 2. Size (i.e. area) of the buildings + 3. Economic sector that the employment is linked to = 4. Employment capacity of the buildings

Figure 24: Second requirement of base dataset

As illustrated in Figure 24, the next step in the data preparation was to determine the size of the buildings. The main dataset that was used to determine the size of each of the buildings was data related to building footprints. These building footprints consisted of the area of the building and the approximate number of floors that a building has. Figures 25 and 26 show a sample of the building footprints.

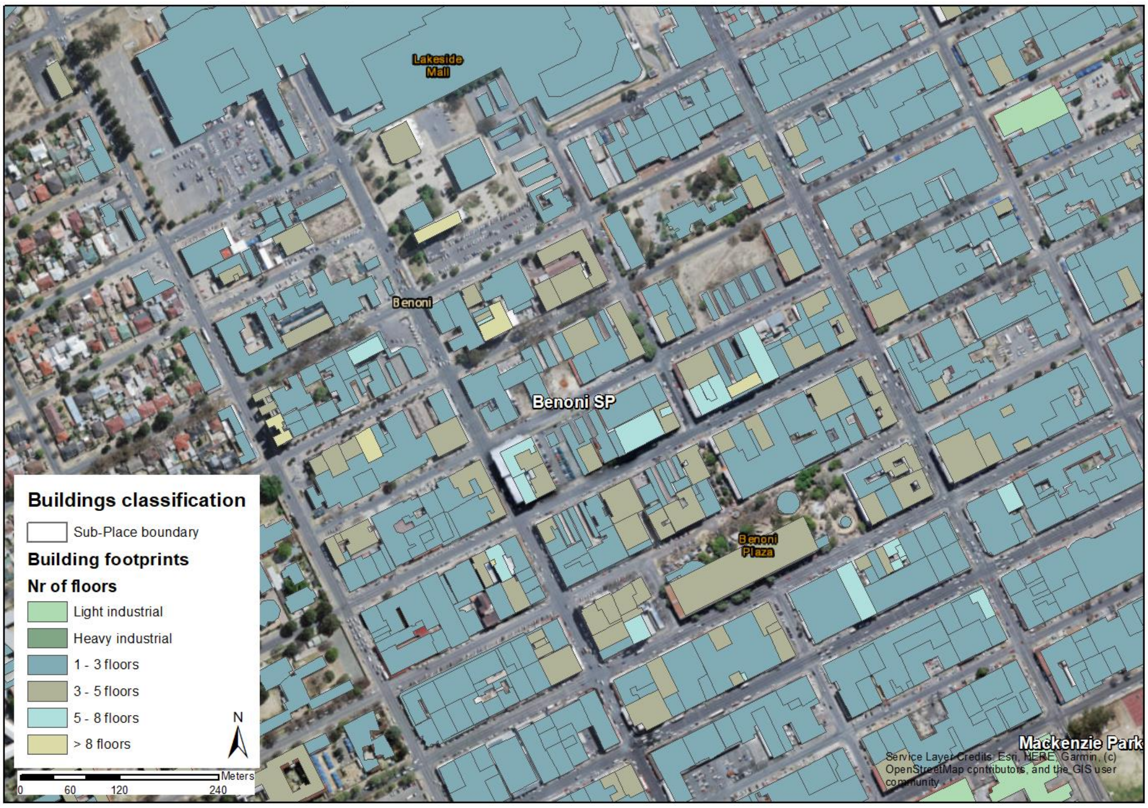


Figure 25: Building footprints

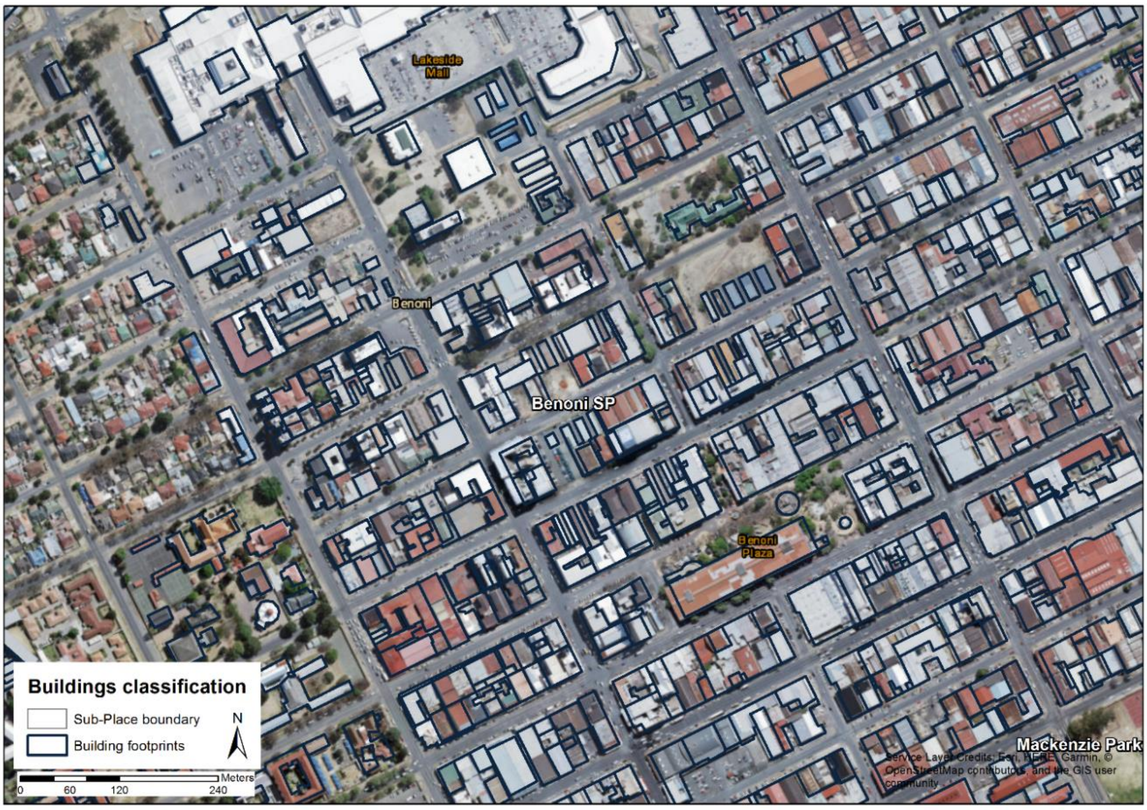


Figure 26: Building footprints outline

The building footprints were not available for residential buildings and there were also not building footprints for all the buildings in the dataset. Therefore a few processing steps were performed to ensure that all buildings had attributes relating to floor area and number of floors. The first step was to join the footprint data to those buildings that intersected with the building footprints. There were situations where more than one building intersected with the footprint. Figure 27 provides an example of multiple intersecting buildings. To solve this problem, the number of buildings that intersected with a footprint was counted and the area of that footprint was then divided by the count. The divided area was then linked to each of the buildings that intersected that specific footprint.

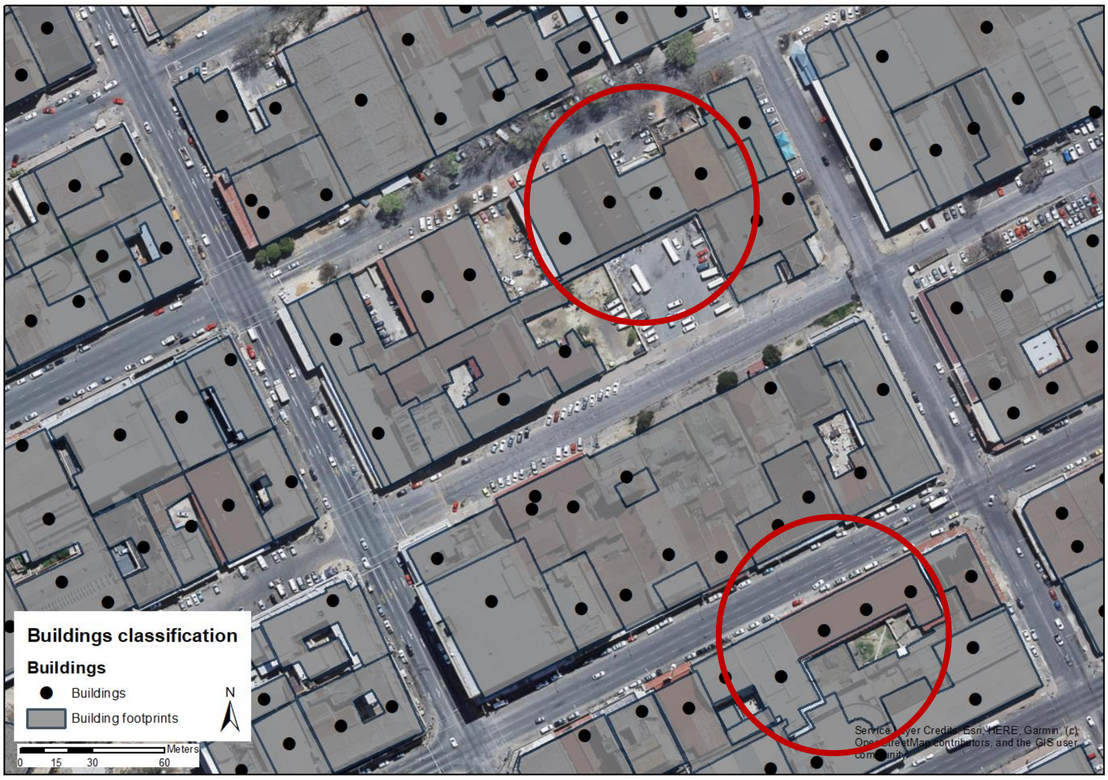

Buildings classification
Buildings
● Buildings
▨ Building footprints
N
Meters
0 15 30 60
Service Layer Credits: Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

Figure 27: Example of more than one building intersecting footprint

Subsequently, the average building footprint size was calculated for each sub-place. This information was then joined to the buildings that did not intersect with any building footprints. Therefore, these buildings were assigned the average size of the buildings that fall within its sub-place. There were still a few buildings that did not have a building size linked to them. This happened when sub-places did not contain any data related to building footprints, therefore an average building area could not be calculated. The average building footprint size for the entire CoE was subsequently calculated and joined to the remaining buildings. After the join, each building now had the following attributes:

1. Primary land use class,
2. Secondary land use class,
3. The floor area of the building, and
4. The number of floors.

The next step was to calculate the floor space. The floor space of a building is calculated by multiplying the area of the building with the number of floors of the building. The floors that were provided in the dataset were provided as a range (ex. 1-3 floors) and not the exact number of floors. Therefore, the maximum value in the range of number of floors was used to calculate the size of the building. The reason for using the maximum rather than the minimum or average, was because the aim was to use this data to calculate the maximum capacity that a building has. The six different floor classes and the final number of floors that were assigned in each class are shown in Table 5.

Table 5: Different floor ranges

Number of floors	Maximum number of floors used
1-3	3 floors
3-5	5 floors
5-8	8 floors
>8	12 floors
Light industrial	1 floor
Heavy industrial	1 floor

The maximum number of floors for more than 8 floors was chosen as 12 floors. This number was determined by using Google Maps street view to manually look at the buildings within the study area that had more than 8 floors linked to them and determining what the approximate maximum number of floors were within this group of buildings.

4.3.3 Economic sectors

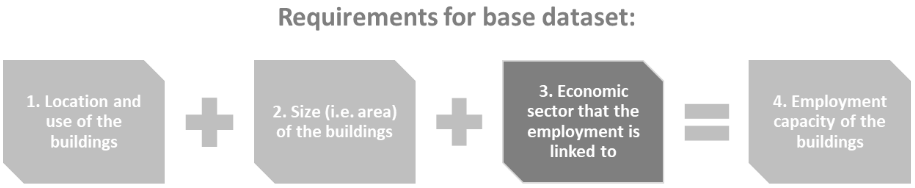

Requirements for base dataset:
1. Location and use of the buildings
2. Size (i.e. area) of the buildings
3. Economic sector that the employment is linked to
4. Employment capacity of the buildings

Figure 28: The third requirement for the base dataset

The third requirement for the base dataset is illustrated in the darker shade in Figure 28 and was the economic sector that the employment opportunity would belong to. In South Africa, there are eleven main economic sectors that most economic activities belong to. Two extra sectors that cover the economic activities that do not fit into the nine main sectors. Table 6 shows the various economic sectors.

Table 6: Economic sectors

SIC code	Economic sector
1	Agriculture, hunting, forestry, and fishing
2	Mining and quarrying
3	Manufacturing
4	Electricity, gas, and water supply
5	Construction
6	Wholesale and retail trade, repairs, hotels, and restaurant
7	Transport, storage, and communication
8	Financial, insurance, real estate, and business services
9	General government
10	Community, social, and personal services
0	Private households

Since all economic activities are grouped into these sectors, all types of employment in South Africa also relate to one of the eleven sectors. To determine which sector an employment opportunity is related to, each of the building classes was linked to a relevant economic sector. Therefore, all the employment opportunities that are located within a building were then linked to that same economic sector.

The Standard Industrial Classification (SIC) of all Economic Activities was used to determine which building classes could be linked to which economic sector (Statistics South Africa, 2012). A detailed description of the different sub-sectors in each of the main economic sectors and the economic activities that make up these sub-sectors are provided in this document and these descriptions were used to determine which buildings to link to which sector. Table 7 shows the economic sectors that were linked to the various building classes.

Table 7: Economic sectors linked to building classes

Primary building class	SIC code
Mining	2
Transport	7
Utilities and infrastructure	4, 7, 8, and 10
Health care facilities	10
Education	10
Commercial	6 and 8
Industrial	3
Recreation and leisure	10
Tourism	6
Institutions	8, 9, and 10
Residential	1, 6, 8, and 10

With many of the building classes, it was clear how to link a building class to an economic sector. For example, *MINING* buildings should be linked to the MINING AND QUARRYING ECONOMIC SECTOR. In these situations, the primary building class was used to link the buildings. For some of the classes, it was not a straightforward match, as one building class could fall into multiple economic sectors. An example of this is the *INSTITUTIONS* buildings class, which could be linked to the COMMUNITY, SOCIAL, AND PERSONAL SERVICE ECONOMIC SECTOR, but these buildings could also be linked to the FINANCIAL, INSURANCE, REAL ESTATE, AND BUSINESS SERVICES ECONOMIC SECTOR or the GENERAL GOVERNMENT ECONOMIC SECTOR. In these situations where the buildings could be grouped into multiple sectors, the secondary building class was used to link the buildings as they provided clearer classes. For example, the *GOVERNMENT* buildings could be linked to the GENERAL GOVERNMENT ECONOMIC SECTOR and the *POLICE AND EMERGENCY SERVICES* buildings could be linked to the COMMUNITY, SOCIAL, AND PERSONAL SERVICE ECONOMIC SECTOR.

In Table 7 it can be seen that the AGRICULTURE, HUNTING, FORESTRY, FISHERIES ECONOMIC SECTOR was linked to the *RESIDENTIAL* building class. Within the *RESIDENTIAL* building class, there is a secondary class for small holdings or agricultural housing. It was decided to link the agriculture employment opportunities to these *RESIDENTIAL* buildings, rather than the buildings classified under *AGRICULTURE* in the buildings data. In the buildings data, some buildings are classified under *AGRICULTURE*, but these buildings only include major outbuildings, which are barns, sheds and other agricultural buildings on farms and smallholdings. Since it did not make sense to link employment opportunities to outbuildings (these are mostly used for storage), the only other option was to link the employment opportunities to the *RESIDENTIAL* buildings.

The only economic sector that was not linked to any of the buildings and that is not covered by this study is the CONSTRUCTION ECONOMIC SECTOR. This economic sector contributes 5 percent of the total employment in the CoE. The CONSTRUCTION ECONOMIC SECTOR was left out as this is a sector in which there are often temporary employment opportunities and not permanent employment positions. Because of this unsteadiness in employment opportunities, it was decided to exclude this sector in this study. Another factor that contributed to the decision to exclude this sector is that the employment opportunities in this sector are not always linked to a specific building but to the site where the construction is taking place.

4.3.4 Employment capacity

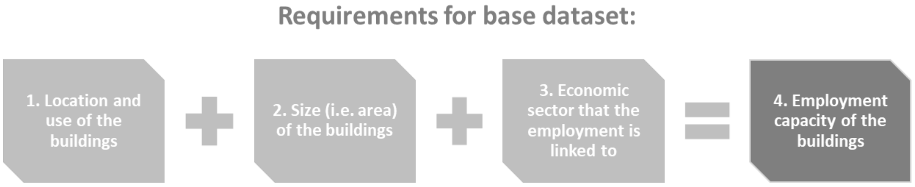

Requirements for base dataset:
1. Location and use of the buildings
2. Size (i.e. area) of the buildings
3. Economic sector that the employment is linked to
4. Employment capacity of the buildings

Figure 29: Final requirement of base dataset

From Figure 29 it can be seen that at this point each of the buildings then had a usage, a size, and what economic sector it is related to, thus the final step was to use this information to calculate the employment capacity for each building. Various methods were used to calculate the capacity of each building. The eleven primary building classes were each individually inspected to determine the best method to calculate an approximate employment capacity for each building. As previously shown in Table 4, these classes are:

1. Mining,
2. Transport,
3. Utilities and infrastructure,
4. Health care facilities,
5. Education,
6. Commercial,
7. Industrial,
8. Recreation and leisure,
9. Tourism,
10. Institutions, and
11. Residential.

In some situations, the same method was used to calculate the capacity for all the buildings in a primary building class. In other situations, different methods were used for the different secondary classes within a primary building class. For many of the classes, the employment capacity was determined using the secondary classes rather than the primary classes, as the main class was too general to define one method of calculating an employment capacity for all the buildings that are grouped within a class. For example, the *MINING* primary class consisted of only one secondary class, thus the primary class could be used when calculating the capacity. The *COMMERCIAL* primary class consisted of six secondary classes, each varying largely in the number of employment opportunities that could be available. Therefore, the secondary class was used to calculate the capacity of each building.

Mining

The employment capacity for buildings classified as mining was calculated using the total number of people working in the MINING AND QUARRYING ECONOMIC SECTOR (As defined by Statistics South Africa (2012)) in CoE for 2011. The total employment within the mining sector was proportionally divided, using building floor space, between all the buildings that had a primary building classification of *MINING*. Therefore, larger buildings would have a larger employment capacity than those buildings with a smaller size.

Transport

The employment capacity for those buildings that were classified as *TRANSPORT* was calculated using the total number of employment opportunities in the TRANSPORT, STORAGE, AND COMMUNICATION ECONOMIC SECTOR (As defined by Statistics South Africa (2012)) in CoE for 2011. The number of employment opportunities in the sector was proportionally divided using the building floor space for all the buildings classified as *TRANSPORT* buildings.

Utilities and infrastructure

The employment capacity for buildings classified as *UTILITIES AND INFRASTRUCTURE* were calculated using various methods, as the various secondary classes within the *UTILITIES AND INFRASTRUCTURE* primary class belong to different economic sectors. The three classes that are linked to the ELECTRICITY, GAS, AND WATER SUPPLY ECONOMIC SECTOR (As defined by Statistics South Africa (2012)) includes the water storage and sewerage treatment plants, the energy production and distribution buildings, and access control buildings. Therefore, the total employment in this economic sector for 2011 was proportionally divided between these three classes, based on the building floor space.

The refuse disposal buildings form part of the COMMUNITY, SOCIAL AND PERSONAL SERVICES ECONOMIC SECTOR (As defined by Statistics South Africa (2012)). Therefore, the total number of employment opportunities in this sector was proportionally divided between the refuse buildings and all the other buildings that form part of this economic sector. Once again, the building floor space was used for the proportional division. The posts and telecommunications buildings form part of the TRANSPORT, STORAGE, AND COMMUNICATION ECONOMIC SECTOR (As defined by Statistics South Africa (2012)) and therefore the total employment of this sector was proportionally divided between the post and telecommunications buildings, as well as all the other buildings that form part of this economic sector.

Health care facilities

The primary building class was also used to calculate the employment capacity for the *HEALTH CARE FACILITIES*. As with the mining and transport classes, the total number of people employed in the HEALTH AND SOCIAL WORK ECONOMIC SECTOR (As defined by Statistics South Africa (2012)) for the CoE was used. This sector includes human health activities such as hospital activities, medical and dental practice activities, and other human health activities. This sector also includes veterinary activities and social work activities (Statistics South Africa, 2012). The employment total was proportionally divided between all the buildings that had *HEALTH CARE FACILITIES* as the primary building class. As with the other classes, the size of the buildings was used for the proportional division.

Education

Two different methods were used for the *EDUCATION* buildings, therefore the secondary building classes were used to calculate the employment capacity. The employment capacity for the buildings that were substituted in earlier in the data preparation, using the GDE schools, were calculated differently to those classes that were not substituted. The location of each school representing a building, the number of classrooms and the number of learners were used to determine the employment capacity for the schools that were replaced.

The Education district profile (2015) determined that the ratio of the number of educators to a classroom is 1.2 and the education statistics for Gauteng (2013) determined that for every 246 learners at a school, there is 1 administrative staff member (Department of Basic Education, 2015; Department of Basic Education, 2013). Using this information the employment capacity was calculated by multiplying the number of classrooms with a ratio of 1.2 and dividing the number of learners by 246. These two values were then summed to determine the final employment capacity for each school.

The employment capacity for tertiary education institutions and other educational institutions was determined using the size of the buildings and the employment totals by occupation. The three occupations that were considered included teaching professionals for college, university and higher education institutions, other teaching institution professionals, and other education professionals that were not classified elsewhere (As defined by Statistics South Africa (2012)). The total employment for these three occupations was proportionally divided between the two secondary classes by making use of the size of the buildings.

Commercial

An estimate of area per employee for each of the secondary classes within the *COMMERCIAL* class, along with the building floor space was used to determine the employment capacity. The area per employee was calculated by using development plans for new developments within the CoE. These plans outline a proposed land use (e.g. office, retails, industrial, mixed-use), bulk floor area of the development, and the number of jobs that the developed area will be able to accommodate (DEMACON Market Studies, 2015). This information was then used to determine the estimated area per employee for each of the sub-classes. Table 8 shows the estimated area per employee that was used for each of the secondary classes.

Table 8: Estimated area per employee for each secondary class of the commercial primary class

Secondary building class	Area per employee
Shopping mall	29 m^2
Shopping centre	36 m^2
Commerce	25 m^2
Petrol station, service station	29 m^2
Office parks	22 m^2

The amount of floor space was divided by the estimated area per employee to calculate the final employment capacity for each of the secondary building classes. The only secondary building class that was handled differently was the informal trading class. The employment capacity for this class was determined by using the total informal employment for the CoE in the WHOLESALE AND RETAIL TRADE, CATERING AND ACCOMMODATION ECONOMIC SECTOR (As defined by Statistics South Africa

(2012)). The total employment was proportionally divided between all the buildings classified as informal trading by making use of the size of the buildings.

Industrial

The capacity for the *INDUSTRIAL* class was calculated using the same method as the *COMMERCIAL* class. The capacity was also calculated using the secondary building classes and the estimated area per employee that was determined using the development plans for new developments within the CoE (DEMACON Market Studies, 2015). Table 9 illustrates the estimated area per employee for each of the secondary classes. Once again, the amount of floor space was divided by the estimated area per employee to calculate the final employment capacity for each of the secondary building classes.

Table 9: Estimated area per employee for each secondary class of the industrial primary class

Secondary building class	**Area per employee**
Light industries and warehousing/distribution	114 m^2
Heavy industries	171 m^2
Fuel depots	171 m^2

Recreation and leisure

The primary building class for *THE RECREATION AND LEISURE* buildings was used to calculate the employment capacity. The total number of people that are employed in the COMMUNITY, SOCIAL AND PERSONAL SERVICES ECONOMIC SECTOR (As defined by Statistics South Africa (2012)) was proportionally divided between the *RECREATION AND LEISURE* buildings, as well as all the other buildings that fall within this economic sector. As with the other classes, the sizes of the buildings were used for the division.

Tourism

The employment capacity for the *TOURISM* class was also calculated using the primary building class. The total number of people employed in the CATERING AND ACCOMMODATION SERVICES ECONOMIC SECTOR, which is a sub-sector of the WHOLESALE AND RETAIL TRADE, REPAIRS, HOTELS, AND RESTAURANTS ECONOMIC SECTOR (As defined by Statistics South Africa (2012)), was proportionally divided between all the buildings that fall in the *TOURISM* building class. As with the other classes, the sizes of the buildings were used for the division.

Institutions

Various methods were used to calculate the employment capacity of the *INSTITUTIONS* class. The method used differed based on the secondary building class. The capacity for the police building class was calculated by using the size of the population that the stations serve. The employment capacity was calculated using an estimated ratio of the number of police officials that are required for a certain population size. In Gauteng, the standard for this ratio is: 1 police official is required for every 293 individuals that live in the specific police district (Le Grange, 2013). Therefore the population that each police station serves, was divided by 293 to determine the employment capacity for each point.

The employment capacity for the government and foreign government secondary classes was determined using the total employment in THE GENERAL GOVERNMENT ECONOMIC SECTOR (As defined by Statistics South Africa (2012)) for the CoE. The total employment was proportionally divided between all the buildings classified as government or foreign government by making use of the size of the buildings.

The employment capacity for the rest of the secondary classes within the *institutions* primary class was calculated using the total employment in the Community, Social, and Personal Services economic sector (excluding the education and health sub-sectors that are within Community, Social, and Personal Services economic sector) (As defined by Statistics South Africa (2012)). The size of the buildings was used to proportionally divide the total employment between:

1. Research institutions,
2. Defense,
3. Emergency services,
4. Correctional services, and
5. All the other classes that fall in the Community, Social, and Personal Services economic sector.

Residential

The *residential* buildings dataset already had a pre-calculated employment capacity for each building. Therefore, the employment capacity for the *residential* classes was not calculated as part of the data preparation, but the existing capacity was rather used. The *residential* buildings that are available for employment allocation include:

1. Formal houses,
2. Informal structures,
3. Cluster or complexes,
4. Small holdings or agriculture, and
5. Security estates.

These classes all form part of three economic sectors. The formal houses, informal structures, complexes, and security estates form part of the Private Households economic sector (As defined by Statistics South Africa (2012)). Private households include all domestic workers. The private households also include home-based employment in the Wholesale and Retail trade, Repairs, Hotels, and Restaurant economic sector and the Financial, Insurance, Real Estate, and Business Services economic sector. The small holdings and farmstead housing classes form part of the Agriculture, Hunting, Forestry, and Fishing economic sector (As defined by Statistics South Africa (2012)).

After the employment capacity was determined for each building class, all four of the requirements for the base dataset was satisfied. The next step in the dissertation was to use this information to develop the employment allocation algorithm.

Chapter 5: Development of algorithm

5.1 Introduction

The chapter discusses the different types of algorithms that are available and then goes into more detail about how the employment allocation algorithm was developed and validated, as well as a discussion on the results.

5.2 Employment allocation algorithm development

The Distributed Evolutionary Algorithm in Python (DEAP) (https://deap.readthedocs.io/en/master/) was used as the framework to develop the algorithm. DEAP is an open-source evolutionary algorithm framework that is developed in Python and enables users to rapidly create prototypes by allowing them to customise the algorithm to suit their needs. DEAP provides two main structures to its users to develop their algorithm, namely a [CREATOR] module and a [TOOLBOX]. The [CREATOR] module allows the user to create individuals and populations from any data structure, including lists and dictionaries and is a vital part of enabling the implementation of the algorithm. The [CREATOR] is also used to initialise the fitness functions (Fortin, et al., 2012; DEAP, 2018).

The [TOOLBOX] is a container for the various tools used in the algorithm and the user can specify which tools are used. This allows a user to add only the tools that are relevant to the problem, it also provides the capability to easily make changes to the tools if required. DEAP allows for multi-objective evaluation and it allows the use of multiprocessing when it is necessary to process a large amount of data. DEAP provides a basic framework that can be used to develop the algorithm (Fortin, et al., 2012; DEAP, 2018). Figure 30 provides pseudo-code of the basic steps followed with an evolutionary algorithm

```
Begin
  Initialise individual
  Evaluate fitness
  Repeat until (i = Number of generations) Do
    Select
    Crossover
    Mutate
    Evaluate
    Select
  Write out results
End
```

Figure 30: Pseudo-code for evolutionary algorithm

The main reasons that the DEAP framework was chosen as the basis for the employment allocation algorithm were because of the multi-objective functionality, the customisation capabilities, the ease of use, and the built-in toolbox. The pseudo-code translated into four main components:

1. Initialising the population,
2. Evaluating the fitness,
3. Specifying the crossover and mutation rules,
4. Selection of the best offspring.

The sections that follow will discuss how each of the components was adapted and implemented for the employment allocation algorithm. Figure 31 provides the steps followed in the development of the algorithm and in which subsection of the chapter each of these steps is discussed.

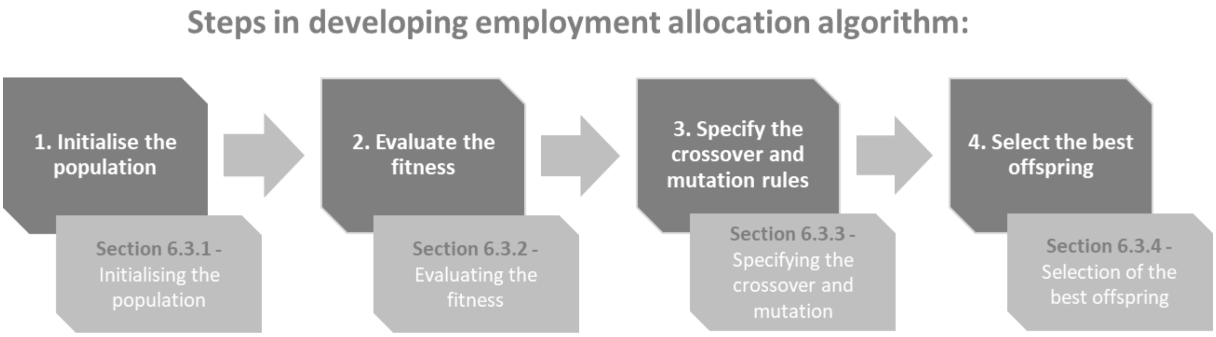

Steps in developing employment allocation algorithm:
1. Initialise the population
Section 6.3.1 - Initialising the population
2. Evaluate the fitness
Section 6.3.2 - Evaluating the fitness
3. Specify the crossover and mutation rules
Section 6.3.3 - Specifying the crossover and mutation
4. Select the best offspring
Section 6.3.4 - Selection of the best offspring

Figure 31: Steps followed in developing algorithm

5.2.1 Initialising the population

Steps in developing employment allocation algorithm:
1. Initialise the population
2. Evaluate the fitness
3. Specify the crossover and mutation rules
4. Select the best offspring

Figure 32: First step in development of algorithm

As illustrated in Figure 32 using the darker shade, the first step of developing the algorithm was to initialise the population. The initial population is made up of a number of individuals. As mentioned in Section 2.6, each individual represents a tentative solution. To create the initial population, the type of individuals that would form the population first needs to be determined. DEAP has various types or formats that could be used to create the individuals. Some of these formats include lists, permutations, tree of mathematical expressions, and particles (DEAP, 2018). Figure 33 provides an example of each of the formats. The format used to create the individual would depend on the type and complexity of the problem that is being solved. In DEAP lists are usually used as the basis for all the other formats, just with added complexity. For example, particles are created by inheriting from the basic LIST type but have the added attributes of speed, speed limits, and best position.

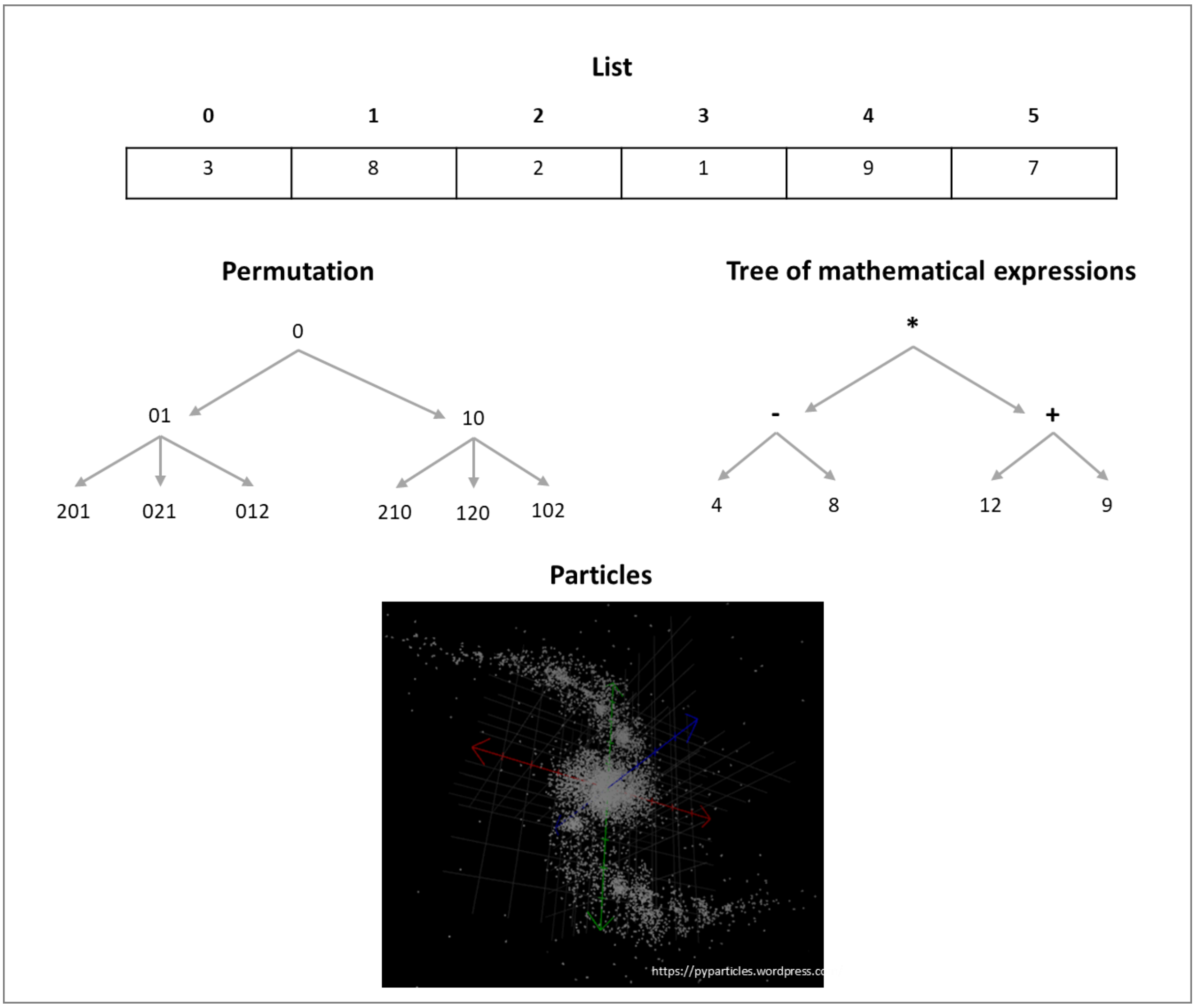
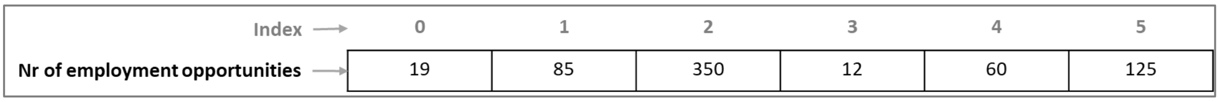


Figure 33: Examples of the different formats available in DEAP

A list of integers was identified to be the most appropriate format to use for solving the problem of this project. When using this format, the index of the list will represent a building ID. The integer value at each point in the list will be the number of allocated employment opportunities. The index of the list can then be linked back to the building ID of the base data to link the allocated employment opportunities to a specific building. Figure 34 provides an extract of the base dataset created in the previous chapter. Figure 35 provides an example of the list format used to allocate the employment opportunities. Figure 36 illustrated how the number of employment opportunities in the list is linked back to the original dataset.

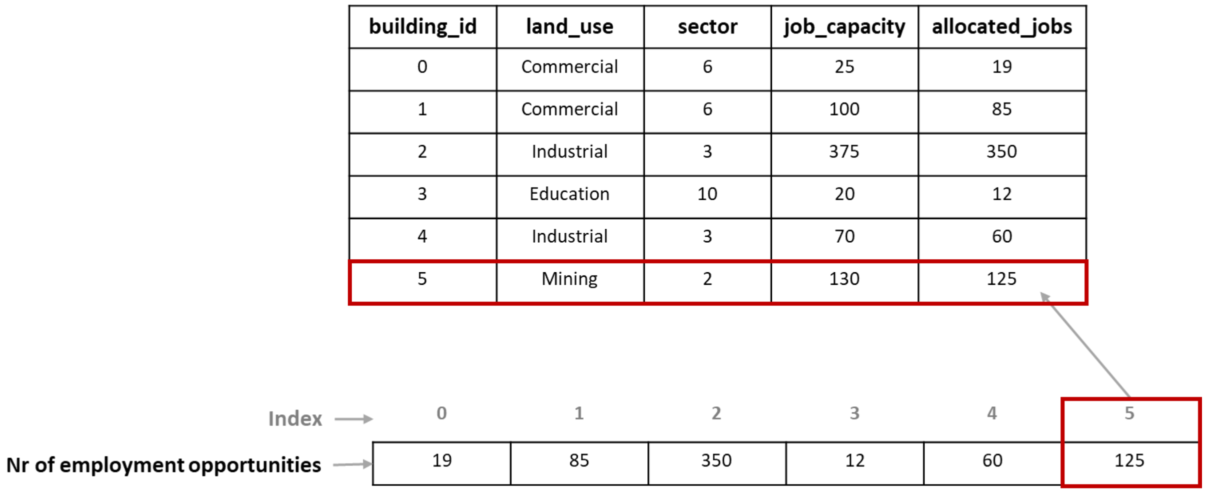
building_id	land_use	sector	job_capacity
0	Commercial	6	25
1	Commercial	6	100
2	Industrial	3	375
3	Education	10	20
4	Industrial	3	70
5	Mining	2	130

Figure 34: Extract from base dataset



Index	0	1	2	3	4	5
Nr of employment opportunities	19	85	350	12	60	125

Figure 35: Example of list format



building_id	land_use	sector	job_capacity	allocated_jobs
0	Commercial	6	25	19
1	Commercial	6	100	85
2	Industrial	3	375	350
3	Education	10	20	12
4	Industrial	3	70	60
5	Mining	2	130	125

Index	0	1	2	3	4	5
Nr of employment opportunities	19	85	350	12	60	125

Figure 36: Example of linking list back to original data

Once the type of individual has been established, the second step is deciding how each individual within the population will be generated. To create the individuals that form part of the initial population, each building was allocated a certain proportion of its employment capacity. Initially, the proportion that was allocated was calculated by randomly generating a number between 0 and 1, which represented a portion of a whole. A random number was used because as mentioned in Section 2.6, the initial population is usually created by making use of random value. This value would then be multiplied with the capacity of each building. The same proportion value was used for all the buildings in an individual, but a new proportion would randomly be generated each time a new individual was created.

Through the development of the algorithm, the process that was used to generate the initial population was updated to allow for more variation between the individuals, which would return a larger selection from which the final solution could be developed. This ensured that premature convergence did not occur in the algorithm. The randomly generated proportion that was used to allocate the initial employment opportunities was calculated by just using one random value for all economic sectors. This section of the algorithm was updated to rather generate random values for each of the economic sectors. The specific random value per economic sector was then multiplied with the capacity of the buildings that are located in each of those economic sectors.

This approach was taken because some sectors had a lot more capacity than others. The added capacity for some sectors was a result of how the capacity was calculated for the buildings in each sector when the base dataset was developed, which resulted in some classes having a much higher allocation of employment than others. This, in the end, led to the algorithm over-allocating employment to buildings in some sectors and under allocating employment to buildings in other sectors. With the change to generating proportions per sector, the likelihood of the same sectors always having an over allocation of employment was reduced.

5.2.2 Evaluating the fitness

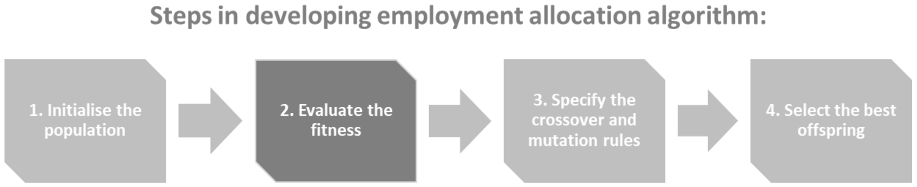

Steps in developing employment allocation algorithm:
1. Initialise the population
2. Evaluate the fitness
3. Specify the crossover and mutation rules
4. Select the best offspring

Figure 37: Second step in development of algorithm

As shown in Figure 37, the second step in developing the algorithm was to evaluate the fitness by creating fitness functions. The first step of creating the fitness functions is to decide what variables could be used to evaluate the fitness of an individual. This decision also influences whether a single objective or multi-objective evaluation function is used. A single objective function only uses one measure of fitness to determine the fitness of an individual. A multi-objective function uses a combination of more than one measure to determine the fitness of a function. For the employment allocation two measures of fitness were identified, therefore a multi-objective fitness function was used.

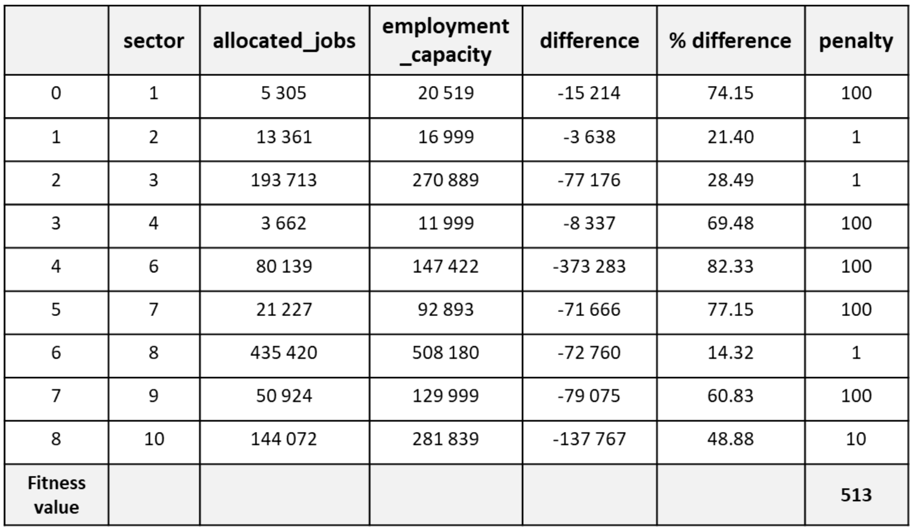
The first fitness function or objective considers the number of employment opportunities allocated per economic sector at municipal level. These values are compared to the actual number of employment opportunities per economic sector for 2011. The absolute difference between the allocated number of employment opportunities and the actual number of employment opportunities was calculated per sector. The total difference was then calculated by summing the differences. This value was then used for the first objective. Figure 38 provides an example of the calculation of the fitness value of the first objective for one individual.

	sector	allocated_jobs	actual_jobs	difference
0	1	5 305	11 733	6 427
1	2	13 361	11 692	1 669
2	3	193 713	165 376	28 337
3	4	3 662	5 662	1 999
4	6	80 139	211 024	130 884
5	7	21 227	75 360	54 132
6	8	435 420	205 343	230 077
7	9	50 924	123 930	73 005
8	10	144 072	190 856	46 783
Fitness value				**573 319**

Figure 38: Example of the first objective calculation

The second fitness function or objective considers the difference between the number of allocated employment opportunities per economic sector and the employment capacity of each economic sector based on summarised building data. This fitness function is used to ensure that the algorithm does not allocate more employment opportunities than a building has the capacity for. It also ensures that it does not allocate more employment opportunities than all the buildings in each of the economic sectors have the capacity for. The total number of allocated employment opportunities per economic sector and the total employment capacity per economic sector was summarised at municipal level by using the attributes in the building data.

The difference between the number of allocated employment opportunities and the employment capacity per economic sector was then calculated. Since the difference values could differ a lot based on the different capacity sizes for each economic sector, the percentage difference was calculated. The percentage difference was then used to assign a “penalty” score. Negative percentages would receive the largest penalty, as there are more employment opportunities allocated than there is capacity for. Values between 0 and 60 would receive a lower score, as this is a tolerable difference. Values above 60 would again receive a larger penalty, as this means that the allocated employment opportunities are a lot less than the capacity. The score for each economic sector is finally summarised to provide the final value for the second objective. Figure 39 provides an example of the calculation of the fitness value of the second objective for one individual.



	sector	allocated_jobs	employment _capacity	difference	% difference	penalty
0	1	5 305	20 519	-15 214	74.15	100
1	2	13 361	16 999	-3 638	21.40	1
2	3	193 713	270 889	-77 176	28.49	1
3	4	3 662	11 999	-8 337	69.48	100
4	6	80 139	147 422	-373 283	82.33	100
5	7	21 227	92 893	-71 666	77.15	100
6	8	435 420	508 180	-72 760	14.32	1
7	9	50 924	129 999	-79 075	60.83	100
8	10	144 072	281 839	-137 767	48.88	10
Fitness value						**513**

Figure 39: Example of the second objective calculation

Once a decision has been made on which variables will be used, the type of fitness function needs to be established. DEAP allows for either minimising or maximising a function. It also allows the user to assign weights to each of the fitness values. This allows a user to assign a higher weight to a fitness value that is more important (DEAP, 2018). The sign of the weight that is assigned determines whether a function is a minimising or maximising function. A negative weight is indicative of a minimising function and a positive weight indicates a maximising function.

For the algorithm, both objective functions were minimised. The first objective received a weight of –10 and the second objective received a weight of -1. The first objective function received a higher weight because it was important to ensure that the algorithm did not allocate more employment opportunities than there was in 2011.

5.2.3 Specifying the crossover and mutation

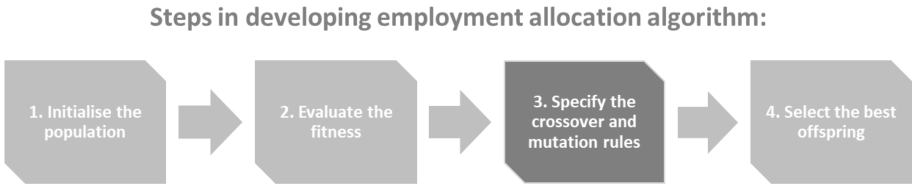

Steps in developing employment allocation algorithm:
1. Initialise the population
2. Evaluate the fitness
3. Specify the crossover and mutation rules
4. Select the best offspring

Figure 40: Third step in developing algorithm

Figure 40 illustrates the third step in developing the algorithm in a darker shade. As mentioned in Section 2.6, crossover and mutation need to be implemented to allow changes to be made to the individuals to create better offspring. DEAP has various built-in crossover and mutation operators. Throughout the process of developing the algorithm, many of the built-in crossover and mutation operators were tested. Some of the built-in mutation operators that were tested included the shuffle index mutation and the Gaussian mutation. The shuffle index mutation moves the values of the individual around within the individual. Figure 41 gives an example of how the shuffle index mutation works. The Gaussian mutation randomly selects a position within the individual and adds a Gaussian random value to the value at that position. Figure 42 gives an example of how Gaussian mutation works. In the end, the shuffle index provided the best solution as was used as the final mutation operator.

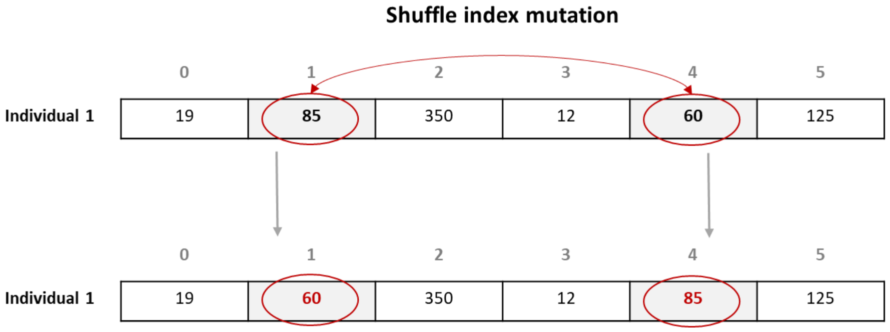


Figure 41: Example of shuffle index mutation

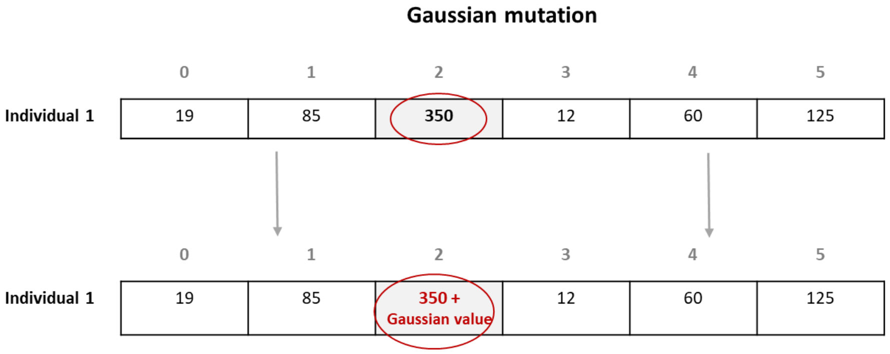


Figure 42: Example of Gaussian mutation

Initially, only mutation was used in the algorithm, but this caused a problem that not enough changes were made to the individuals. Therefore, there was not enough variation between the individuals, which ultimately led to the solution space not being explored adequately. Using only mutation led to the algorithm finding a final solution very quickly, but the solution was not very accurate. The total number of employment opportunities that were allocated was only off by a few 100 employment opportunities compared to the actual number of employment opportunities. Therefore, the algorithm under or allocated around 0.01% of the total number of employment opportunities. However, when the total employment opportunities per economic sector were inspected some sectors had a lot more allocated employment opportunities than the actual employment opportunities that were available. Other sectors had a lot less allocated than the total that was available. In one of the runs, the algorithm under allocated around 70% on the total number of employment opportunities in the COMMUNITY, SOCIAL, AND PERSONAL SERVICES SECTOR. On the other hand, it over allocated around 60% on the total number of employment opportunities in THE WHOLESALE AND RETAIL TRADE, REPAIRS, HOTELS, AND RESTAURANTS SECTOR.

In Section 2.6 it was identified that crossover is necessary to ensure that the offspring is not similar to the parent. Therefore implementing crossover would solve this problem of not having enough variation in the solution space. Three of the relevant built-in crossover operators were tested. These operators included a one-point crossover and a two-point crossover. When a one-point crossover is used, one portion of an individual is swapped with the portion that is at the same position of a different individual. Figure 43 provides an example of one-point crossover.

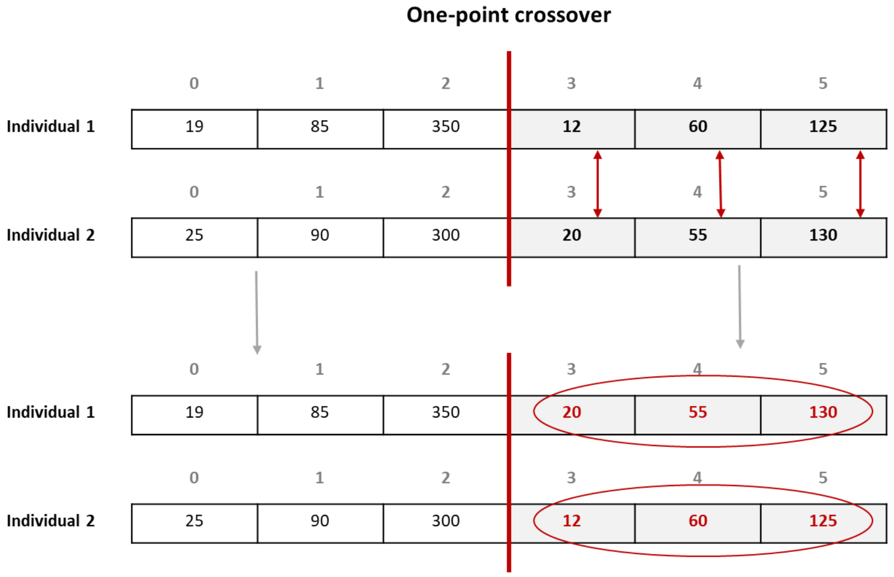


Figure 43: Example of one-point crossover

Two-point crossover works the same, but instead of just one portion of the individual being swapped, two portions of the individuals are swapped. Figure 44 provides an example of two-point crossover. In the end, the two-point crossover led to a better solution and was used as the crossover operation.

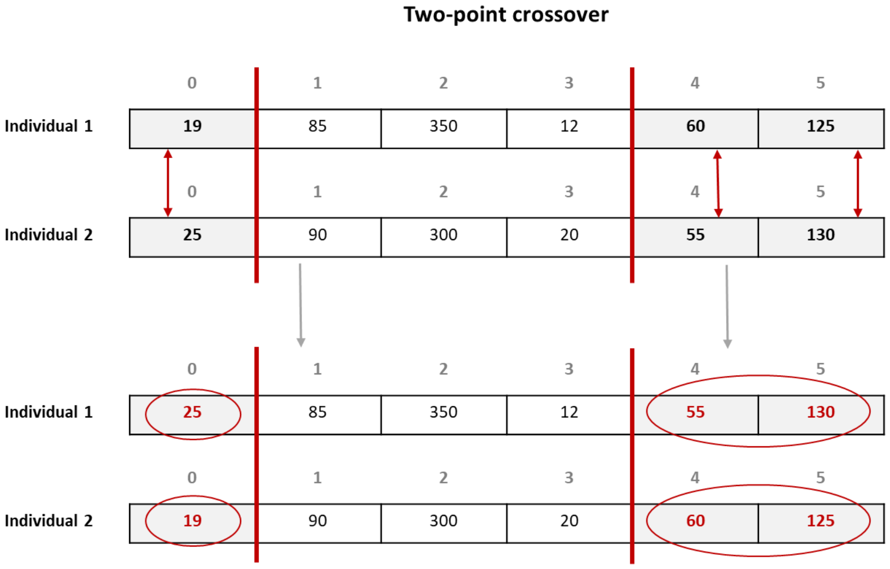


Figure 44: Example of two-point crossover

5.2.4 Selection of the best offspring

Steps in developing employment allocation algorithm:
1. Initialise the population
2. Evaluate the fitness
3. Specify the crossover and mutation rules
4. Select the best offspring

Figure 45: Fourth step in developing algorithm

The final step in developing the algorithm is showing in a darker shade in Figure 45 and this step was deciding on the method that would be used to select the best offspring of which the next generation would consist. DEAP also provided various built-in selection operators. Once again, various operators were tested, some of these include best selection and lexicase selection. The best selection operator selects a certain number of best individuals basing this on a specified value. The lexicase selection operator selects the best individuals based on the fitness cases. In the end, the best selection operator provided the best results.

5.3 Validation of employment allocation

Various factors were considered to validate the accuracy and performance of the employment allocation algorithm. Throughout the process of developing the algorithm, the performance of the algorithm was also visualised using various graphs to determine whether the algorithm was performing as expected or not. The first measure and graph that was used were plotting the minimum and maximum values for both Objective 1 and Objective 2 to see if the algorithm results converge to a final solution. As mentioned in Section 5.2, Objective 1 considers the difference between the actual number of employment opportunities and the allocated employment opportunities. Objective 2 considers the difference between the number of allocated employment opportunities and the employment capacity. Figure 46 and 47 show the change in the minimum and maximum values of the two objectives of the fitness function from the two different runs mentioned in Section 5.2 that were generated during the algorithm development.

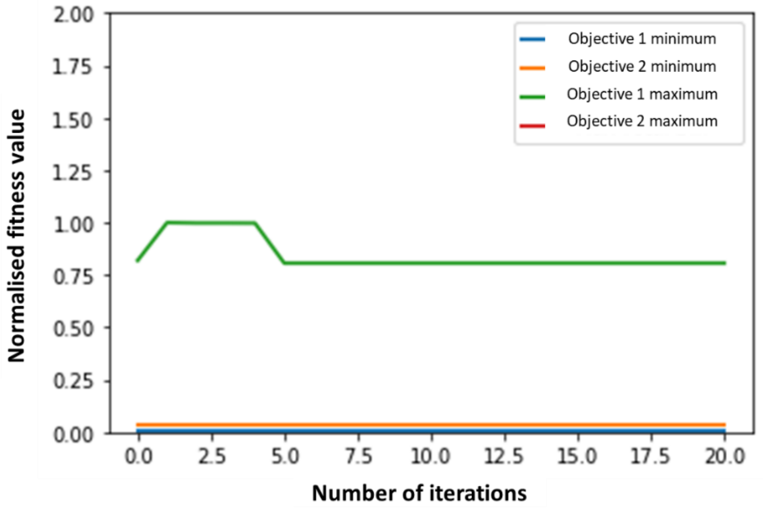


Figure 46: Change in minimum and maximum values for the initial run

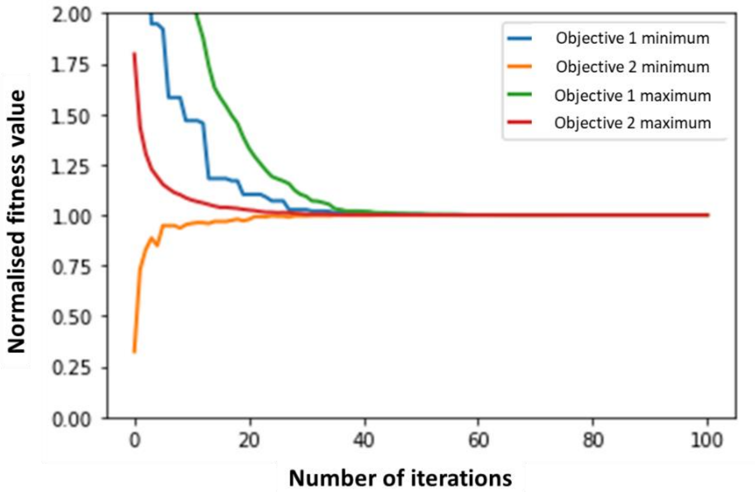


Figure 47: Change in minimum and maximum values for improved run

Figure 46 illustrates the initial algorithm that only implemented mutation and no crossover. This run also used only one randomly generated number to create the initial population. The algorithm had very little variation in the solutions that were found. The behaviour of the initial run was not in line with what is expected behaviour for a typical evolutionary algorithm, which is that when the algorithm initially starts its run there is a lot of variation within the initial individuals as this allows for a large range of individuals to be evaluated to find the best solution. As the algorithm run progresses, the variation should become less as the options for better solutions become less. This means the algorithm should start converging to a final solution in which the marginal improvement is not big enough to warrant continuing the run of the algorithm and therefore the best solution is found.

Figure 47 shows the improved run when crossover was added to the algorithm. This run also took into consideration the change from using one random value when creating the initial population, to using a random value per economic sector. This graph shows immense improvement in the variation within the first few iterations of the algorithm and then as the number of iterations increases, the algorithm starts to converge towards a final solution. The behaviour of the improved run is more in line with the expected behaviour of a typical evolutionary algorithm.

Although the improved run showed better behaviour in terms of performance, it was still showing some error in generating an accurate allocation of employment opportunities. There were two measures of the accuracy of the algorithm. The first measure was the difference between the total number of employment opportunities that were allocated on the municipal level and the actual number of employment opportunities in 2011 for the CoE. The second measure was the difference between the total number of allocated employment opportunities per economic sector and the actual number of employment opportunities per economic sector in 2011 for the CoE. Table 10 shows the evaluation of these two measures for the improved run.

Table 10: Validation results for improved run

SIC code	Economic sector	Allocated employment	Actual employment	Difference	%
1	Agriculture, forestry, and fisheries	11 244	11 733	-489	-4.17
2	Mining and quarrying	8 323	11 692	-3 369	-28.81
3	Manufacturing	165 774	165 376	398	0.24
4	Electricity, gas, and water	3 494	5 662	-2 168	-38.29
6	Wholesale and retail trade, catering, and accommodation	208 353	211 024	-2 671	-1.27
7	Transport, storage, and communication	77 876	75 360	2 516	3.34
8	Finance, insurance, real estate, and business services	206 936	205 343	1 593	0.78
9	General government	107 229	123 930	-16 701	-13.48
10	Community, social, and personal services	191 281	190 856	425	0.22
	Overall allocation				
	Total employment	**980 510**	**1 198 005**	**-20 466**	**2.04**

Table 10 illustrates that the algorithm had an acceptable overall allocation, but that the allocation per economic sector was not as accurate. Although different crossover, mutation, and selection methods were tested after the improved run, there was not an increase in the accuracy of the algorithm per economic sector. Since the algorithm was not the problem, the only other component that could be causing the error was the base data. When inspecting Table 10, it can be seen that the Mining and Quarrying sector, the Electricity, Gas, and Water sector, and the General Government sector show the largest difference between the allocated employment opportunities and the actual employment opportunities.

When the employment capacity for the buildings in these sectors was inspected, it was identified that these sectors had the least amount of extra capacity. When the base dataset was created, various methods were used to determine the capacity of the buildings. This led to some sectors having a lot more excess capacity than others. Since the algorithm uses the mean difference between the allocated employment capacity and the actual employment capacity when evaluating the fitness, the differences in the amount of excess capacity skewed the fitness function. This ultimately led to some sectors having a less accurate allocation than others.

The total number of employment opportunities per economic sector that was used to calculate the employment capacity for the various sectors were rounded up to the nearest thousand or ten thousand depending on the number of employment opportunities. This allowed for a larger excess of capacity for all the economic sectors. The same crossover, mutation, and selection methods that was used for the improved run was used with the updated base dataset and the final allocation was then created. Figure 48 depicts the minimum and maximum values for the two objectives from the final run.

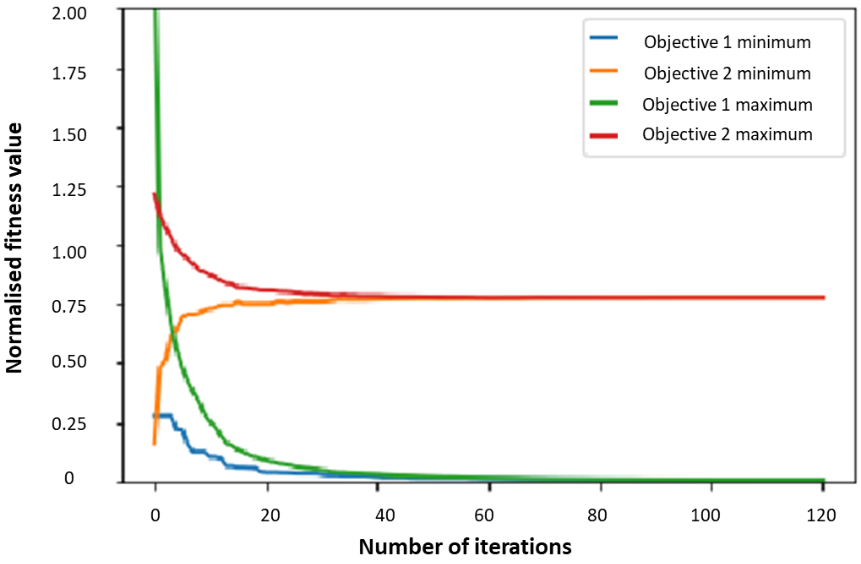


Figure 48: Evaluation criteria graph for the final run

Figure 48 illustrates that the final run algorithm also showed the behaviour of a typical evolutionary algorithm. Initially, there is a lot of variation between the solutions but as the algorithm continues its run, there is a convergence to a final solution. At about the 30th iteration, the amount of improvement of the algorithm started to decrease. The algorithm stopped showing any improvement at about the 110th iteration. At this point, the algorithm could not find any more improved solutions.

Figure 49 and 50 depicts the percentage of improvement of Objective 1 and Objective 2 throughout the algorithms run. It shows a similar trend to that of Figure 48 but from a different perspective. Figure 49 illustrates that through the first number of iterations, Objective 1 showed a large percentage of improvement, but as the number of iterations increased, the percentage of improvement decreased. The graph also shows that no more improvement occurred about the 110th iteration.

Figure 50 illustrates that Objective 2 also showed a lot of improvement initially, although the percentage of improvement was a lot less than that of Objective 1. At about the 15th iteration, Objective 2 had very little improvement until it also stopped improving at about the 110th iteration. The continuous improvement of Objective 1, while Objective 2 shows very little improvement in the later iterations could occur because the total number of employment opportunities is not being increased and the employment opportunities are only being moved between sectors. Throughout the run, there were times where Objective 2 got worse and the percentage improved went into the negative. This most likely occurred when the number of employment opportunities in one or more of the sectors was more than the employment capacity that the buildings in those sectors had.

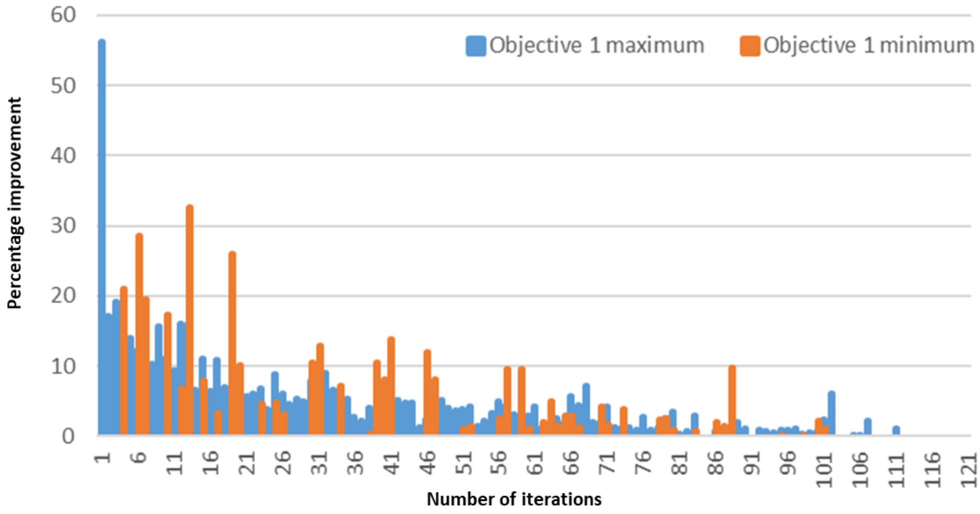


Figure 49: Percentage improved per iteration - Objective 1

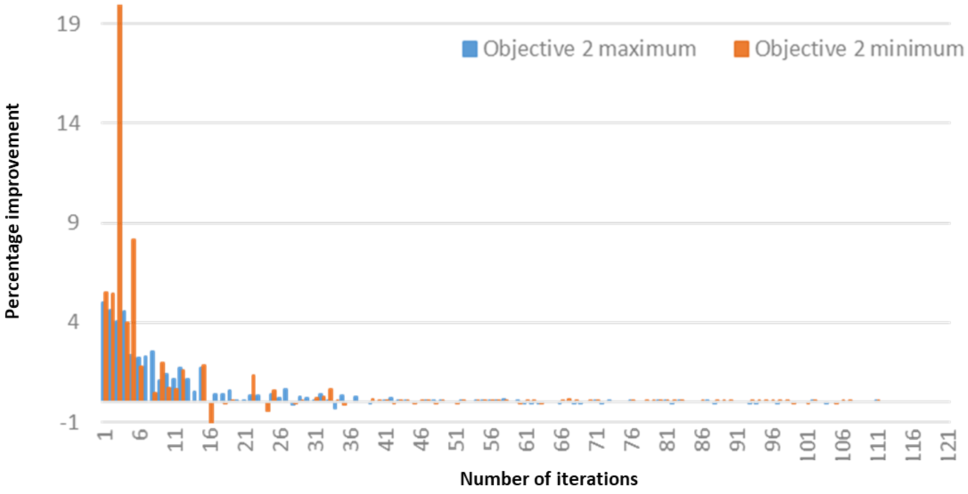


Figure 50: Percentage improved per iteration - Objective 2

Table 11 shows the evaluation of the difference between the allocated employment and the actual employment for the final run. Table 11 includes home-based jobs, which were not part of the previous runs. While the algorithm was developed, only the non-residential buildings and the small holdings and farmstead buildings from the residential buildings were used to test the algorithm. The residential buildings made up the bulk of the total buildings (93.91%) but accounted for less of the overall employment capacity (16.45%). Therefore, it was decided to test the algorithm using the non-residential buildings and then when the algorithm was finalised add the residential buildings.

Table 11: Validation results for final run

SIC code	Economic sector	Allocated employment	Actual employment	Difference	%
1	Agriculture, forestry, and fisheries	12 408	11 733	675	5.75
2	Mining and quarrying	11 815	11 692	123	1.05
3	Manufacturing	165 364	165 376	-12	-0.01
4	Electricity, gas, and water	5 779	5 662	117	2.07
6	Wholesale and retail trade, catering, and accommodation	212 229	211 024	1 205	0.57
7	Transport, storage, and communication	73 268	75 360	-2 092	-2.78
8	Finance, insurance, real estate, and business services	205 375	205 343	32	0.02
9	General government	117 836	123 930	-6 094	-4.92
10	Community, social, and personal services	194 396	190 856	3 540	1.85
	Home based jobs				
0	Private households (includes domestic workers)	87 800	87 800	0	0.00
6	Wholesale and retail trade, catering, and accommodation	54 645	54 645	0	0.00
8	Finance, insurance, real estate, and business services	54 584	54 584	0	0.00
	Overall allocation				
	Total employment	**1 195 499**	**1 198 005**	**-2 082**	**-0.21**

From Table 11 it can be seen that overall the algorithm under allocated 2082 employment opportunities for the entire CoE. This led to an overall accuracy of 0.21%. An overall accuracy of 1% over or under allocation was deemed acceptable for the purposes of this project. Therefore, the final allocation was deemed acceptable. The algorithm was less accurate per economic sector than for the overall allocation. Many economic sectors had an accuracy of less than 1%. There were six economic sectors that did not fall under this threshold.

The AGRICULTURE, FORESTRY, AND FISHERIES SECTOR and the GENERAL GOVERNMENT SECTOR had the largest difference between the allocated and actual employment opportunities. The AGRICULTURE, FORESTRY, AND FISHERIES SECTOR over allocated by 5.75%. The GENERAL GOVERNMENT SECTOR under allocated by 4.92%. Overall the six sectors along with the small differences within the other sectors balanced each other out, which led to a more accurate overall allocation. Because of this, the final allocation was deemed accurate enough.

The final validation measure was examining the algorithm's Pareto front for the final run. Figure 51 depicts the resulting Pareto front. From the figure, it can be seen that the Pareto front did not consist of many solutions. This means that only a few of the individuals or solutions were not dominated and could not be improved upon by the algorithm. The Pareto front depicts the trade-offs that are available between the non-dominated solutions. Therefore, it provides the best solutions for the problem, but the final solution that is chosen would depend on what exactly a user deems as most the most important factor to consider in the final solution.

For the employment allocation, the trade-off is between the total number of employment opportunities per economic sector and the total number of employment opportunities per building. Thus, the trade-off is that the solution could either be more accurate per building and less accurate per economic sector or vice versa. Solution A on the graph is a solution where the results would be more accurate in terms of the number of employment opportunities per economic sector but not all the buildings would be filled to capacity. Solution D on the graph is the opposite, with this solution the buildings will be filled to capacity but the total number of employment opportunities per economic sector would be less accurate. With solutions B and C the trade-off is not as skewed to favour one objective. With these solutions, there is a balance between staying accurate in the total number of employment opportunities per economic sector and ensuring that the allocated employment opportunities are as close to capacity as possible.

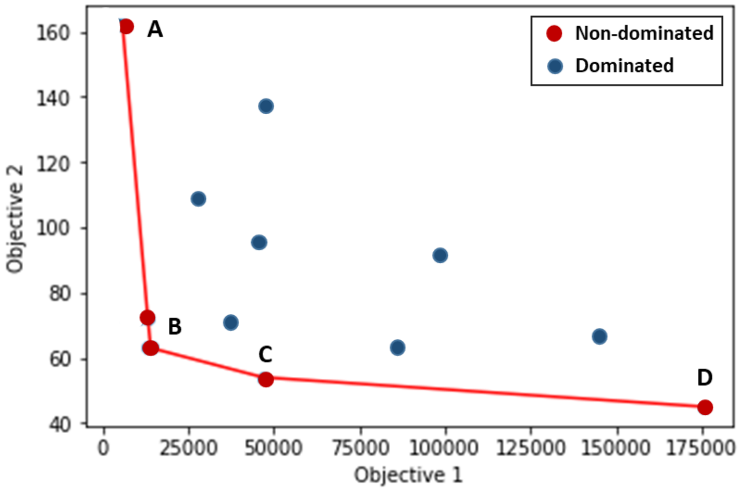


Figure 51: Pareto front graph for the final run

5.4 Discussion of results

One of the benefits of having employment data at such a fine scale is the fact that the data can be aggregated to different scales or levels. The different levels allow different information to be extracted from the data. The results of the aggregated data also provide a different type of information than the employment data that is provided by Census. In the Census data, the number of employment opportunities that are provided per sub-place or main place is the information of the individuals that lived in the specific sub-place or main place. Thus, the information is indicative of the number of individuals that live in the sub-place or main place that is employed.

This does not necessarily mean that those individuals work in the same sub-place as they live. Since the employment total per sub-place or main place are not the total number of employment that is available in the sub-place or main place, the Census data does not provide an accurate spread of employment opportunities across the municipality. Therefore, the results are presented at various levels, which includes sub-place level, main place level, and ward level. The number of allocated employment opportunities are shown at each of the levels. An employment density map was also created at the sub-place level. Figure 52 shows the total number of employment opportunities per sub-place for the CoE.

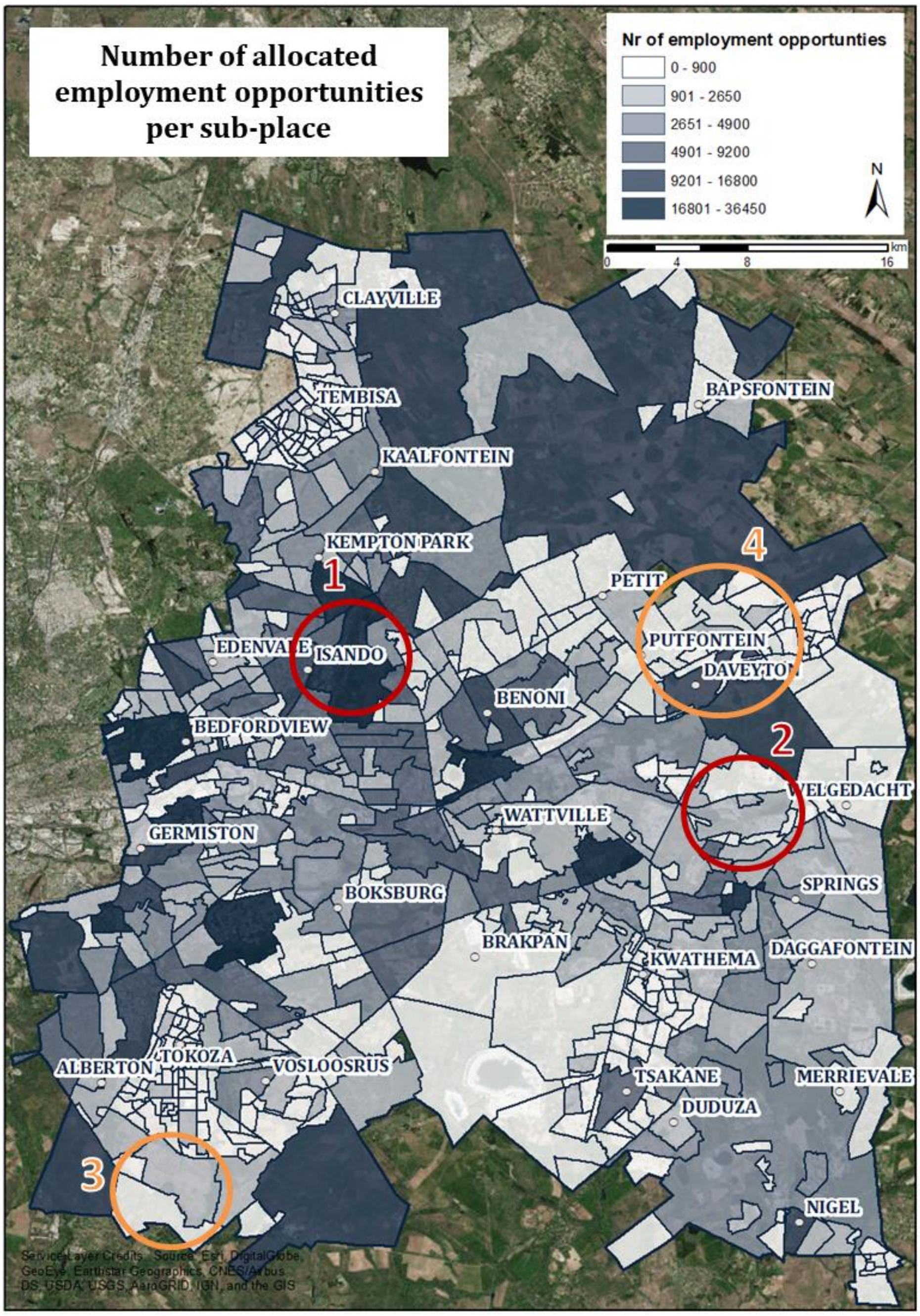


Figure 52: Number of allocated employment opportunities per sub-place for the City of Ekurhuleni

The two red circles on the map show two areas where there is a major difference between the Census data and the allocated employment opportunities. The first circle to the East is the sub-place where O.R Tambo International Airport is located. In the original Census data, there is 0 number of individuals employed in this sub-place. This is because there are no people that live in this sub-place. In the map, it can be seen that in the sub-place there are between 16801 and 36450 number of employment opportunities. This is more accurate as the O.R Tambo airport is one of the major employment hubs in the municipality. There are also many businesses surrounding the airport that provide products and services to the airport.

The second circle shows the sub-place where the Geduld Proprietary Mines are located. Once again, this is not an area where any individuals live, therefore there are no employment opportunities in this sub-place in the Census data. On the map, it can be seen that this sub-place fall in the second classification area of between 901 and 2650 employment opportunities. This is a significant increase in employment opportunities. It is, therefore, an area of importance when looking at sub-places that provide a large number of employment opportunities. There are many other examples of areas that have a large number of employment opportunities that are not represented in the Census data.

The reverse of this situation is also true. There are also areas where the Census data has a large number of individuals that are employed in the area, but this is only because the sub-place is predominantly residential. This means there are not actually that many employment opportunities available in the sub-place. The orange circles show two areas where this is the situation. The circle numbered 3 in the South-West part of the municipality shows the Palm Ridge sub-place. In the Census data, this is an area that shows a significant number of employed individuals, but it is also an area that is mostly residential. In the map, it can be seen that there are a lower number of employment opportunities and it falls in the second classification class of 901 - 2650 employment opportunities.

The circle numbered 4 in the East part of the municipality includes multiple sub-places. Once again, this is a mostly residential area and therefore the sub-places has a higher number of employed individuals. In reality, there are not as many non-residential buildings in the area, therefore on the map, it can be seen that many of the sub-places within the circle have a lower number of employment opportunities.

Figure 53 shows the total number of employment opportunities per square kilometre on sub-place level for the CoE. This map shows the density of employment opportunities per sub-place. The density map was only calculated on sub-place level because there is a lot of variation in the sizes of the sub-places. This means that larger sized sub-place will have more employment opportunities than those of smaller size, although these larger areas are not necessarily an employment hotspot in the municipality. Identifying the employment hotspots also made more sense with the sub-place level where more detail is available.

Therefore calculating the density of the employment opportunities will highlight the employment hotspot in the municipality where there are more employment opportunities per square kilometre. The figure indicates that there is a higher density of employment opportunities in the West of the CoE and the Eastern parts of the CoE. This in line with the areas of high non-residential density that were identified in Figure 23 in Section 4.3.1.

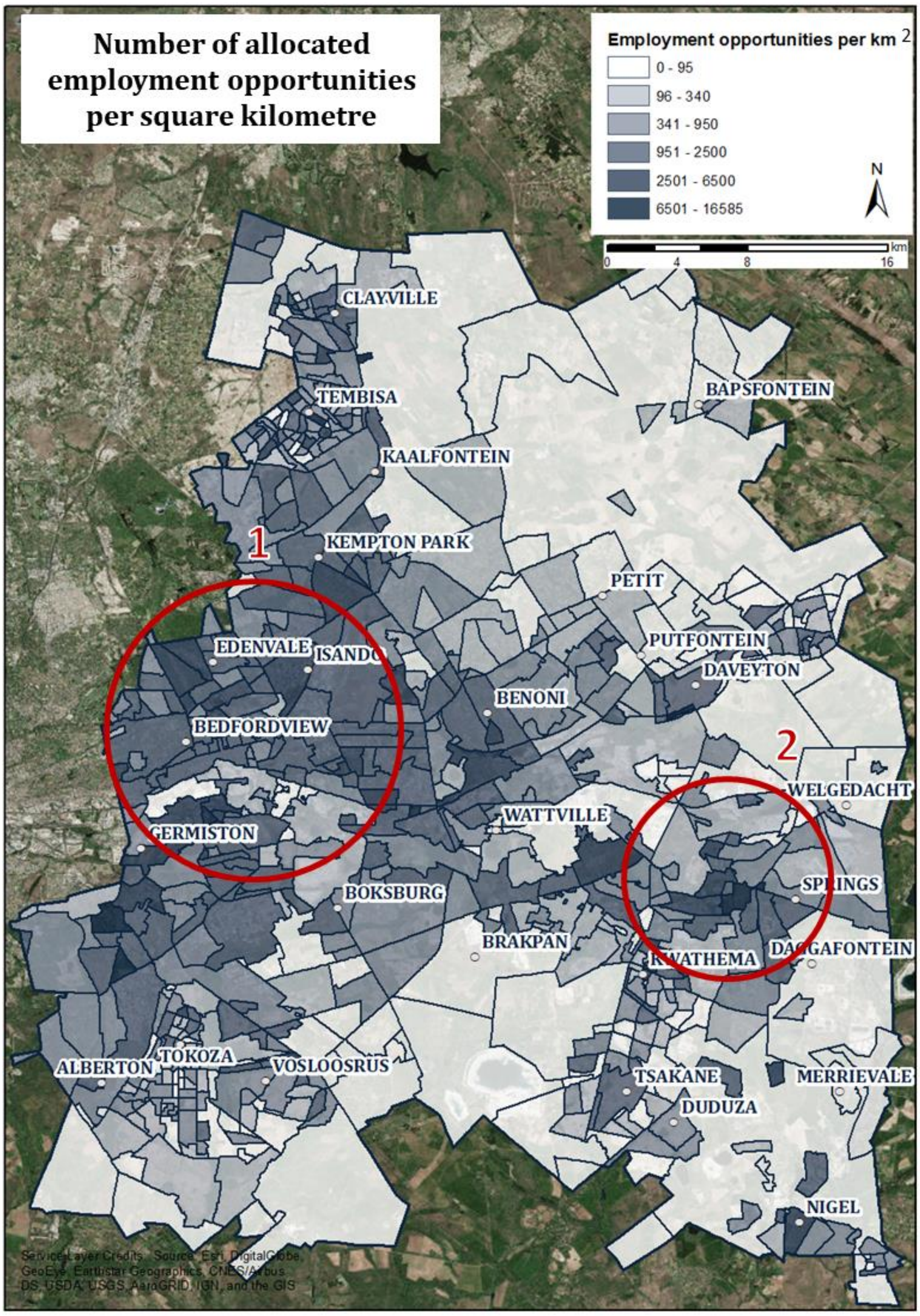


Figure 53: Number of allocated employment opportunities per square kilometre for the City of Ekurhuleni

Figure 54 illustrates the total number of employment opportunities per main place. Kempton Park and Germiston are the areas that have the highest number of employment opportunities. This is to be expected as these are the two areas with some of the largest CBDs in the CoE. These two main places are also close to the Johannesburg CBD in the City of Johannesburg, which is one of the largest employment nodes in Gauteng. Following Kempton Park and Germiston with the highest number of employment opportunities are the Boksburg, Benoni, Springs, and Alberton main places.

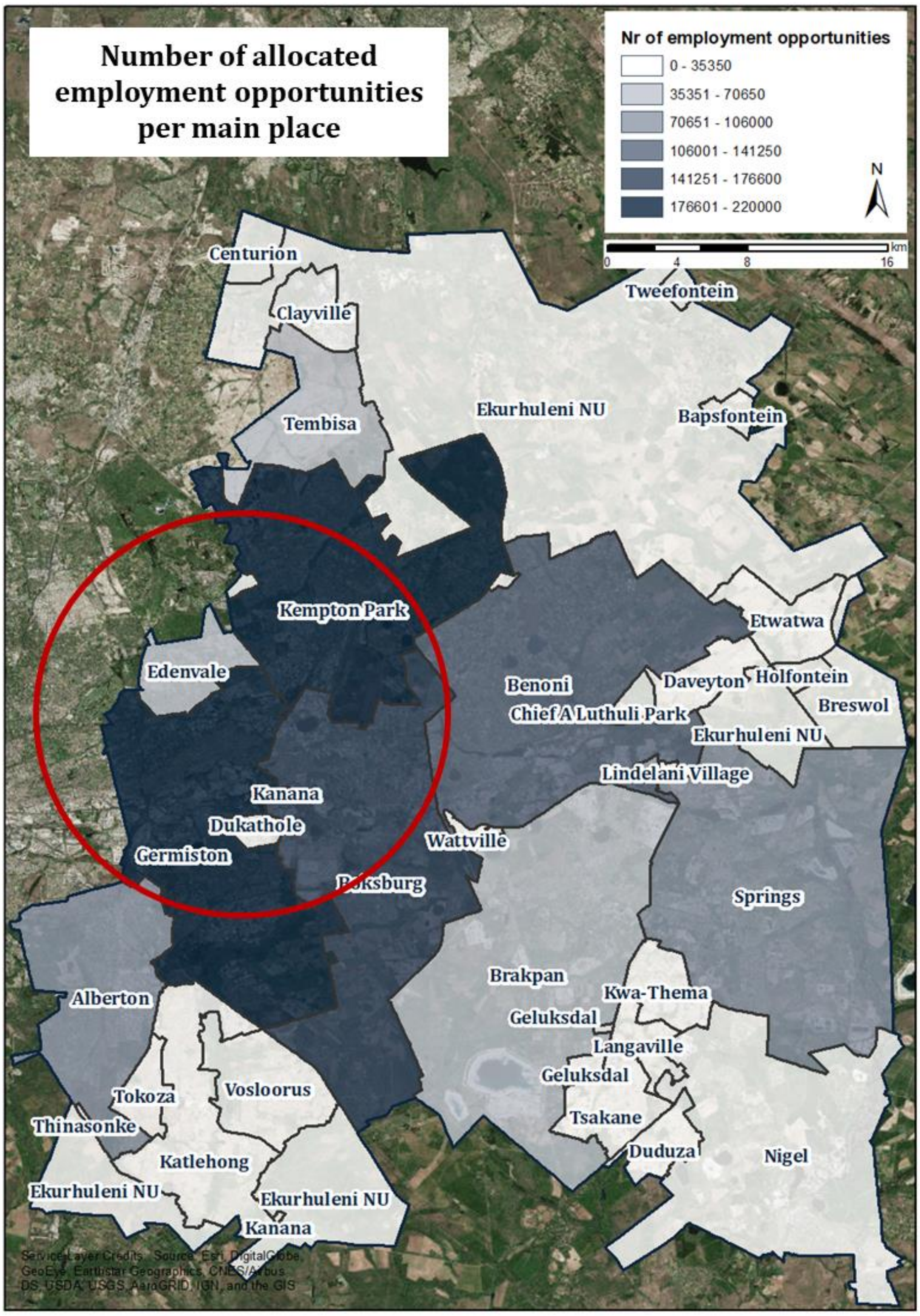


Figure 54: Number of allocated employment opportunities per main place for the City of Ekurhuleni

Figure 55 illustrates the difference between the actual number of employment opportunities and the allocated number of employment opportunities per metro region. The map on the left is the total number of actual employment opportunities. The map in the middle is the number of allocated employment opportunities. The map on the right shows the difference between the previous two maps. On the difference map, the green shades are indicative of areas where the actual number of employment opportunities were less than the allocated employment opportunities. The blue shades are indicative of areas where the actual number of employment opportunities are more than the number of allocated employment opportunities.

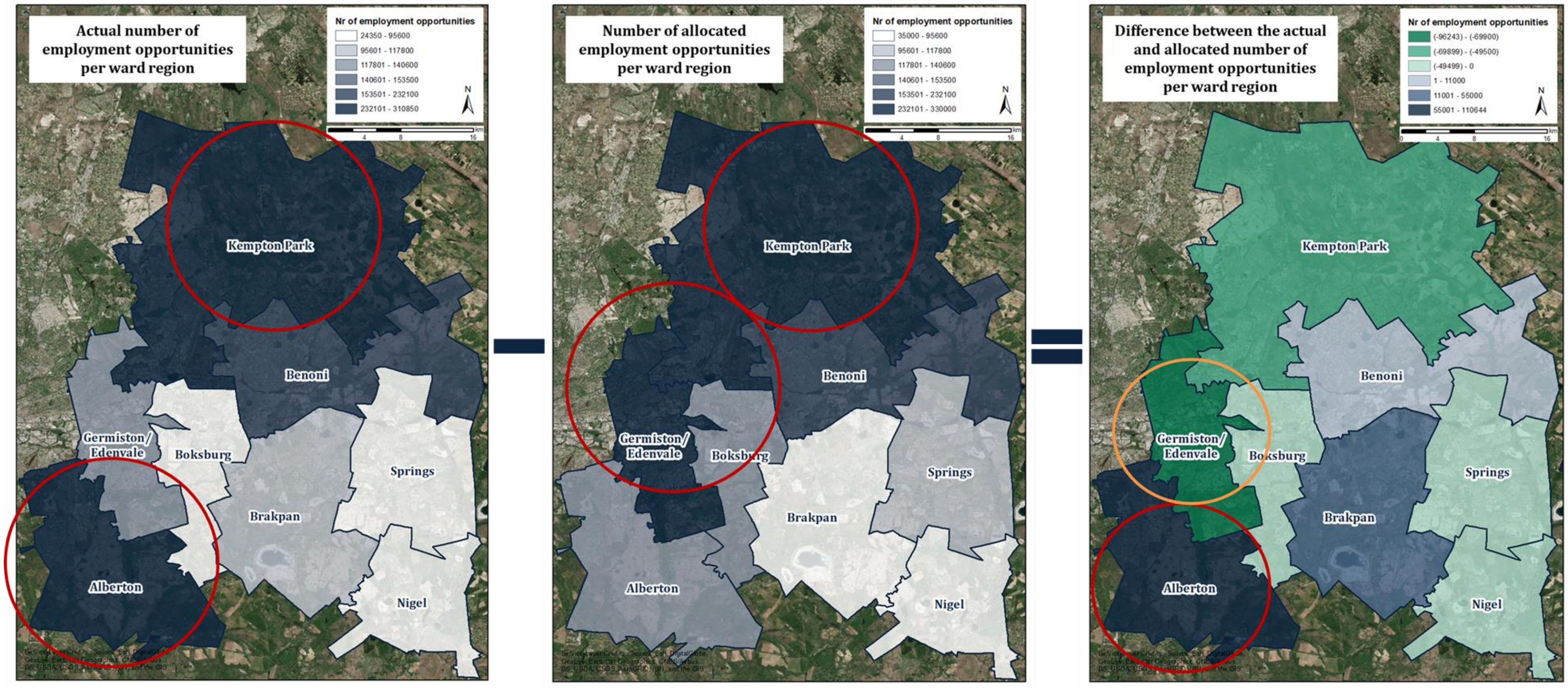


Figure 55: Difference between actual and allocated number of employment opportunities for the City of Ekurhuleni.

As was the case with the sub-place Census data, the map on the left with the actual number of employment opportunities per metro region (i.e. Census data) illustrates the total number of individuals that live in a metro region and that is employed. Alberton and Kempton Park are the two metro regions that have the highest number of individuals that are employed that live in those metro regions. This is in line with what was shown on the high density residential buildings map in Section 4.3.1. On this density map, Alberton had the highest density of residential buildings in the entire CoE. Therefore, it makes sense that it would have the highest number of individuals living there that are employed.

The allocated employment opportunities map in the middle illustrates that the Kempton Park and Germiston/Edenvale metro regions have the highest number of employment opportunities. In the high density non-residential buildings map in Section 4.3.1, both the Germiston/Edenvale and the Kempton Park regions had a very high density of non-residential buildings. Therefore, it is expected that these would be the regions with the highest number of employment opportunities. Kempton Park is also the largest metro regions in terms of size (i.e. area), thus it would allow for a large number of employment opportunities.

The difference map between the actual and allocated number of employment opportunities on the right it shows that Alberton has the largest difference where the actual number of employment opportunities is more than the allocated value. This is to be expected, as Alberton is the regions with the highest density of residential buildings, but also one of the areas with the lowest density of non-residential buildings. From this, it can be deduced that there are many individuals who live in the Alberton metro region, but travel to a different area for work.

Other areas that show a large positive difference between the actual and allocated number of employment opportunities are the Benoni and Brakpan metro regions. It can be expected that a similar situation occurs in these regions as in the Alberton metro region. Germiston/Edenvale has the largest negative value where the actual number of employment opportunities are less than the allocated number of employment opportunities. From this it can be deducted that these are the areas where individuals from other metro regions travel to for work.

With the algorithm developed and validated for the City of Ekurhuleni, the next step in the development of the overall methodology was to test the algorithm on other study areas.

Chapter 6: Testing the algorithm in other case study areas

6.1 Introduction

One of the requirements for the dissertation was to develop a methodology to disaggregate employment data that could be applied in any study area. Therefore, the methodology was applied in two other study areas, namely the City of Johannesburg Municipality and the City of Tshwane Municipality. This chapter gives an overview of the two areas; it also discusses the validation of the employment allocation results in the area, and then finally provides the results for each of the two municipalities.

6.2 Overview of case study areas

6.2.1 Overview of City of Johannesburg

The City of Johannesburg (CoJ) is a metropolitan municipality that is located in Gauteng of South Africa. Figure 56 shows the location of the CoJ within South Africa. The CoJ is made up of eighteen cities. These cities include Johannesburg, which is the largest city in South Africa. The municipality covers a geographical area of 1 645km^2 (Municipalities of South Africa, 2019). Figure 57 shows the seven metro regions located in the CoJ.

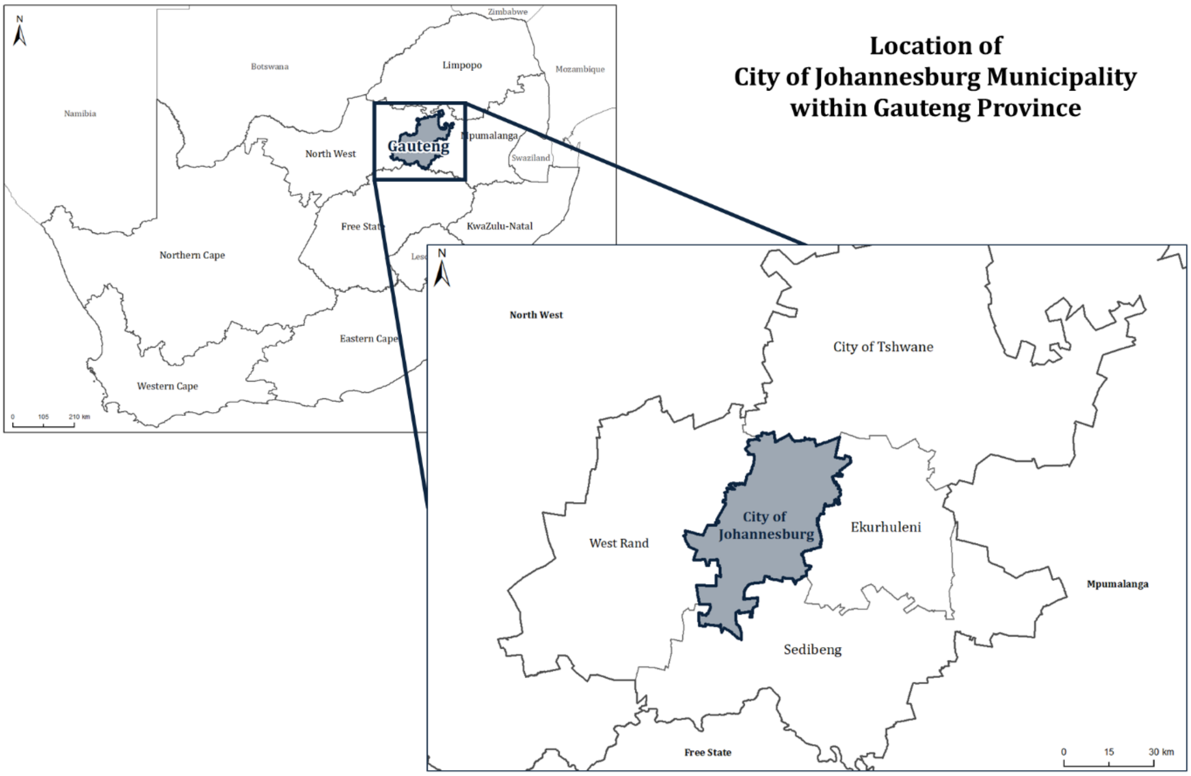


Figure 56: Location of City of Johannesburg within Gauteng province

The CoJ had an estimated population of 5 367 688 people in 2017, with a population density of around 3 263 people per km^2. An estimated 3 814 772 of the population in 2017 were of working age. Out of this an estimated 2 016 906 of the population were employed in 2017. This gives an estimated 25.49% unemployment rate in 2017 (Municipalities of South Africa, 2019; Quantec, 2019). The CoJ is projected to have an increase of 3.6 million people by 2050. This is an estimated population growth of 84%. It is also projected that the CoJ will be the fastest-growing municipality in Gauteng (Le Roux, et al., 2019).

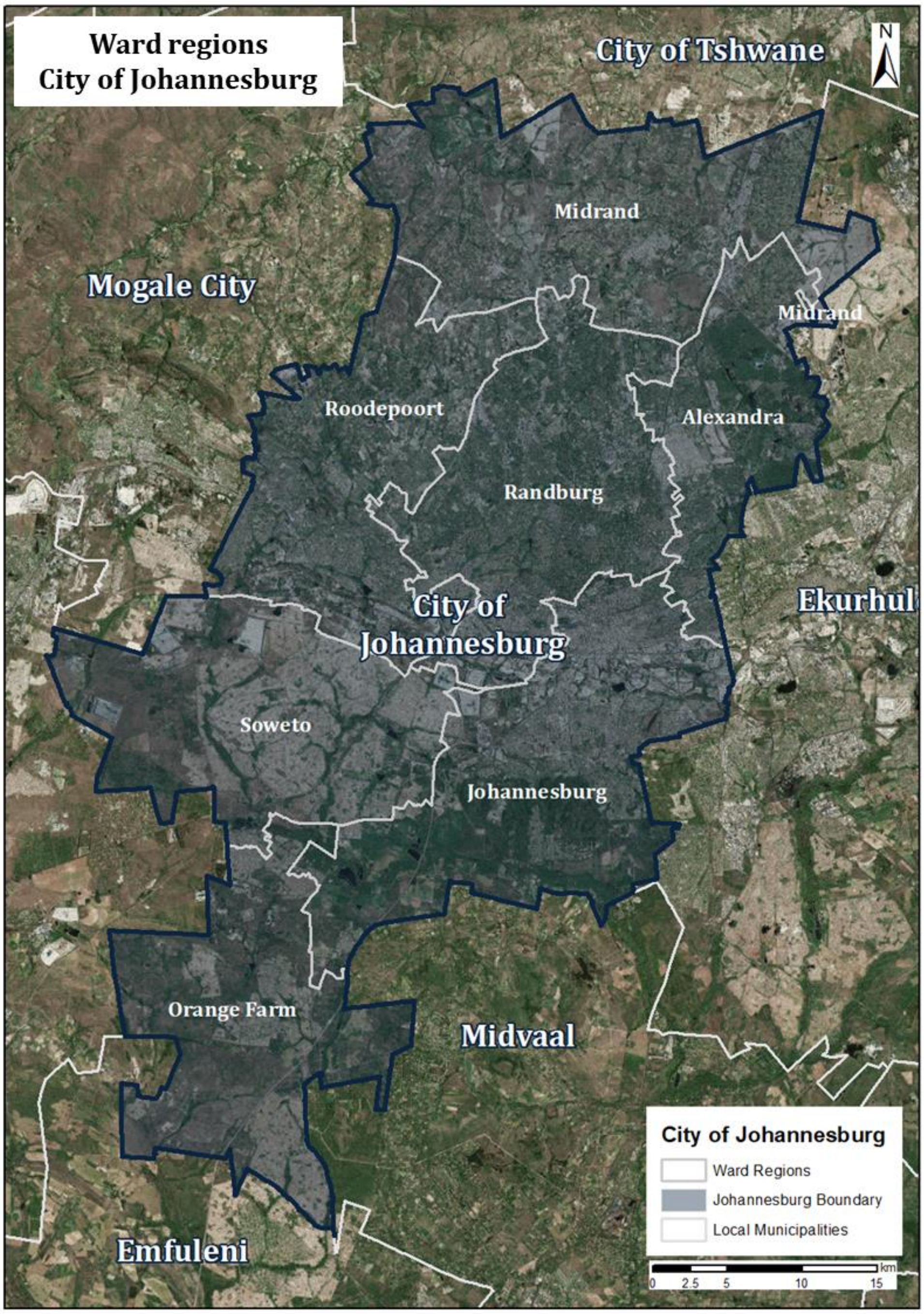


Figure 57: Map showing the metro regions in the City of Johannesburg

The CoJ is known as the economic hub of South Africa as it is not only the largest contributor to Gauteng's economy but also the largest contributor to the economy of South Africa. The CoJ has four main economic sectors that makeup around 80% of the municipality's economy. These sectors include the FINANCE AND BUSINESS SECTOR, MANUFACTURING SECTOR, TRADE SECTOR, and COMMUNITY SERVICES SECTOR (Municipalities of South Africa, 2019).

6.2.2 Overview of City of Tshwane

The City of Tshwane (CoT) is a metropolitan municipality that is located in the Gauteng Province of South Africa. Figure 58 shows the location of the CoT within South Africa. The CoT consists of twenty-one cities, including Pretoria, which is the capital city of South Africa. Since the capital city is located here, many embassies are also located here. The municipality covers a geographical area of 6 298km^2 which makes it the largest size municipality in Gauteng (Municipalities of South Africa, 2019).

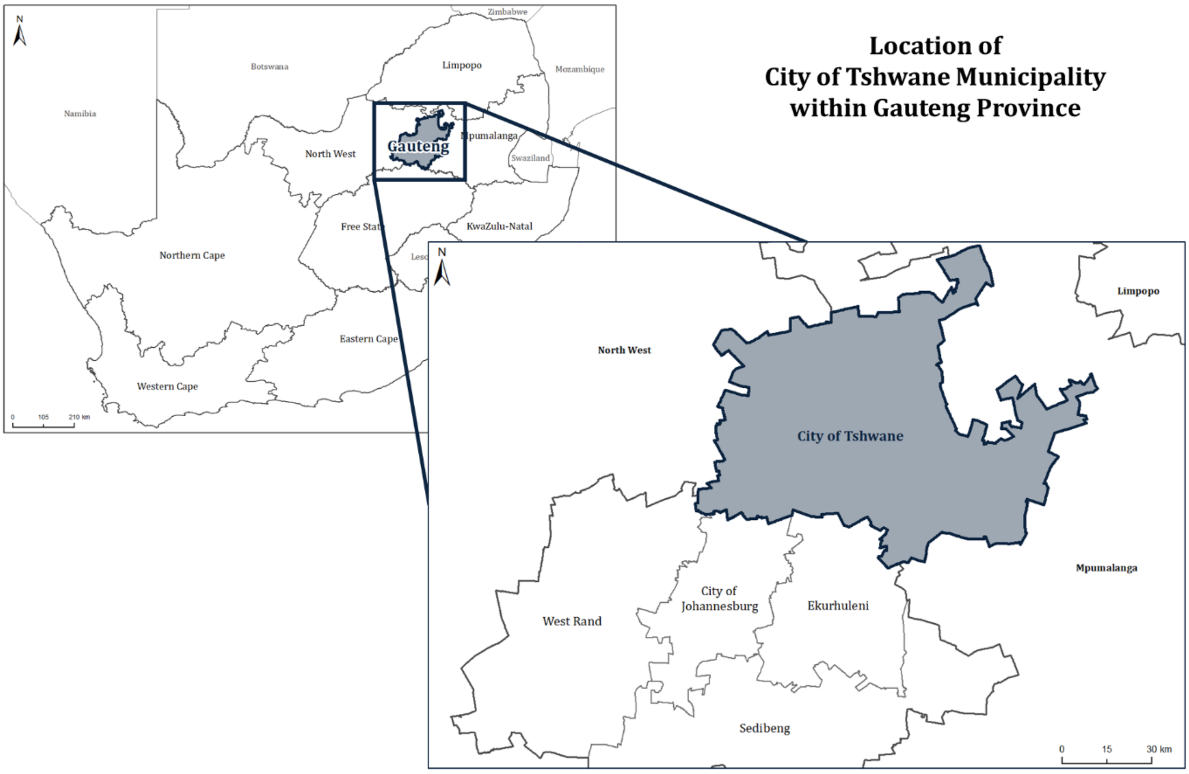


Figure 58: Location of City of Tshwane within Gauteng province

The CoT had an estimated population of 3 429 235 people in 2017, with a population density of around 545 people per km^2. An estimated 2 423 046 of the population in 2017 were of working age. Out of this an estimated 1 264 719 of the population were employed in 2017. This gives an estimated 24.39% unemployment rate in 2017 (Quantec, 2019). The CoT is projected to have increase an of 2.2 million people by 2050. This is an estimated population growth of 76%. It is also projected that the CoT will be the second-fastest-growing municipality in Gauteng (Le Roux, et al., 2019). Figure 59 displays the seven metro regions located in the CoT.

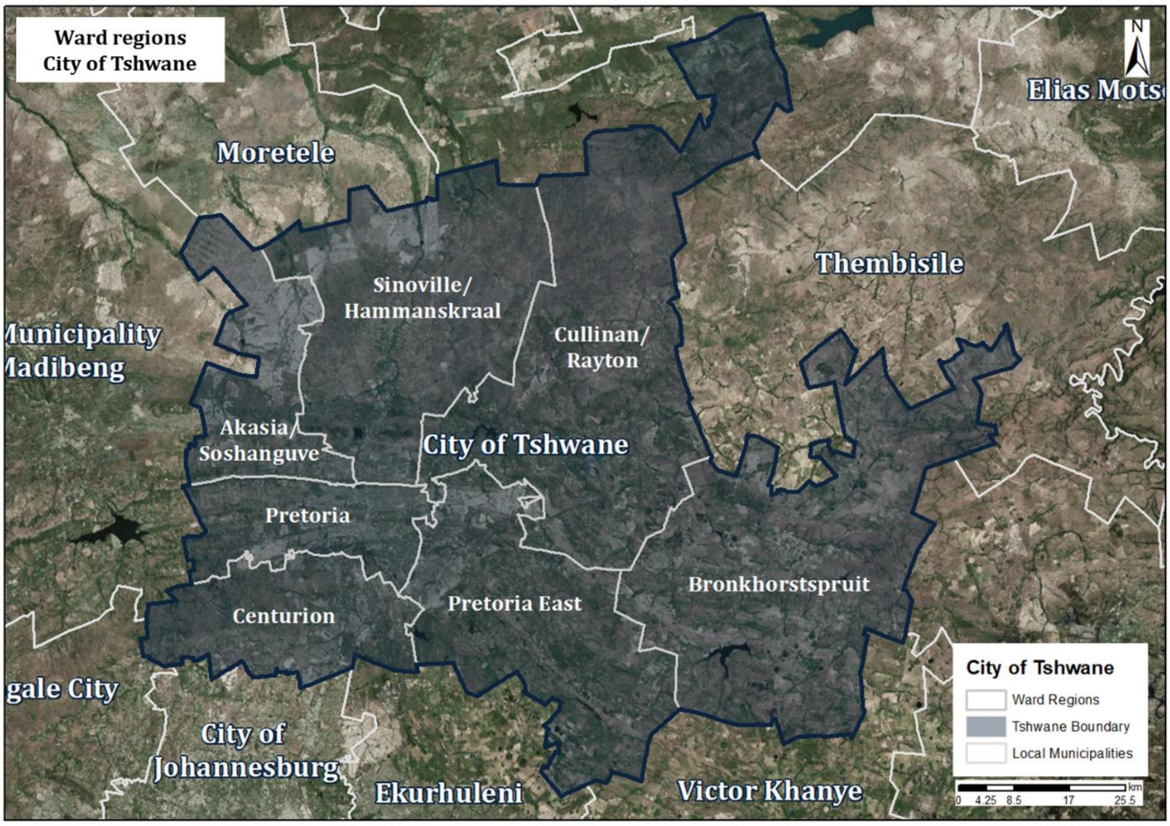


Figure 59: Map showing the metro regions in the City of Tshwane

Along with the many embassies that were mentioned earlier, the Union buildings are also located in the CoT. This is why the CoT is known as the administrative hub of South Africa. The CoT is the second largest contributor to Gauteng's economy and is one of the top contributors to the South African economy. The CoT has a very diverse economy with various economic sector contributing to the overall economy. The economic sectors with the highest contributing include GENERAL GOVERNMENT SECTOR and FINANCE AND BUSINESS SECTOR that contributes to around 50% of the CoT's total economy. Other high contributing sectors include MANUFACTURING, WHOLESALE AND RETAIL, and TRANSPORT (Municipalities of South Africa, 2019).

6.3 Implementation of the algorithm for case study areas

The entire methodology that was followed during the job allocation process for the CoE was performed for the CoJ and CoT. This includes the data preparation and the implementation of the algorithm. The exact steps were followed during the data preparation, as the same datasets that were available for the CoE were also available for CoJ and CoT. Table 12 and 13 show the datasets that were used for the CoJ and CoT respectively.

Table 12: Datasets used for the City of Johannesburg

	Dataset	Data custodian
1	City of Johannesburg residential building dataset	CSIR
2	City of Johannesburg non-residential building dataset	CSIR
3	City of Johannesburg building footprints	CSIR
4	Sub place boundaries for the City of Johannesburg	StatsSA
5	Main place boundaries for the City of Johannesburg	StatsSA
6	Metro regions in the City of Johannesburg	StatsSA
7	Local municipal boundary for the City of Johannesburg	StatsSA
8	Primary and High schools in the City of Johannesburg	Gauteng Department of Education
9	Police stations in the City of Johannesburg	South African Police Service
10	Employment per economic sector for the City of Johannesburg	Quantec

Table 13: Datasets used for the City of Tshwane

	Dataset	Data custodian
1	City of Tshwane residential building	CSIR
2	City of Tshwane non-residential building	CSIR
3	City of Tshwane building footprints	CSIR
4	Sub place boundaries for the City of Tshwane	StatsSA
5	Main place boundaries for the City of Tshwane	StatsSA
6	Metro regions in the City of Tshwane	StatsSA
7	Local municipal boundary for the City of Tshwane	StatsSA
8	Primary and High schools in the City of Tshwane	Gauteng Department of Education
9	Police stations in City of Tshwane	South African Police Service
10	Employment per economic sector for the City of Tshwane	Quantec

The main steps that were performed during the data preparation included (as in Section 4.3):

1. Preparing the buildings datasets,
2. Adding the footprint size for the non-residential buildings,
3. Linking each building to an economic sector,
4. Calculating the final employment capacity for each of the buildings.

After these steps were performed, there was a base dataset for each of the municipalities. The CoJ had a total of around 1 070 000 buildings within the base dataset and the CoT had a total of around 760 000 buildings. Figure 60 shows the distribution of residential buildings across CoJ. As was the case for the CoE, the darker shade of blue is indicative of areas with a higher density of buildings. From this map, it can be seen that the highest density of residential buildings is located in the Soweto metro region and parts of the Midrand, Alexandra, and Orange Farm metro regions. Many of these highly dense areas are high-density informal areas within the CoJ. The density of residential buildings is much lower in the Johannesburg, Randburg, and Roodepoort metro regions.

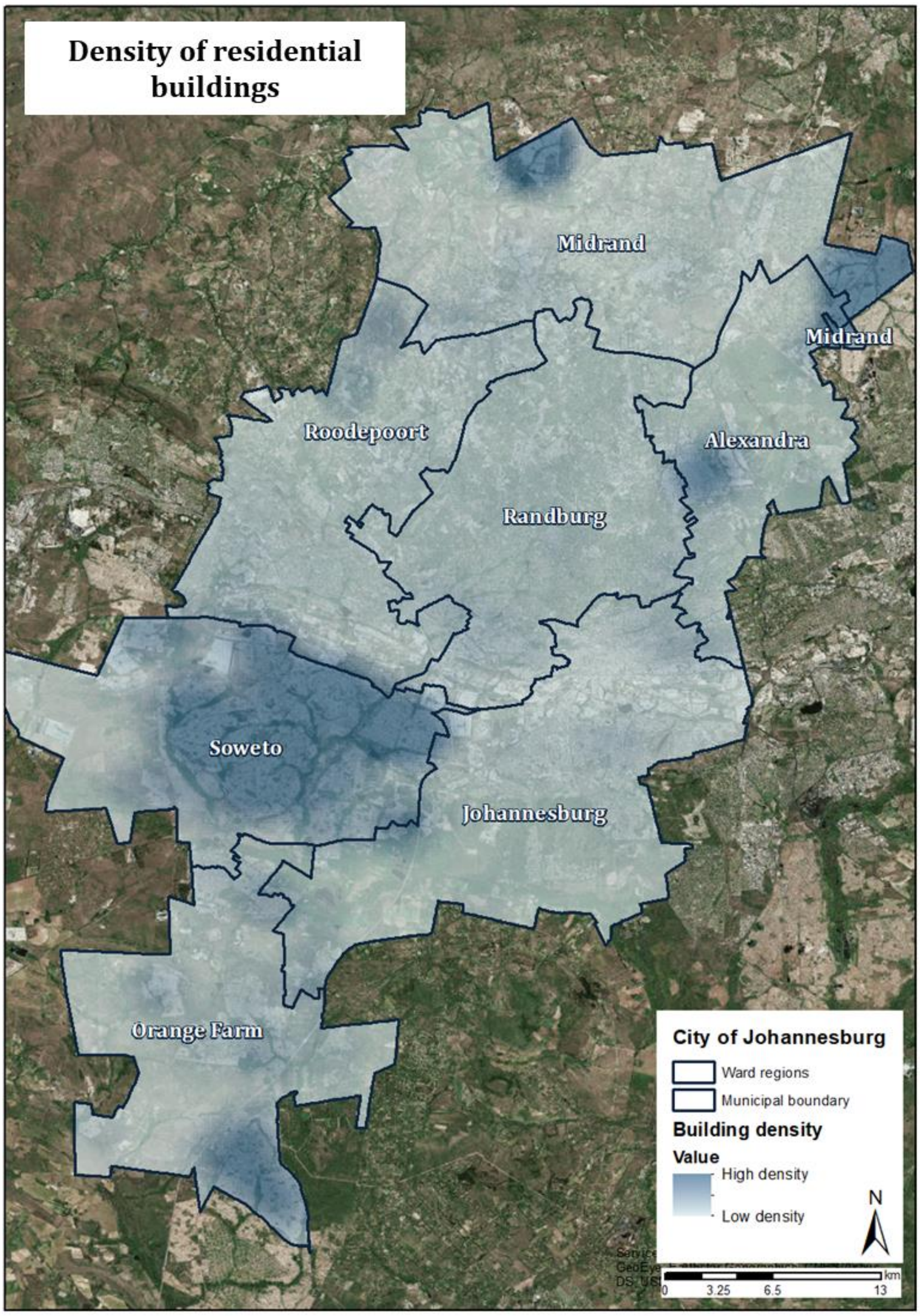


Figure 60: Map showing areas with a high density of residential buildings

Figure 61 shows the distribution of the non-residential buildings across CoJ. The darker shade of red is indicative of areas with a higher density of buildings. The areas of higher density for the non-residential buildings are more spread across the municipality, while the areas of higher density of the residential buildings were more isolated and more prominent in certain metro regions. Each of the metro regions displays some areas of higher density in non-residential buildings, but the Johannesburg, Randburg, and Roodepoort metro regions show the highest density of non-residential buildings.

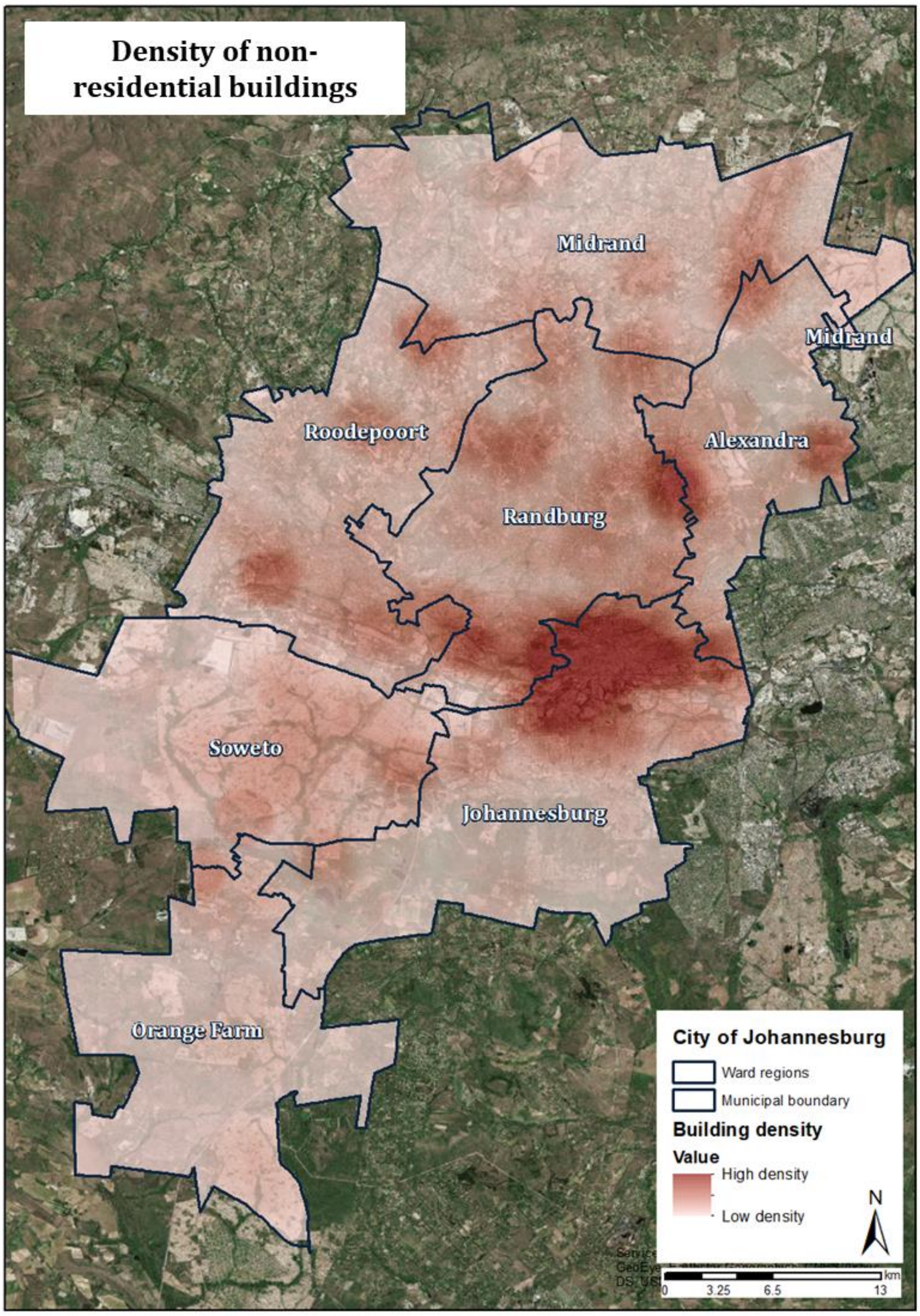


Figure 61: Map showing areas with a high density of non-residential buildings

Figure 62 depicts the distribution of residential buildings across CoT. As with the earlier maps, the darker shade of blue is indicative of areas with a higher density of buildings. From this map, it can be seen that the highest density of residential buildings is located in the Akasia/Soshanguve metro region and parts of the Pretoria, Centurion, and Pretoria East metro regions. As with the CoJ, many of these dense areas are high-density informal areas. The density of residential buildings is much lower in the Cullinan/Rayton metro region and parts of Sinoville/Hammanskraal and Bronkhortspruit metro regions.

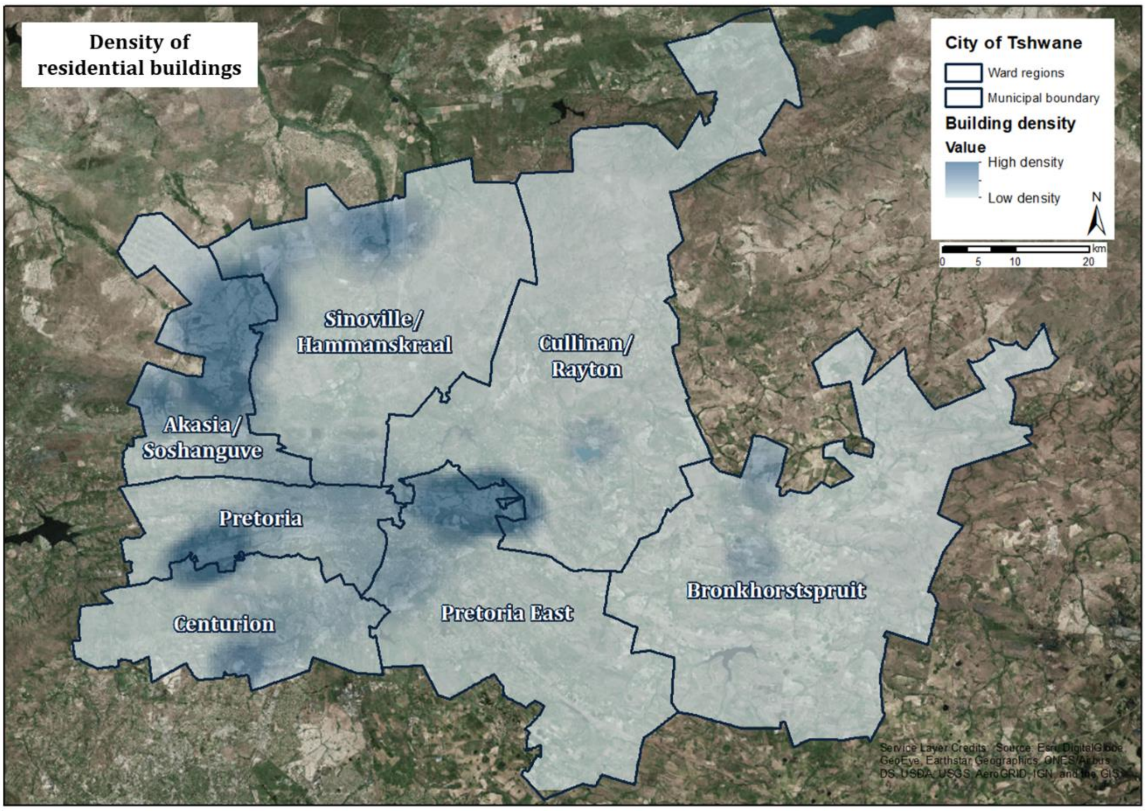


Figure 62: Map showing areas with a high density of residential buildings

Figure 63 illustrates the distribution of the non-residential buildings across CoT. The darker shade of red is indicative of areas with a higher density of buildings. The areas of higher density can be seen the Pretoria and Centurion metro regions and in parts of the Akasia/Shoshonguve and Pretoria East metro regions. Bronkhorstspruit, Cullinan/Rayton, and Sinoville/Hammanskraal have a few hotspot areas where there is a higher density of buildings.

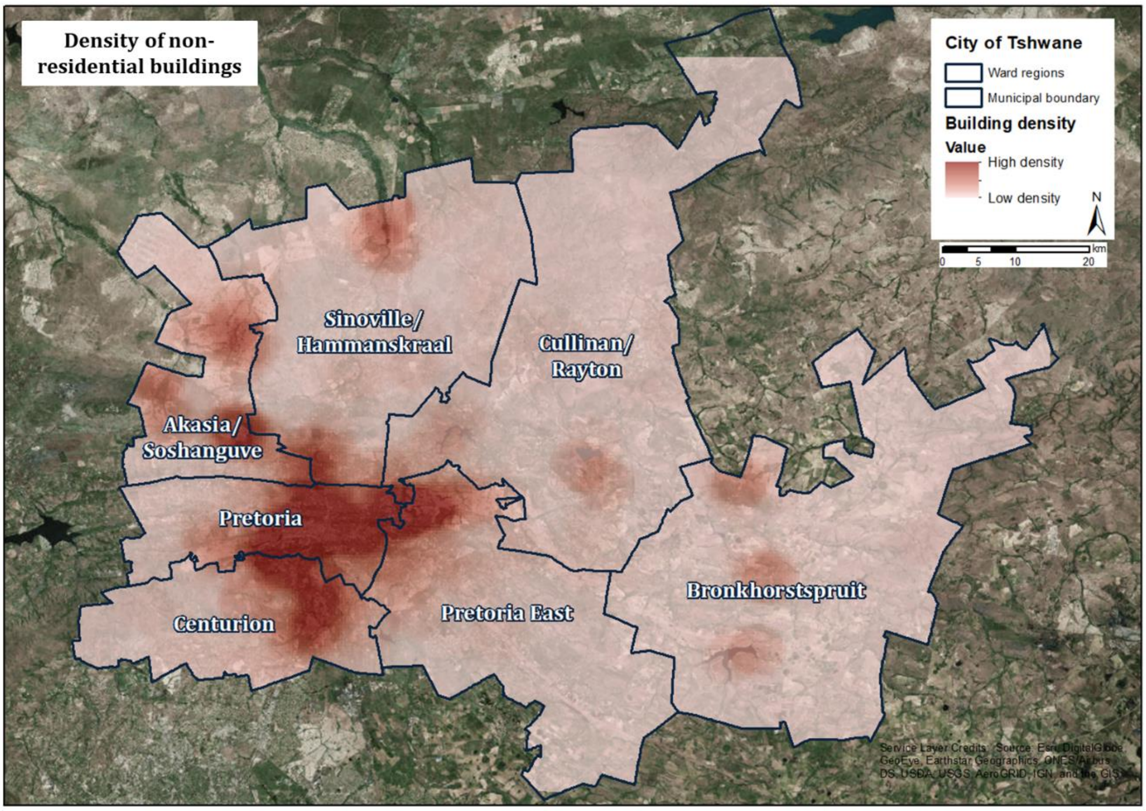


Figure 63: Map showing areas with a high density of non-residential buildings

Once the data preparation was completed and the base dataset was created, the employment allocation algorithm was implemented. The sections that follow will discuss the validation of the algorithm results for both municipalities, as well as the final results for both municipalities.

6.4 Validation of algorithm

Since the same approach and algorithm was used for all three of the municipalities, the same validation techniques were used to evaluate the results for the CoJ and CoT. The first validation technique looks at the performance and behaviour of the algorithm. Figure 64 shows the change in the minimum and maximum values of the two objectives of the fitness function for the CoJ. Figure 65 shows the change in the minimum and maximum values of the two objectives of the fitness function for the CoT.

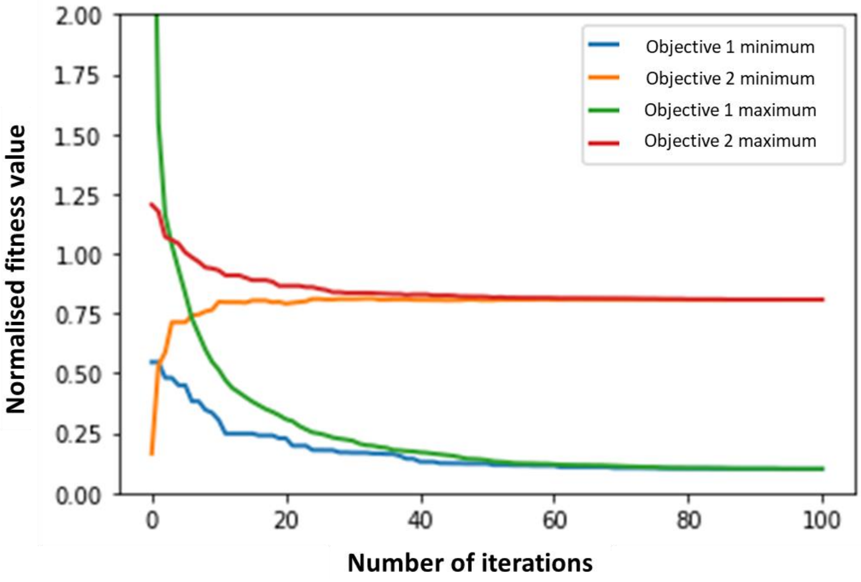


Figure 64: Evaluation criteria graph for the City of Johannesburg

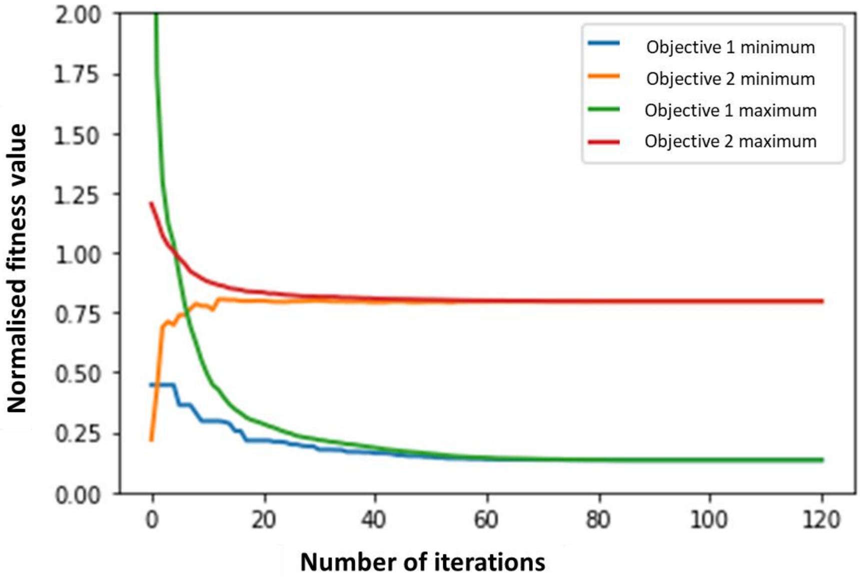


Figure 65: Evaluation criteria graph for the City of Tshwane

Both graphs show a very similar structure to that of the CoE. Initially, the graphs show a lot of variation and during the run of the algorithm, the values start to converge until the point where a better solution could not be found. With the CoJ the amount of improvement started to decrease at around the 20th iteration. The algorithm stopped showing any improvement around the 80th iteration. With the CoT, the amount of improvement started to reduce at around the 15th iteration. The algorithm stopped showing any improvement at around the 100th iteration.

The second validation technique was to compare the total allocated employment opportunities to the total number of actual employment opportunities. The final validation was comparing the allocated employment opportunity per economic sector to the total number of actual employment opportunities per economic sector. Table 14 and 15 show the validation results for the CoJ and CoT respectively. Once again, both municipalities showed similar results to that of the CoE. For both municipalities, the allocation for most of the economic sectors was quite accurate, but again there were two or three sectors where the differences were larger. As with the CoE, the differences balanced each other out, which led to better overall accuracy.

Table 14: Validation results for the City of Johannesburg

SIC code	Economic sector	Allocated employment	Actual employment	Difference	%
1	Agriculture, forestry, and fisheries	13 195	12 981	214	1.65
2	Mining and quarrying	5 163	6 378	-1 215	-19.05
3	Manufacturing	178 977	182 641	-3 664	-2.01
4	Electricity, gas, and water	9 548	7 999	1 549	19.36
6	Wholesale and retail trade, catering, and accommodation	300 451	300 583	-132	-0.04
7	Transport, storage, and communication	86 016	85 638	378	0.44
8	Finance, insurance, real estate, and business services	369 082	369 090	-8	0.00
9	General government	164 788	173 339	-8 551	-4.93
10	Community, social, and personal services	313 229	312 745	484	0.15
	Home based jobs				
0	Private households	140 157	140 157	0	0.00
6	Wholesale and retail trade, catering, and accommodation	104 214	104 214	0	0.00
8	Finance, insurance, real estate, and business services	103 987	103 987	0	0.00
	Overall allocation				
	Total employment	**1 788 807**	**1 799 752**	**-10 945**	**-0.61**

For the CoJ the algorithm under projected by 10 945 employment opportunities. This is 0.61% less than the actual number of employment opportunities. Three sectors had the most deviation from the actual number of employment opportunities. These sectors are the Mining and Quarrying sector; Electricity, Gas, and Water sector; and the General Government sector. The percentage difference between the mining and quarrying sector and the electricity, gas, and water sector is not only the two largest differences, but the two differences are also very similar in size. This could be caused by the fact that the two sectors have a similar number of actual employment opportunities and thus the excess employment capacity for both are similar. These closely related values could balance each other out when the fitness in the algorithm is measured and as a result, the values are not improved, as this error is not highlighted by the fitness value.

For the CoT the algorithm under projected by 3 526 employment opportunities. This is 0.31% less than the actual number of employment opportunities. As with the CoJ, three sectors showed the most deviation. These sectors are the Mining and Quarrying sector; Agriculture, Forestry, and Fisheries sector; and the Manufacturing sector.

Table 15: Validation results for the City of Tshwane

SIC code	Economic sector	Allocated employment	Actual employment	Difference	%
1	Agriculture, forestry, and fisheries	15 134	15 826	-692	-4.37
2	Mining and quarrying	1 595	1 914	-319	-16.67
3	Manufacturing	109 653	112 286	-2 633	-2.34
4	Electricity, gas, and water	4 457	4 373	84	1.92
6	Wholesale and retail trade, catering, and accommodation	158 454	158 700	-246	-0.16
7	Transport, storage, and communication	53 632	53 392	240	0.45
8	Finance, insurance, real estate, and business services	161 273	161 431	-158	-0.10
9	General government	191 895	191 376	519	0.27
10	Community, social, and personal services	200 121	200 442	-321	-0.16
	Home based jobs				
0	Private households	83 867	83 867	0	0.00
6	Wholesale and retail trade, catering, and accommodation	77 572	77 572	0	0.00
8	Finance, insurance, real estate, and business services	77 864	77 864	0	0.00
	Overall allocation				
	Total employment	**1 135 517**	**1 139 043**	**-3 526**	**-0.31**

6.5 Discussion of results

The results for the CoJ and CoT are presented at various scales, which includes sub-place level, main place level, and at metro region level.

6.5.1 Results for City of Johannesburg

Figure 66 depicts the total number of employment opportunities per sub place for the CoJ. From this map, it can be seen that the sub-places with the highest number of employment opportunities are in the Southern parts of the CoJ. Other areas with high numbers of employment opportunities are around the Sandton area and the Johannesburg area. These are all areas where the sub-place is larger in size (i.e. area) and therefore can accommodate a higher number of employment opportunities.

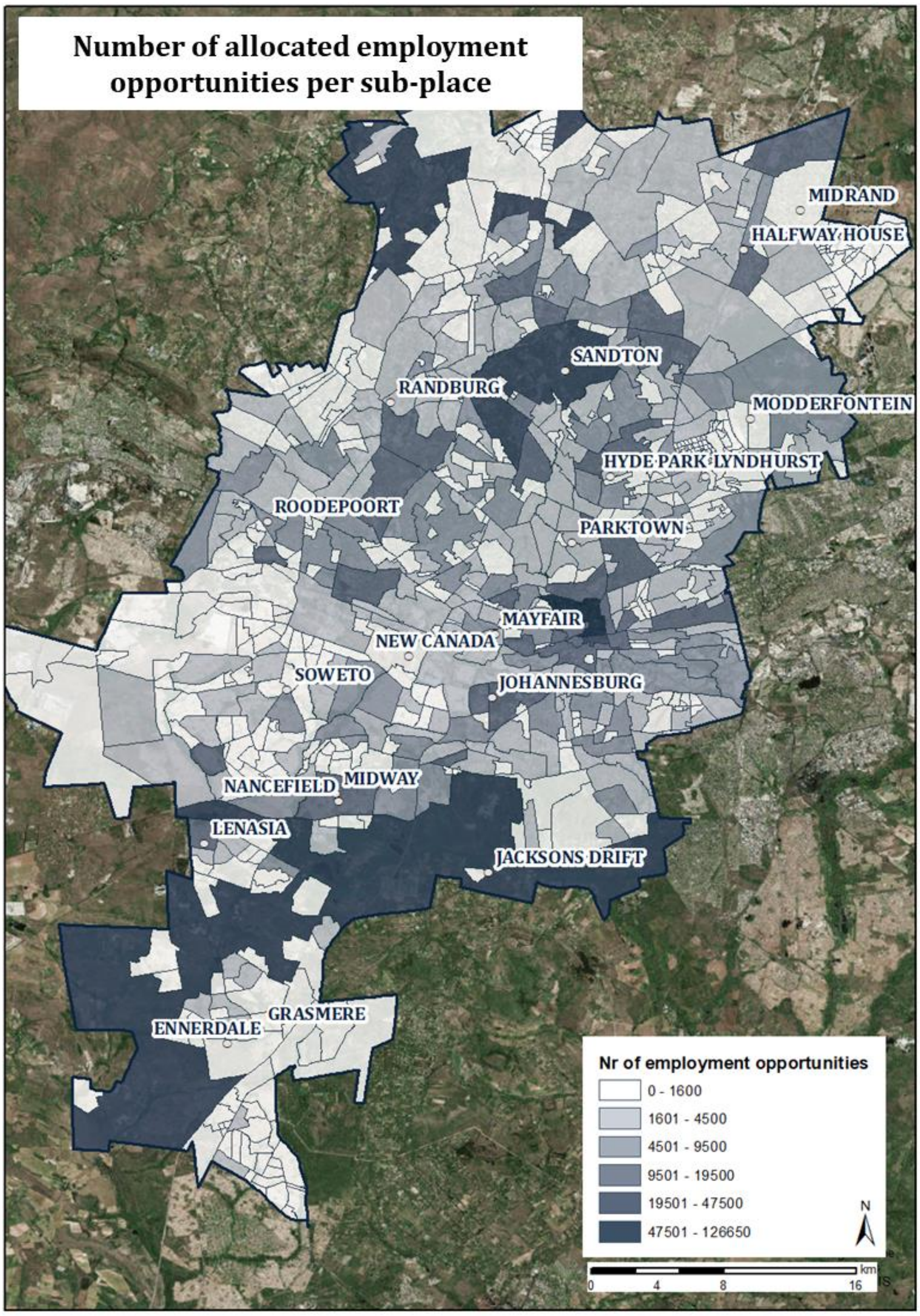


Figure 66: Number of allocated employment opportunities per sub-place for the City of Johannesburg

Figure 67 shows the total number of employment opportunities per square kilometre on sub-place level for the CoJ. From this map, it can be seen that the highest density of employment opportunities are located around the Central business district (CBD) area in central Johannesburg. If this map is compared to Figure 61 in Section 6.3, which shows the density of non-residential buildings, very similar hotspots are identified. The higher density of employment opportunities is expected in this area, as the Johannesburg CBD is one of the largest employment nodes in the entire Gauteng.

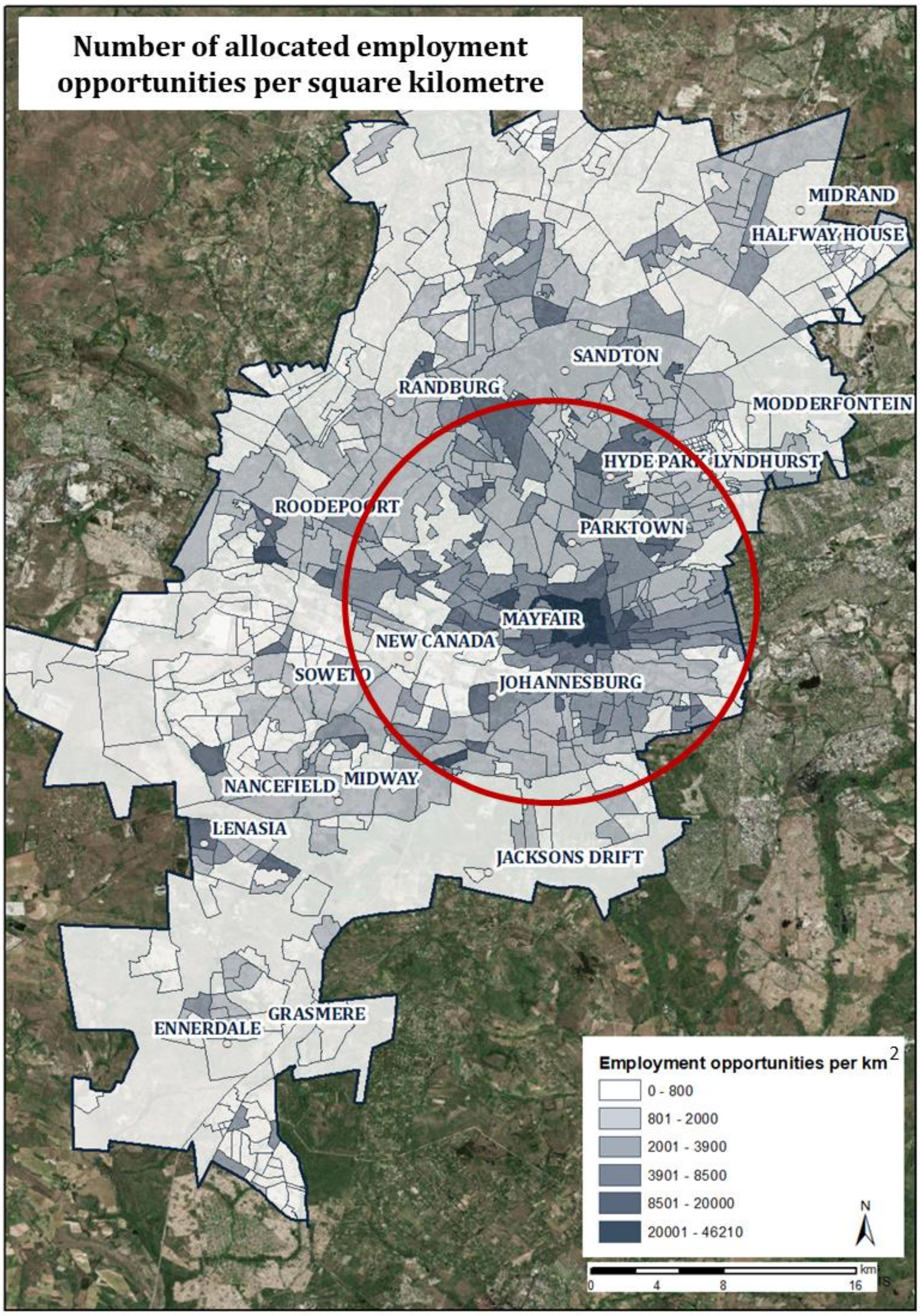


Figure 67: Number of allocated employment opportunities per square kilometre for the City of Johannesburg

Figure 68 depicts the total number of employment opportunities per main place for the CoJ. The Johannesburg main place has the largest number of employment opportunities, followed by Randburg and Sandton. The areas with the lowest number of employment opportunities are located on the fringes of the CoJ. These areas include Zevenfontein, Chartwell, Farmall in the North of CoJ and Lawley, Orange Farm, and Lakeside in the South of CoJ. One of the reasons for the large number of employment opportunities on the Eastern side of the CoJ is because these areas share a border with the areas of high employment in the CoE. Closely related economic activities in these areas are most likely what led to the close proximity of the employment nodes for the two metros.

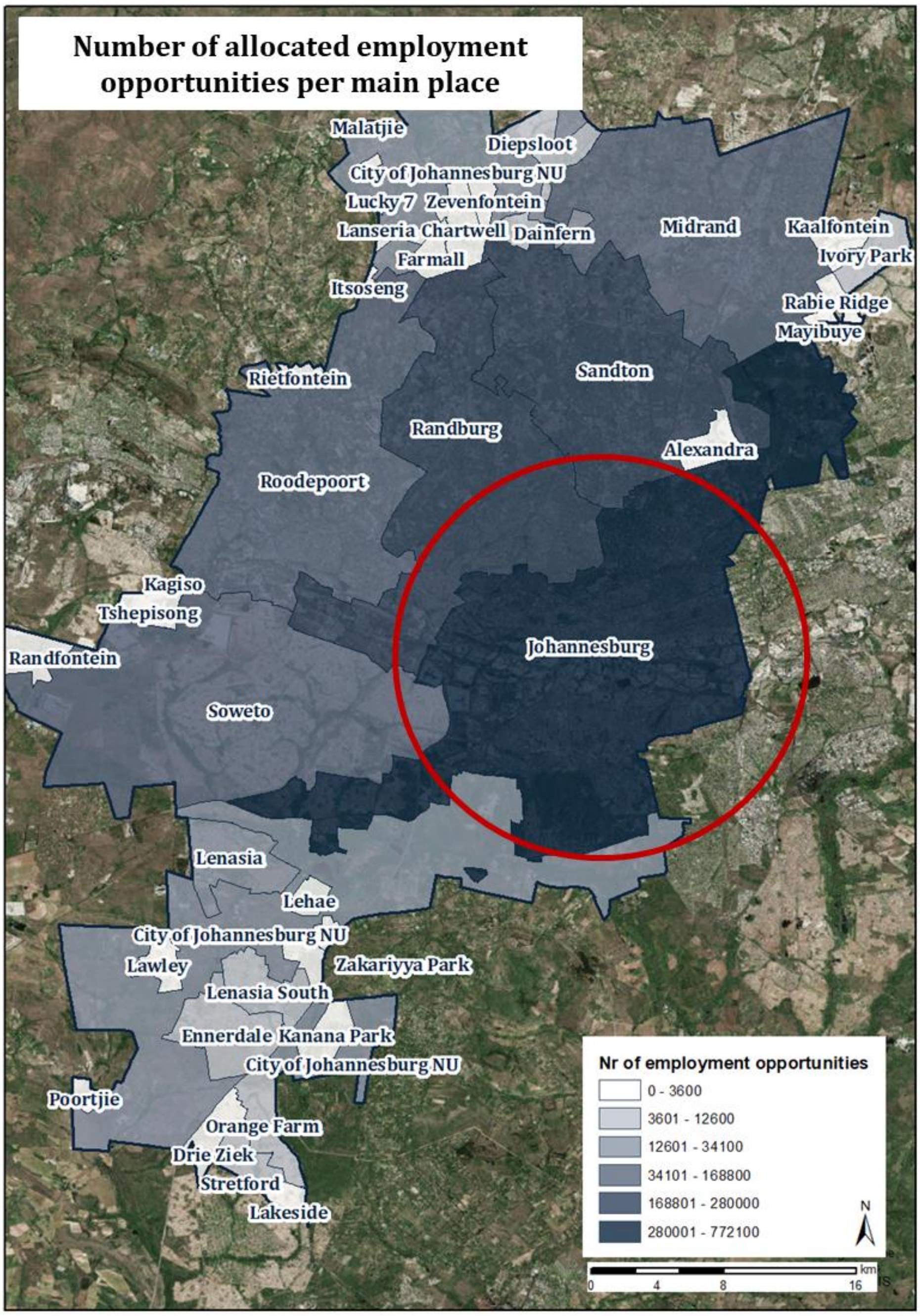


Figure 68: Number of allocated employment opportunities per main place for the City of Johannesburg

Figure 69 illustrates the difference between the actual number of employment opportunities and the allocated number of employment opportunities per metro region. The map on the left is the total number of actual employment opportunities. The map in the middle is the number of allocated employment opportunities. The map on the right shows the difference between the previous two maps. On the difference map, the green shades are indicative of areas where the actual number of employment opportunities were less than the allocated employment opportunities. The blue shades are indicative of areas where the actual number of employment opportunities are more than the number of allocated employment opportunities.

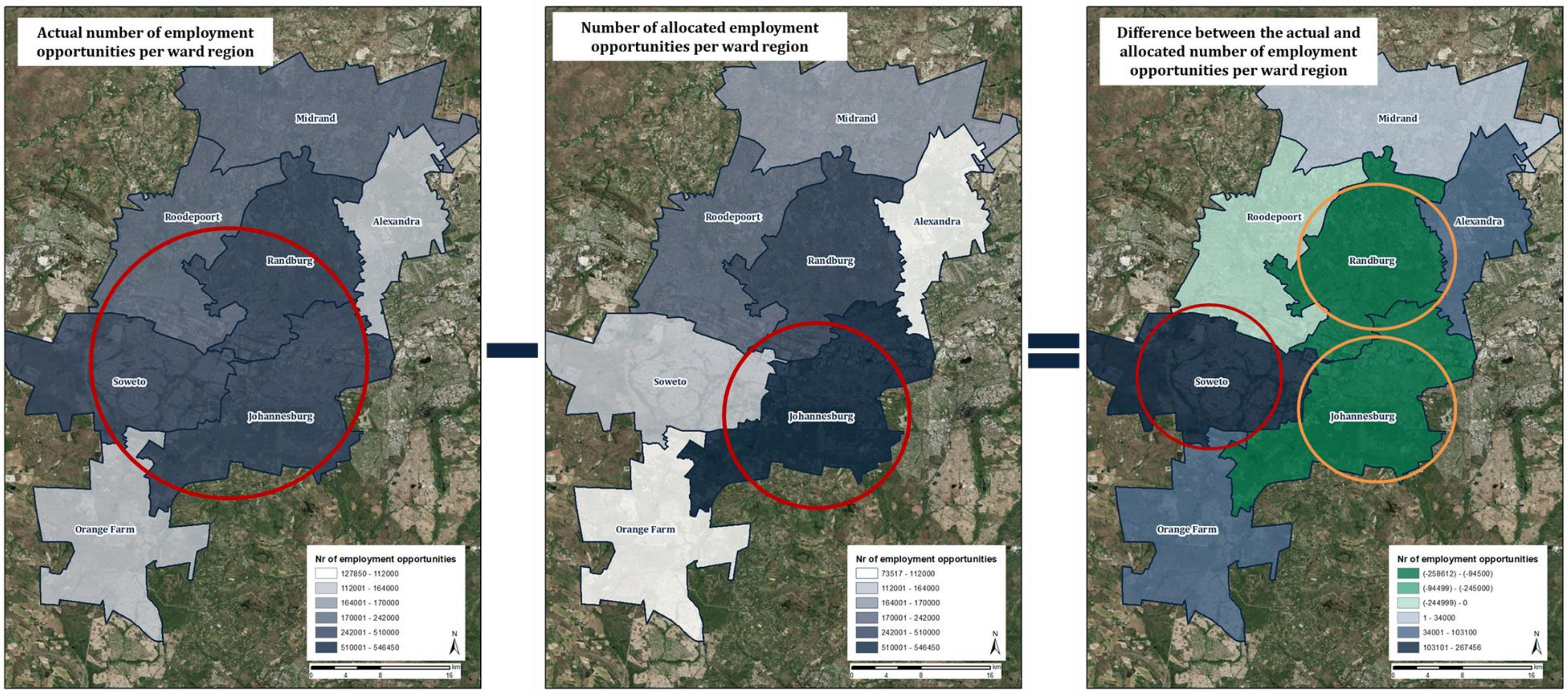


Figure 69: Difference between actual and allocated number of employment opportunities for the City of Johannesburg.

The actual number of employment opportunities map on the left shows that the Johannesburg, Randburg and Soweto metro regions have the highest number of individuals that live in those regions and are employed. The Soweto metro region was the area with the highest density of residential buildings in the density map in Section 6.4. Thus, this high number compared to other metro regions is expected. Both the Johannesburg and Randburg metro regions are larger sized regions in the area and as a result, have a higher capacity or larger area for individuals to live.

The allocated employment opportunities map in the middle show that the Johannesburg metro region has the highest number of employment opportunities followed by the Randburg and Roodepoort metro regions. This is in line with the map depicting the density of non-residential buildings in Section 6.4. In this map, the Johannesburg metro region had areas with the highest density of non-residential buildings, followed by the Randburg and Roodepoort metro regions. Therefore, it is expected that these areas would have the highest number of employment opportunities.

The map with the difference between the actual and allocated employment opportunities on the right illustrates that the largest difference in values was the actual is more than the allocated is in Soweto, followed by Alexandria and Orange farm. This is in line with what is expected as all three of these metro regions are known for their high density suburbs. Therefore, they would have many people who live there and are employed but many of these individuals would travel to different metro regions for work. Johannesburg and Randburg are the two metro regions with the highest negative difference where the actual is less than the allocated. As with the CoE, these two areas are probably the areas where people travel to for work.

6.5.2 Results for City of Tshwane

Figure 70 shows the total number of employment opportunities per sub-place for the CoT. From this map, it can be seen that the larger sub-places to the North-East and South-East of the CoT have the largest number of employment opportunities. The rest of the higher employment areas are located in the central parts of the CoT. The map also shows that the CoT has a few sub-places that are a lot larger in size than many of the other sub-places in the area. This makes it difficult to identify the sub-places in the CoT that form part of the employment hubs in the municipality.

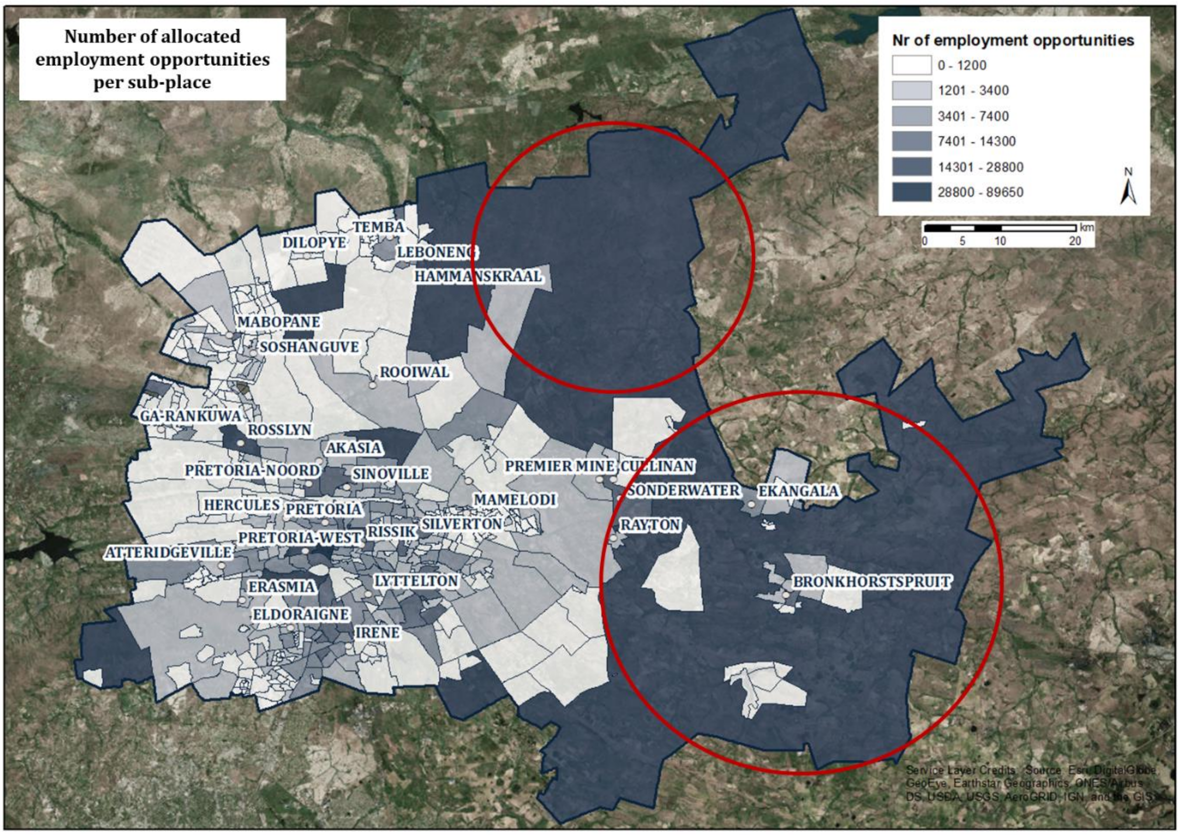


Figure 70: Number of allocated employment opportunities per sub-place for the City of Tshwane

Figure 71 shows the total number of employment opportunities per square kilometres on sub-place level for the CoT. From this map, it can be seen that the highest density in employment opportunities are located in the South-Western part of the CoT. This is where the CBD is located. Surrounding the CBD are other areas that are also the main providers of employment in the metro. It is expected that the CBD and its surrounding areas would have the highest number of employment opportunities.

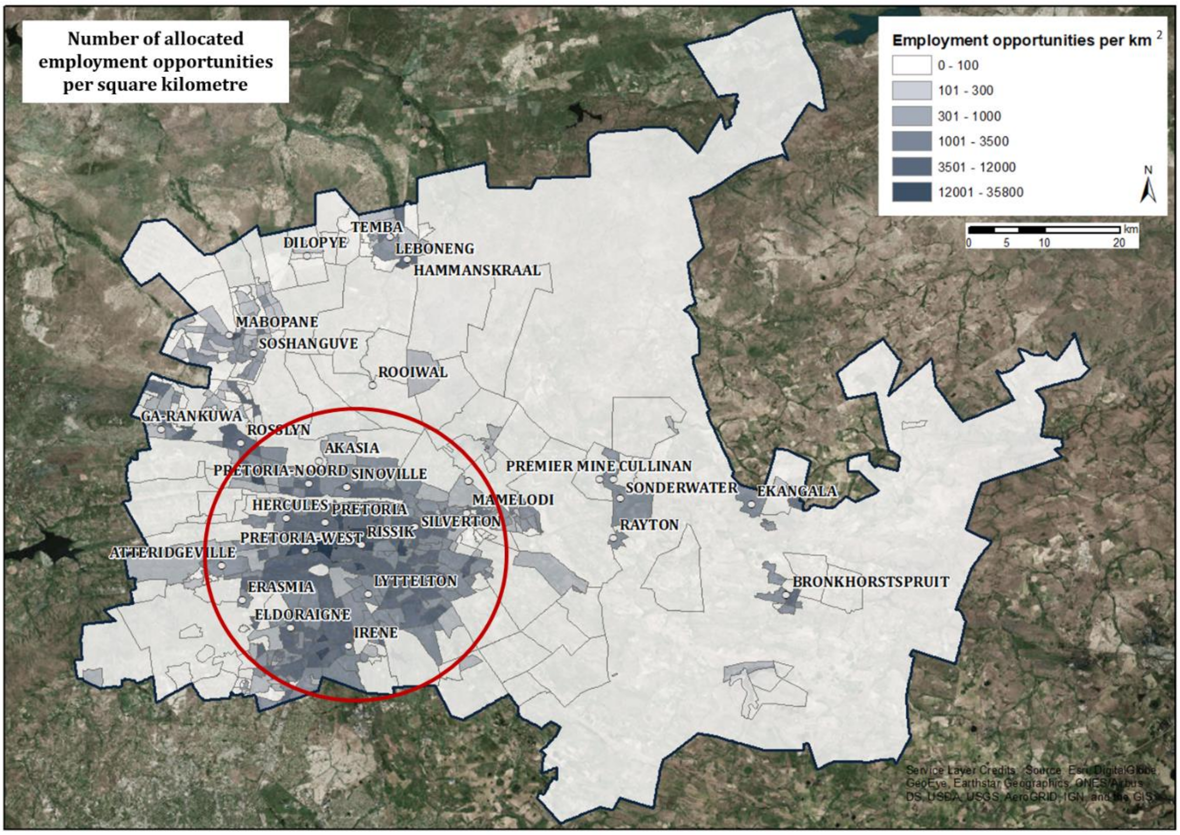


Figure 71: Number of allocated employment opportunities per square kilometre for the City of Tshwane

Figure 72 depicts the total number of employment opportunities per main place for the CoT. The Pretoria main place in the South-West of CoT has the highest number of employment opportunities. It is followed by the Centurion main place as the area with the second-highest number of employment opportunities. In the map of the high density non-residential areas in Section 6.4, these areas also have the highest density of non-residential buildings. Both these areas are known as highly developed areas that provide a large number of employment opportunities in the metro.

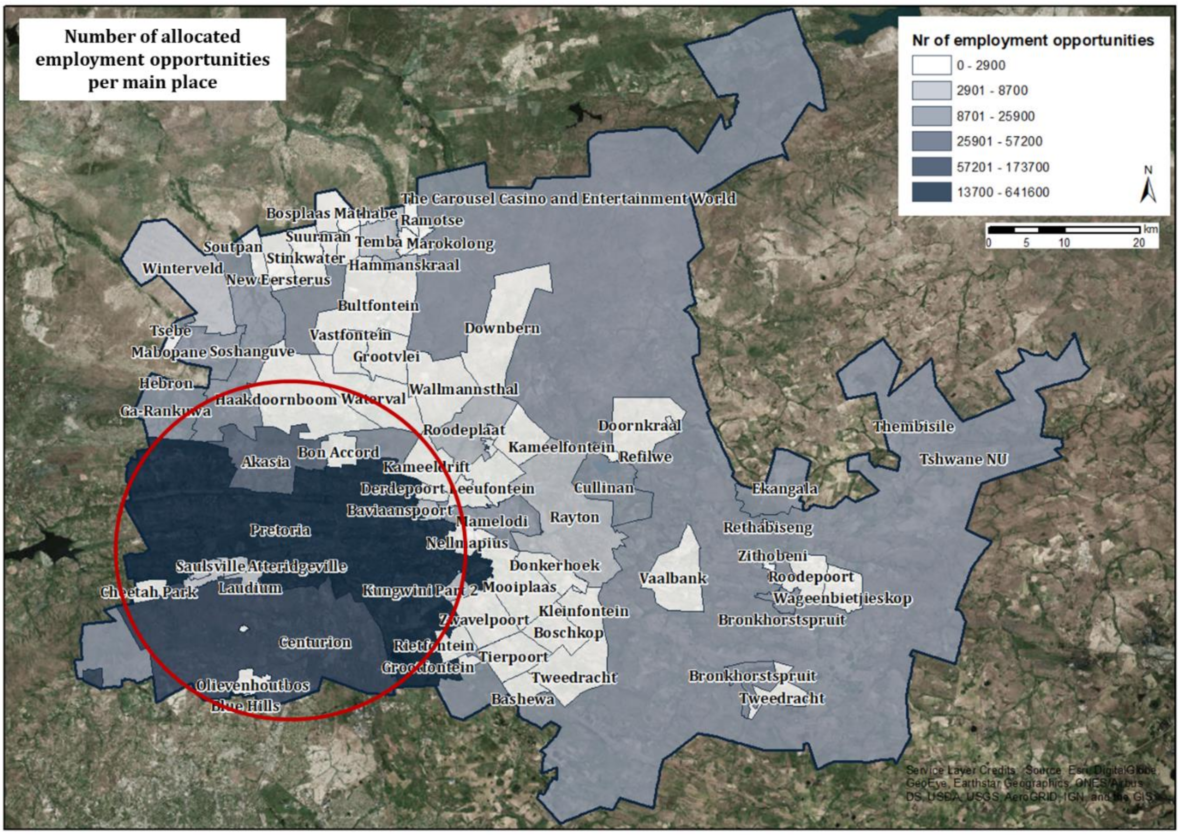


Figure 72: Number of allocated employment opportunities per main place for the City of Tshwane

Figure 73 illustrates the difference between the actual number of employment opportunities and the allocated number of employment opportunities per metro region. The map on the top is the total number of actual employment opportunities. The map in the middle is the number of allocated employment opportunities. The map on the bottom depicts the difference between the previous two maps. On the difference map, the green shades are indicative of areas where the actual number of employment opportunities were less than the allocated employment opportunities. The blue shades are indicative of areas where the actual number of employment opportunities are more than the number of allocated employment opportunities.

The Pretoria, Pretoria East and Akasia/Soshanguve metro regions have the highest number of individuals who live in the area and are employed. This is consistent with the high density residential buildings map in Section 6.4 where these are the three areas with the highest density of residential buildings. These are also areas that are known to have a high population density and as a result that would have a high number of individuals who are employed.

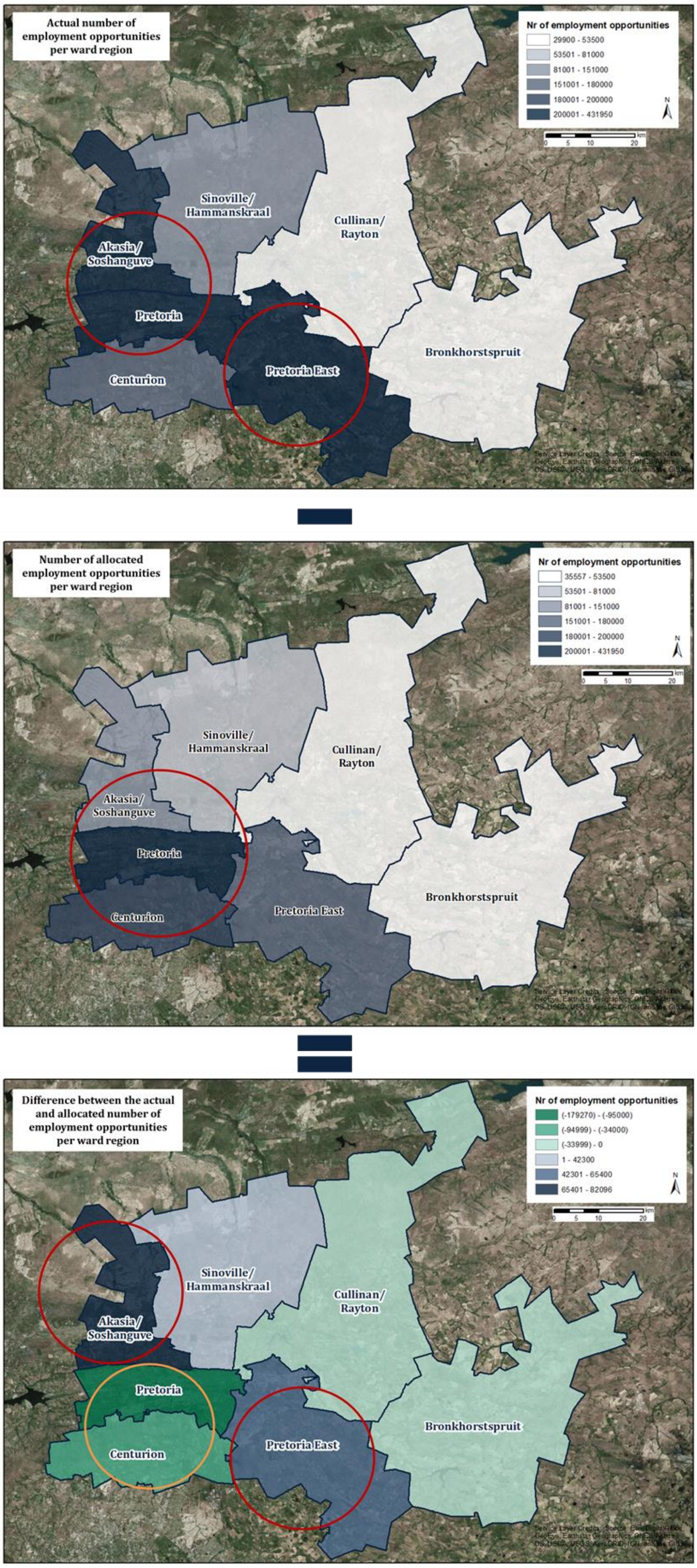


Figure 73: Difference between actual and allocated number of employment opportunities for the City of Tshwane.

The allocated employment opportunities map in the middle shows that the Pretoria metro region, followed by the Centurion metro region and the Pretoria East metro region has the highest number of employment opportunities. This is consistent with the high density non-residential buildings map in Section 6.4. On this map, these three areas have the highest density of non-residential buildings and as a result, have a high capacity for employment. The CoT's CBD is located in the Pretoria metro region and is one of the largest employment hubs in the area, thus the high number of employment in this metro region is plausible. The Centurion and Pretoria East regions share a border with the CoE and the CoJ metros. As mentioned in Section 6.5.1, the closely related economic activities in all these areas are most likely what led to the close proximity of the employment nodes.

The map with the difference between the actual and allocated number of employment opportunities at the bottom indicates that the Pretoria East and Akasia/Soshanguve metro regions have the highest positive difference between the actual and allocated values. This is expected, as both Pretoria East and Akasia/Soshanguve are the two areas with the highest density of residential buildings and lower densities in non-residential buildings. Both these areas are known to be more residential and thus a large amount of the population travel to other areas for employment. The Pretoria metro region, followed by the Centurion metro region are the two regions that have the largest negative difference where the actual number of employment opportunities are less than the allocated number of employment opportunities. Once again, this indicates that these two metro regions are most likely the areas where people travel to for work.

Chapter 7: Conclusion

7.1 Introduction

The final chapter provides an overview of the main results and it offers recommendations on possible future research that can emanate from this research study.

7.2 Main results from the dissertation

The aim of this dissertation was to create a methodology that could be used to disaggregate employment data that is at municipal level to building level. Throughout the dissertation, various results are presented. In the section that follows, each of the objectives of the dissertation is discussed.

Objective 1: Perform a literature review of existing theory and related work on disaggregation methods used in land use change modelling.

The objective was to review any existing methods for disaggregating employment data. This also included methods for disaggregating other data that is similar to the format of the available employment data. The first group of methods that were identified for disaggregating data was mentioned in Chapter 2 during the literature review. Various population synthesis methods were identified that are usually used with disaggregating data with land use change models. These methods, however, could not be applied in the South African context, as a detailed sample of Census data is required for the methods. This sample is not available for employment data in South Africa.

Another factor that needs to be taken into consideration when using the available employment data in South Africa is that the location of employment is captured where a individual lives and not necessarily were they work. Hence, it is required that a higher level of employment data is used for disaggregation. Because of these factors, further research was done for disaggregation methods that did not require a sample dataset and these methods were also presented in Chapter 2. From this second study, a few studies were identified that provided examples of data and techniques that contributed to the methodology of this dissertation.

Objective 2: Create an employment capacity dataset that can be used in the development of the employment allocation algorithm.

The development of the dataset consists of the following processes:

a. Using the literature reviewed as part of Objective 1, identify datasets that can be used for disaggregation.

From the five studies that were reviewed as part of Objective 1, various datasets were identified that could assist with the disaggregation of the employment data. A building dataset formed the basis of the final dataset. Other important datasets included a building footprints dataset, any supplementary datasets that could assist with the calculation of a building’s capacity, and employment totals per economic sector for the entire study area. The supplementary datasets that were implemented with the buildings dataset was a GDE schools dataset and a police station dataset. These two datasets had added information that was used to calculate building capacity.

Data availability is usually a challenge when doing any type of research. This is especially true in South Africa and a lack of data is one of the major factors that contributed to the problem that was solved during this dissertation. This is one of the reasons why a metropolitan municipality was chosen as the study area. As these are larger, more developed areas in the country, there is an improved availability of data compared to most local municipalities. Therefore, many of the datasets used in the dissertation are easily accessible, but some (like the buildings dataset) is not.

b. Develop a procedure that can be used to prepare a dataset (i.e. employment capacity) for a metro in South Africa.

The focus of Objective 2b was to merge all the datasets that were identified as part of Objective 2a to form a base dataset that the algorithm would use for the allocation of the employment opportunities. The main requirement of the base dataset was that each building had to have an employment capacity linked to it. Three elements were required to calculate the final capacity, these included the use of the building (land use), the size of the building, and the economic sector that the employment opportunities in the building are linked to. All three of these elements were then used to create the base dataset where each building had an employment capacity.

Many of the research that was reviewed as part of Objective 2a only considered one factor to disaggregate the data. For this dissertation, various methods were used to determine the employment capacity for the different building types, each customised for the specific building class. This is why supplementary datasets for schools and police stations were added to enhance the accuracy of the employment capacity calculation.

Objective 3: Design and implement the employment allocation algorithm for a metro in South Africa.

The design and implementation would consist of the following sections:

a. Develop the algorithm to allocate employment opportunities to buildings.

Objective 3a was to develop an algorithm that could be used to disaggregate the employment data. The algorithm was an evolutionary algorithm, implemented in Python, using the DEAP library to allocate a certain number of employment opportunities to a building. The allocation was based on the employment capacity of each building and the total number of employment opportunities per economic sector that are available for the municipality.

b. Validate the performance of the algorithm in allocating employment opportunities.

The final part of developing the algorithm was validating the results to determine the accuracy of the algorithm. The algorithm was accurate in terms of the total number of employment opportunities that were allocated for the entire municipality. The algorithm was less accurate when the total number of allocated employment opportunities per economic sector were compared to the actual number of employment opportunities. The allocated employment opportunities were accurate in some of the economic sectors, but over or under allocated in others.

Objective 4: Analyse and discuss the results of the employment allocation algorithm and assess the use of the algorithm in two other metros in South Africa.

Objective 4 inspected the results of the algorithm for the City of Ekurhuleni to see the various ways the data can be used and interpreted. For this objective, maps were created that summed the allocated employment opportunities to different scales or levels. The scales or levels that were used were at sub-place level, main place level and metro region level. The density of employment data at the sub-place level was also calculated. From these maps, it could be seen that the Kempton Park metro region and Germiston/Edenvale metro region had the highest number of employment opportunities out of the eight metro regions in the City of Ekurhuleni.

The second part of Objective 4 was to test the use of the algorithm in other case study areas. Therefore, the algorithm was also applied in the City of Johannesburg and the City of Tshwane to see the adaptability of the algorithm to other study areas. The validation for both areas was similar to that of the City of Ekurhuleni. The algorithm was accurate in determining the total number of employment opportunities for the municipalities but less accurate for the total number of employment opportunities per economic sector.

Figure 74 provides a summary map that shows the number of employment opportunities per metro region for all three of the municipalities. From this map, the metro regions from the three municipalities that contribute the highest number of employment opportunities could be identified. The Johannesburg and Randburg metro regions in the City of Johannesburg and the Pretoria metro region in the City of Tshwane have the highest number of employment opportunities. These regions are followed by the Kempton Park and Germiston/Edenvale metro regions in the City of Ekurhuleni and the Roodepoort metro region in the City of Johannesburg. The map offers a good overview of the employment hotspots in each of the municipalities, as well as for Gauteng as these three municipalities provide the bulk of employment for Gauteng.

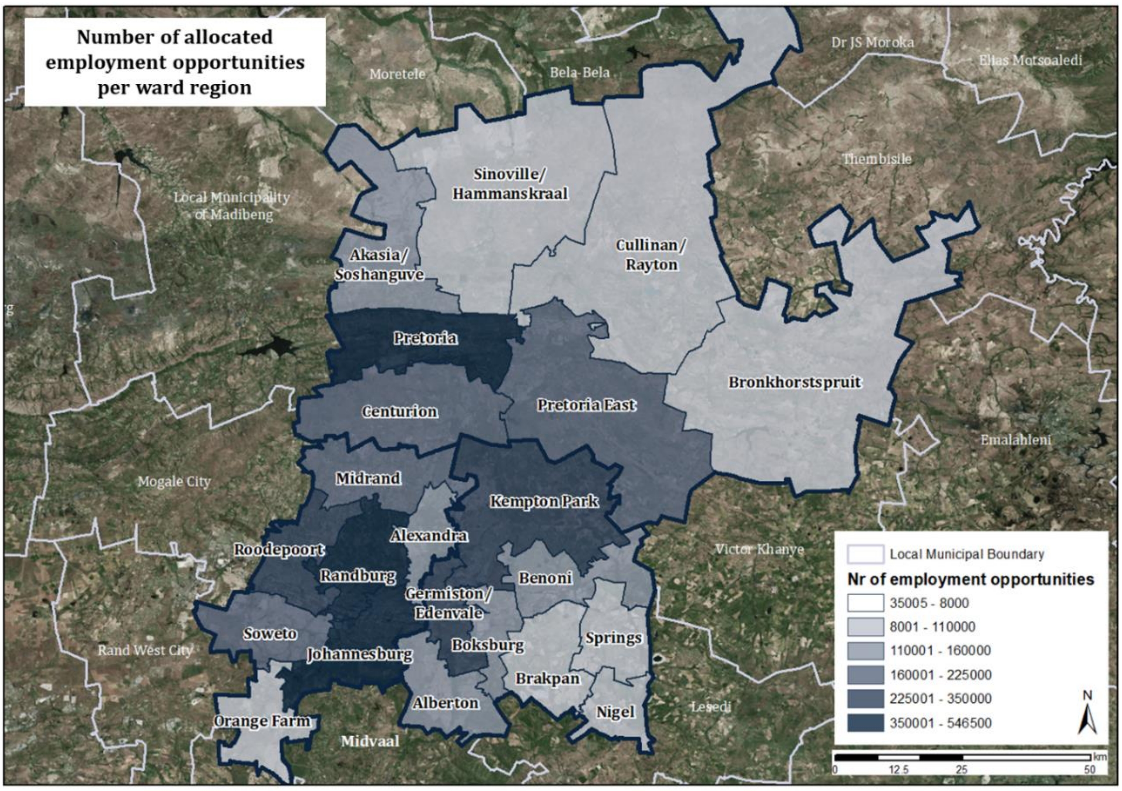


Figure 74: The number of allocated employment opportunities for all three areas

Overall, the results indicate that the employment allocation algorithm was successful in disaggregating employment data from municipal level to building level. All evolutionary algorithms come with some degree of uncertainty as one of the main features of evolutionary algorithms are that they find the most optimal solution, and so there are other solutions available as well. Thus, the results of the algorithm also come with that same level of uncertainty.

When looking at results at a higher level (i.e. metro regions) the results highlight the high employment areas that are expected to form the employment hubs in a municipality. Examples of these areas that are expected to be employment hubs would be the CBD areas within a municipality, highly developed areas with higher densities, and transportation hubs (i.e. airports and harbours). The areas with lower employment numbers are also where they are expected to be, namely areas where the focus is more on residential rather than non-residential.

The algorithm had a similar level of accuracy for the overall employment totals for all three metros. With all three metros, the algorithm allocated fewer employment opportunities than the actual amount of employment. In the CoE example the algorithm under allocated with 0.21%, in the CoJ it under allocated with 0.61% and with the CoT example it under allocated with 0.31%. The algorithm proved to be more accurate for the CoE on sector level than the other two test areas. The largest sector difference for the CoE is 6%, but the largest sector difference for the CoJ is 20% and the CoE is 16%. A solution to this could be to adapt the employment capacity calculations even further to be more specific to each of the metros for those sectors that have a larger difference between the allocated and actual number of employment opportunities.

7.3 Limitations and assumptions

As with most projects, the quality of the data affected the accuracy of the results. As the algorithm was being created, the computational power, as well as the hardware and software used during the research, affected how fast the project moved along. One of the limiting factors of the validation of the algorithm was that the algorithm could only be validated on municipal level and not on a finer scale, because of the nature of how employment data is collected during Census.

The total number of employment opportunities that were used in the dissertation considered a combination of both formal and informal employment opportunities. This can be seen as a limitation as informal employment opportunities cannot always be linked to a building as was done in the dissertation. As the aim of the dissertation was to link employment to buildings, an assumption was made that the percentage of informal employment opportunities in most sectors is not that significant. The sector where it is indeed significant, namely the Wholesale and retail trade, catering and accommodation economic sector, the informal employment opportunities were linked with the informal trading building class.

Another factor that needs to be kept in mind is that the algorithm does not take into consideration the fact that individuals often live in one metro but travel to another for work. This is not taken into consideration because the validation of the algorithm takes place on metro level and not province level. This effect could be reduced if the algorithm is applied to all the municipalities in Gauteng and the employment totals that are used for validation is also for the entire Gauteng. An alternative solution to this problem could be to make use of travel data to determine to where people travel for work. Depending on how detailed the travel information is, it could be used to update the control totals for each metro based on how many people enter and leave the metro for employment opportunities.

7.4 Recommendations for further research

The following areas are recommended for future research:

Linking employment opportunities to households. In the problem statement, it was identified that the relationship between employment opportunities, households and the buildings they are located in could be indicative of several patterns of movement within a city. These patterns could be used to assist in various parts of a land use change model. The relationship between households and buildings is indicative of the spread of households and residential areas across a municipality or city. The various attributes associated with the household allows for even more inferences to be made about an areas demographic composition.

The relationship between employment opportunities and buildings allow inferences to be made about areas with high density employment that could form the employment hubs of a municipality or city. It also allows for the identification of areas where there is a lack of employment and investment into the development of employment opportunities is required. The final relationship, which is between households and employment opportunities, could be indicative of the travel patterns within a municipality or city. This information could in return be used by the transport model to model routes that are the most use or areas where it is important for the development of new transport routes.

Creating this relationship between these three factors is, therefore, a three-part problem. The first part of the problem is linking households to the specific building that they live in. There are already various methods, in the form of population synthesisers, which solve the problem of where specifically households are located. These methods integrate various datasets to create a detailed household dataset. Population synthesisers make use of a sample of detailed Census data, different control attributes and control totals to develop an accurate representation of the households within an area.

The second part of the problem is linking employment opportunities to the buildings that they are located in, which was solved in this dissertation. The final part of the problem is linking the disaggregated employment opportunities to the disaggregated households. Since the relationship between all three these factors need to be known to be beneficial for land use change models, this is a component where further research would be beneficial. A possible way that the link between households and employment opportunities can be established is using a similar approach as was used in this dissertation.

An evolutionary algorithm can be developed that could use travel patterns and distances within a municipality or city to establish the link. It would be beneficial to first link the individual data to employment opportunities and thereafter link the individuals back to households. This is because there could be multiple employment opportunities associated with one household. Linking the individuals to the employment opportunities would remove this one to multiple relationship. An areas demographic attributes that are associated with the individual data could also be used with the travel patterns and distances to establish the link between the individuals and employment opportunities.

References

Abraham, J. E., Stefan, K. J. and Hunt, J. D., 2012. Population Synthesis Using Combinatorial Optimization at Multiple Levels. *Transportation Research Record,* Volume 17.

Abraham, J. E., Weidner T., Gliebe, J., Willison C. and Hunt J., 2005. Three Methods for Synthesizing Baseyear Built Form for use in Integrated Land Use-Transport Models. *Research Record: Journal of the Transportation Research Board,* Issue 1902, pp. 114-123.

Abutaleb, K., Taiwo, O. J., Ahmed, F. and Ngie, A., 2013. *Modeling urban change using cellular automata: the case study of Johannesburg, South Africa.* Johannesburg and Stellenbosch: University of Johannesburg.

Antoni, J., Vuidel, G. and Klein, O., 2017. *Generating a located synthetic population of individuals, households, and dwellings*.

Axhausen, K. W. and Müller, K., 2010. Population synthesis for microsimulation State of the art ETH Library.

Beckman, R. J., Baggerly, K. A. and McKay, M. D., 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice,* 1 11, 30(6), pp. 415-429.

Bishop, Y., Fienberg S., Holland P., Light R., Mosteller F. and Imrey, P., 1977. Discrete Multivariate Analysis: Theory and Practice. *Applied Psychological Measurement Cambridge,* Volume 1.

Black, D. and Henderson, V., 1999. A theory of urban growth. *Journal of political economy,* 107(2), pp. 252-284.

Borning, A., Waddell, P. and Forster, R., 2014. UrbanSim: Using Simulation to Inform Public Deliberation and Decision-Making. pp. 1-24.

Brelsford, C., Lobo, J., Hand, J. and Bettencourt, L., 2017. Heterogeneity and scale of sustainable development in cities. *PNAS,* 114(34), pp. 8963-8968.

Britz, W., Verburg, P. H. and Leip, A., 2011. Modelling of land cover and agricultural change in Europe: Combining the CLUE and CAPRI-Spat approaches. *Agriculture, Ecosystems & Environment,* 7, 142(1-2), pp. 40-50.

Brueckner, J., 2000. Urban sprawl: diagnosis and remedies. *International regional science review,* Volume 23, pp. 160-171.

Cenaero, 2019. *Pareto Front*. [Online] Available at: http://www.cenaero.be/Page.asp?docid=27103&langue=EN

Chaudhuri, G. and Clarke, K. C., 2013. The SLEUTH Land Use Change Model: A Review. *The International Journal of Environmental Resources Research,* 1(1).

Chibba, S., 2016. *South African cities are rapidly growing.* [Online]
Available at: https://www.brandsouthafrica.com/investments-immigration/economynews/south-african-cities-are-rapidly-growing

Choupani, A. and Mamdoohi, A. R., 2016. Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transportation Research Procedia,* Volume 17, pp. 223-233.

De Palma, A., Picard, N. and Motamedi, K., 2014. Chapter 4.2: Application of UrbanSim in Paris (ile-de-france) Case Study. *Ecole Polytechnique.*

DEAP, 2018. *DEAP 1.2.2 documentation.* [Online]
Available at: https://deap.readthedocs.io/en/master/overview.html

DEMACON Market Studies, 2015. *Financial and Projection Modelling of the Ekurhuleni CIF: Task 10 Economic Impact,* Tshwane.

Demographia, 2019. *Demographia World Urban Areas: 2019,* Belleville, IL, USA.

Department of Basic Education, 2013. *Education statistics in South Africa 2011,* Pretoria.

Department of Basic Education, 2015. *Department of Basic Education: District Profiles*, South Africa.

Dovey, K. and King, R., 2011. Forms of informality: Morphology and visibility of informal settlements. Built Environment, 37(1), pp.11-29.

Edmonds, I., 2013. *Urbanization in South Africa.* [Online]
Available at: https://my.vanderbilt.edu/f13afdevfilm/2013/09/urbanization-in-south-africa/

Eiben, A. and Smith, J., 2015. *Introduction to Evolutionary Computing.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Felsenstein, D., Axhausen, K. and Waddell, P., 2010. Land Use-Transportation Modeling with UrbanSim: Experiences and Progress. *Journal of Transport and Land Use,* 3(2).

Fortin, F., De Rainville, F., Gardner, M., Parizeau, M. and Gagne, C., 2012. *DEAP: Evolutionary Algorithms Made Easy*.

Frumkin, H., 2016. Urban Sprawl and Public Health. *Public Health Reports*.

Gallay, O., 2010. Starting with UrbanSim: On the Creation of an Introductory Project. *Infoscience.*

Geard, N., McCaw, J., Dorin A., and McVernon J., 2013. Synthetic Population Dynamics: A Model of Household Demography. *Journal of Artificial Societies and Social Simulation,* 16(1).

Guan, D., Li, H., Inohae, T., Su, W., Nagaie, T. and Kokao, K., 2011. Modeling urban land use change by the integration of cellular automaton and Markov model. *Ecological Modelling,* 222(20), pp. 3761-3772.

Haaland, C. and Van den Bosch, C. K., 2015. Challenges and strategies for urban green-space planning in cities undergoing densification: a review. *Urban Forestry & Urban Greening,* pp. 1-39.

Hevner, A. R., 2007. A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems,* 19(2), pp. 87-92.

Huang, Z. and Williamson, P., 2001. *A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small Area Microdata,* Liverpool.

Huynh, N. N., Barthelemy, J. and Perez, P., 2016. A heuristic combinatorial optimisation approach to synthesising a population for agent-based modelling purposes.. *Journal of Artificial Societies and Social Simulation,* 19(4).

Ioannides, Y. M. and Rossi-Hansberg, E., 2010. Urban Growth, In *Economic Growth,* pp. 264-269. Palgrave Macmillan, London.

Jjin, J. and Lee, H., 2017. Understanding residential location choices: an application of the UrbanSim residential location model on Suwon, Korea. *International Journal of Urban Sciences,* pp. 1-20.

Kakaraparthi, S. K. and Kockelman, K. M., 2011. An Application of UrbanSim to the Austin, Texas Region: Integrated-Model Forecasts for the Year 20. *Journal of Urban Planning and Development,* 137(3), pp. 238-247.

Kim, J. and Lee, S., 2015. A Reproducibility Analysis of Synthetic Population Generation. *Transportation Research Procedia,* Volume 6, pp. 50-63.

Kok, J., Gonzalez, L., Kelson, N. and Periaux, J., 2011. *An FPGA-based Approach to Multi-Objective Evolutionary Algorithm for Multi-Disciplinary Design Optimisation*.

Konduri, K. C., You, D., Garikapati, V. M. and Pendyala, R. M., 2016. Enhanced Synthetic Population Generator That Accommodates Control Variables at Multiple Geographic Resolutions. *Transportation Research Record: Journal of the Transportation Research Board,* 1, Volume 2563, pp. 40-50.

Le Grange, M., 2013. *Ratio of Police Officers to Population.* [Online]
Available at: http://alphasecurity.co.za/ratio-of-police-officers-to-population/

Le Roux, A., Arnold, K., Makhanya, S. and Mans, G., 2019. *South Africa's urban future: Growth projections for 2050..* [Online]
Available at: https://pta-gis-2-web1.csir.co.za/portal/apps/GBCascade/index.html?appid=5180459a765c4e63bfb3fa527c7302b3.

Le Roux, A. and Augustijn, P., 2017. Quantifying the spatial implications of future land use policies in South Africa. *South African Geographical Journal,* 99(1), pp. 29-51.

Lenormand, M. and Deffuant, G., 2012. Generating a synthetic population of individuals in households: Sample-free vs sample-based methods.

Lin, B. B., Meyers, J. and Barnett, G., 2015. Understanding the potential loss and inequities of green space distribution with urban densification. *Urban Forestry & Urban Greening,* 14(4), pp. 952-958.

MathWorks, 2019. *What Is Multiobjective Optimization?.* [Online]
Available at: https://www.mathworks.com/help/gads/what-is-multiobjective-optimization.html#bscm2p2-3

Moeckel, R., Spiekermann, K., Wegener, M. and Wegener, S., 2003. Creating a Synthetic Population. In *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM),* pp. 1-18.

Mohajeri, N., Gudmundsson, A. and Scartezzini, J., 2015. *Expansion and densification of cities: linking urban form to urban ecology.* Lausanne, Switzerland.

Municipalities of South Africa, 2012. *City of Ekurhuleni Metropolitan Municipality - Overview.* [Online]
Available at: https://municipalities.co.za/overview/4/city-of-ekurhuleni-metropolitan-municipality

Municipalities of South Africa, 2019. *Gauteng Municipalities.* [Online]
Available at: https://municipalities.co.za/provinces/view/3/gauteng

Namazi-Rad, M, Mokhtarian, P. and Perez, P., 2014. Generating a Dynamic Synthetic Population – Using an Age-Structured Two-Sex Model for Household Dynamics. *PLoS ONE,* 14 4, 9(4), p. e94761.

Oxford Dictionary, 2018. *urban sprawl | Definition of urban sprawl in US English by Oxford Dictionaries.* [Online]
Available at: https://en.oxforddictionaries.com/definition/us/urban_sprawl

Patterson, Z., Kryvobokov, M., Marchal, F. and Bierlaire, M., 2010. Disaggregate models with aggregate data: Two UrbanSim applications.. *The Journal of Transport and Land use,* 3(2), pp. 5-37.

Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S., 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems,* 24(3), pp. 45-77.

Pritchard, D. R. and Miller, E. J., 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation,* 14 5, 39(3), pp. 685-704.

Python, 2018. *urbansim 3.1.1 : Python Package Index.* [Online]
Available at: https://pypi.python.org/pypi/urbansim/3.1.1

Quantec, 2019. *EasyData by Quantec*. [Online] Available at: www.easydata.co.za/

Ramaswami, A., Russel A., Culligan, P., Rahul Sharma, K. and Kumar, E., 2016. Meta-principles for developing smart, sustainable, and healthy cities. *American Association for the Advancement of Science*, 352(6288), pp 940-943.

Rasmussen, B., Sussman, A., Siddiqui, C. and Hodges, T., 2015. *Scenario Planning for Sustainability and Resilience: Central New Mexico as National Example*.

Schirmer, P., Zollig, C., Muller, K., Rodenmann, B. and Axhausen, K., 2011. The Zurich Case Study of UrbanSim*, Institute for Transport Planning and System*, pp 1-31.

Seto, K. C., Golden, J. S., Alberti, M. and Turner II, B. L., 2017. Sustainability in an urbanizing planet. *PNAS,* 114(34), p. 8935–8938.

Simpson, L. and Tranmer, M., 2005. Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software. *The Professional Geographer,* 5, 57(2), pp. 222-234.

Statista, 2019. *South Africa: Urbanization from 2007-2017.* [Online]
Available at: https://www.statista.com/statistics/455931/urbanization-in-south-africa/

Statistics South Africa, 2012. *Standard industrial classification of all economic activities*.

Talbi, E. and El-Ghazali, 2009. *Metaheuristics: From Design to Implementation (Vol 74)*. John Wiley & Sons

Tanton, R., 2014. A Review of Spatial Microsimulation Methods. *International Journal of Microsimulation,* 7(71), pp. 4-25.

Technical University Munich, 2016. *SILO Introduction.* [Online]
Available at: http://silo.zone/introduction.html

The World Bank, 2019. *Urban population | Data.* [Online]
Available at: https://data.worldbank.org/indicator/SP.URB.TOTL

U.S. EPA, 2000. *Projecting Land-Use Change A Summary of Models for Assessing the Effects of Community Growth and Change on Land-Use Patterns,* Reston, VA.

UN DESA, 2018. *68% of the world population projected to live in urban areas by 2050, says UN | UN DESA | United Nations Department of Economic and Social Affairs.* [Online]
Available at: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

UrbanSim, 2018. *UrbanSim.* [Online]
Available at: http://www.urbansim.com/urbansim/

Vaishnavi, V., Kuechler, B. and Petter, S., 2017. *Design Science Research in Information Systems*.

Van Heerden, Q. and Waldeck, L., 2015. *Integrated land-use and transport modelling using OTP to determine lowest cost trips.* Pretoria.

Verburg, P. H., Andrzej, T. and Erez, H., 2013. Assessing spatial uncertainties of land allocation using a scenario approach and sensitivity analysis: A study for land use in Europe. *Journal of Environmental Management,* Volume 127, pp. 132-144.

Verburg, P. H. and Overmars, K. P., 2009. Combining top-down and bottom-up dynamics in land use modeling: exploring the future of abandoned farmlands in Europe with the Dyna-CLUE model. *Landscape Ecology,* 8 11, 24(9), pp. 1167-1181.

Waddell, P., 1996. *Accessibility and Residential Location: The Interaction of Workplace, Residential Mobility, Tenure, and Location Choices*, pp. 1-24.

Waddell, P., 1998. *An Urban Simulation Model for Integrated Policy Analysis and Planning: Residential Location and Housing Market Components of UrbanSim.* Antwerp, Belgium, pp. 1-22.

Waddell, P., 2002. UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning. *Journal of the American planning association,* 68(3), pp. 297-314.

Waddell, P., 2009. *Parcel-Level Microsimulation of Land Use and Transportation: The Walking Scale of Urban Sustainability*.

Waddell, P. and Borning, A., 2004. A Case Study in Digital Government: Developing and Applying UrbanSim, a System for Simulating Urban Land Use, Transportation, and Environmental Impacts. *Social science computer review,* 22(1), pp. 37-51.

Waddell, P. and Borning, A., 2004. *Developing and Applying UrbanSim, a System for Simulating Urban Land Use, Transportation, and Environmental Impacts,* University of Washington.

Waldeck, L. and Holloway, J., 2016. *Technical Report on Determining the Place of Work for Urban Growth Simulation*.

Waldeck, L. and Van Heerden, Q., 2017. Integrated land-use and transportation modelling in developing countries: using OpenTripPlanner to determine lowest-cost commute trips. *Transportation, Land Use and Integration: Applications in Developing Countries,* Volume 100, pp. 49-68.

Weber, S., 2010. *Design Science Research: Paradigm or Approach? Design Science Research: Paradigm or Approach?.* Lima, Preu, pp. 1-9.

World Bank, 2016. *Urban Population.* [Online]
Available at:
http://data.worldbank.org/indicator/SP.URB.TOTL?end=2015&locations=ZA&name_desc=false&start=1960&view=chart&year=2015

World Bank, 2018. *Urban population | Data.* [Online]
Available at: https://data.worldbank.org/indicator/SP.URB.TOTL

World Economic Forum, 2017. *Migration and Its Impact on Cities*.

World Health Organization, 2015. *Urban population growth.* [Online]
Available at:
http://www.who.int/gho/urban_health/situation_trends/urban_population_growth_text/en/

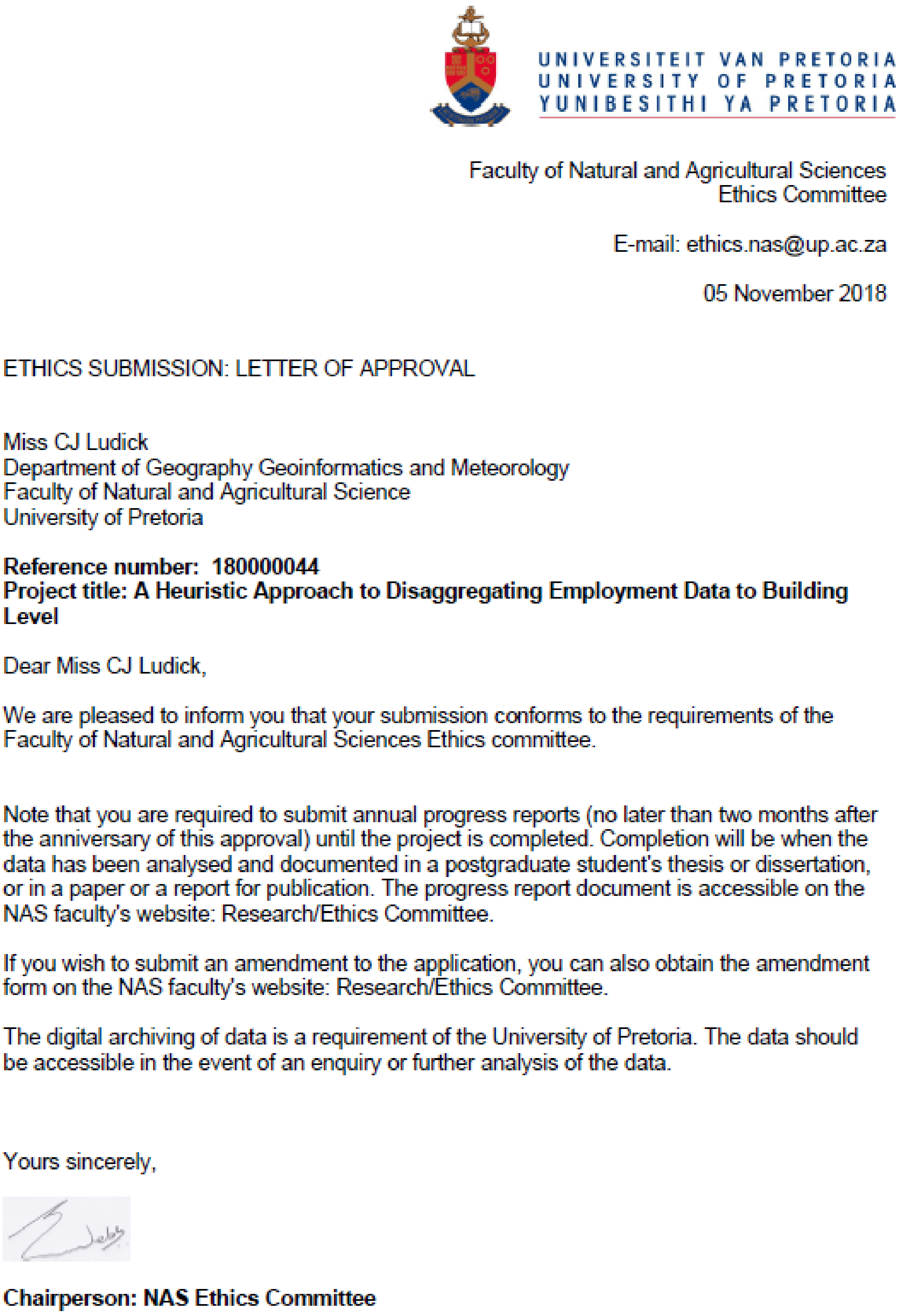
Založnik, M., 2011. *Iterative Proportional Fitting: Theoretical Synthesis and Practical Limitations* (Doctoral dissertation, University of Liverpool).

Appendix A: Acronyms and abbreviations

Table 16: List of abbreviations used

Acronym	Definition
CA	Cellular Automata
CBD	Central business district
CoE	City of Ekurhuleni
CoJ	City of Johannesburg
CoT	City of Tshwane
CSIR	Council for Scientific and Industrial Research
DEAP	Distributed Evolutionary Algorithm in Python
EA	Evolutionary algorithms
GDE	Gauteng Department of Education
HC	Hill climbing
IPF	Iterative proportional fitting
IPU	Iterative proportional updating
LCM	Location Choice Model
MNL	Multinomial Logit model
SR	Synthetic reconstruction
SIC	Standard Industrial Classification
StatsSA	Statistics South Africa

Appendix B: Ethics approval letter



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Natural and Agricultural Sciences
Ethics Committee

E-mail: ethics.nas@up.ac.za

05 November 2018

ETHICS SUBMISSION: LETTER OF APPROVAL

Miss CJ Ludick
Department of Geography Geoinformatics and Meteorology
Faculty of Natural and Agricultural Science
University of Pretoria

Reference number: 180000044
Project title: A Heuristic Approach to Disaggregating Employment Data to Building Level

Dear Miss CJ Ludick,

We are pleased to inform you that your submission conforms to the requirements of the Faculty of Natural and Agricultural Sciences Ethics committee.

Note that you are required to submit annual progress reports (no later than two months after the anniversary of this approval) until the project is completed. Completion will be when the data has been analysed and documented in a postgraduate student's thesis or dissertation, or in a paper or a report for publication. The progress report document is accessible on the NAS faculty's website: Research/Ethics Committee.

If you wish to submit an amendment to the application, you can also obtain the amendment form on the NAS faculty's website: Research/Ethics Committee.

The digital archiving of data is a requirement of the University of Pretoria. The data should be accessible in the event of an enquiry or further analysis of the data.

Yours sincerely,

Chairperson: NAS Ethics Committee