geographical analysis

Geographical Analysis (2022) **0**, n/a

A Multi-objective Optimization Approach for Disaggregating Employment Data

Chantel Ludick[1,2], Quintin van Heerden[2]

[1]Department of Geography, Geomatics and Meteorology, University of Pretoria, Pretoria, South Africa, [2]Council for Scientific and Industrial Research, Pretoria, South Africa

In many countries, including South Africa, data on employment is rarely available on a downscaled level, such as building level, and is only available on less detailed levels, such as municipal level. The aim of this research was to develop a methodology to disaggregate the employment data that is available at an aggregate level to a disaggregate, detailed building level. To achieve this, the methodology consisted of two parts. First, a method was established that could be used to prepare a base data set to be used for disaggregating the employment data. Second, a multiobjective optimization approach was used to allocate the number of employment opportunities within a municipality to building level. The algorithm was developed using an Evolutionary Algorithm framework and applied to a case study in a metropolitan municipality in South Africa. The results showed favorable use of multiobjective optimization to disaggregate employment data to building level. By enhancing the detail of employment data, planners, policy makers, modelers and other users of such data can benefit from understanding employment patterns at a much more detailed level and making improved decisions based on disaggregated data and models.

Introduction

It was estimated that 4.66 billion people globally, which is about 55.7% of the world's population, lived in urban areas by 2019 (Demographia 2019). Furthermore, population projections indicate that approximately 2.5 billion additional people will live in urban areas by 2050, with the majority of the increase being in Asia and Africa. The increase in urban population is not just accounted for by new births, but also the migration of people from rural areas to urban areas in search of new opportunities and improved lifestyles (World Economic Forum 2017; UN DESA 2018; World Bank 2018).

An increase in the urban population raises questions as to what the impact will be on urban areas/cities and their sustainability (Brelsford et al. 2017; Seto et al. 2017). Urbanization could be a key component to creating sustainable cities, but if not managed correctly it could also be

Correspondence: Chantel Ludick, Department of Geography, Geomatics and Meteorology, University of Pretoria, Meiring Naudé Road, Brummeria, Pretoria 0184, South Africa
e-mail: cludick@csir.co.za

Submitted: October 21, 2021. Revised version accepted: April 12, 2022.

doi: 10.1111/gean.12328

a threat to the sustainability of cities. For example, urbanization fosters innovation in terms of technology and sustainability, however, it also contributes to increased pollution, environmental degradation, and inequality. The most prevalent method for managing the increasing urbanization is by making use of development plans and land use policies to allow or restrict development in certain areas. When development plans and land use policies for cities are created, both these positive and negative effects of urbanization need to be considered (Ramaswami et al. 2016; Brelsford et al. 2017; Seto 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 (Guan et al. 2011; Le Roux and Augustijn 2017). Thus, 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 such as UrbanSim (UrbanSim 2021), SLEUTH (UCSB 2021), and CLUE (Environmental Geography 2021) are some of the tools that are currently available to assist in policymaking. Land use change models are computer simulation tools that can be used to understand the causes, mechanisms and significance of urban growth as well as provide a user with the capability to study how land use systems function and what effects they could have (Abutaleb et al. 2013; Chaudhuri and Clarke 2013; Le Roux and Augustijn 2017).

Land use change models rely on a variety of data sets as inputs to perform simulations that produce projections on how an urban area will develop or change. These input data sets range from information about households, to employment, transport, land use, and development plans. For the simulations to be as accurate as possible, the input data should be of high quality and detailed. Currently, in South Africa, there is a lack of detailed data on place of work and how these employment opportunities are linked to where individuals stay. The lowest level at which employment data is provided in South Africa is at the sub-place level (i.e., demarcation by Statistics SA that is more or less the size of a neighborhood) and many models require data sets to be at a much more disaggregated and detailed level (e.g., agent-based models require data to be at building level).

To address this gap, this article offers a new employment disaggregation algorithm that can be used to disaggregate employment data that is available at a more aggregate level to a more detailed, building level. This article outlines not only the development of the algorithm for a case study area, but also the preprocessing that went into developing the initial data set from which the algorithm is developed. Therefore, the article is organized as follows. Data and study area section introduces the case study area and the data sets used, as well as the preprocessing element for developing the initial or base data set. Employment allocation algorithm section describes the development process of the employment allocation algorithm. Discussion of results section presents the discussion of the results from the application of the algorithm to a case study area. Finally, Conclusion section provides concluding remarks.

Data and study area

Case study area

The City of Ekurhuleni is a metropolitan municipality located in Gauteng, South Africa and is one of only eight metropolitan municipalities in the country (Municipalities of South Africa 2019). Fig. 1 illustrates the location of the Ekurhuleni within the Gauteng province. The municipality covers a geographical area of 1,975 km^2 and consists of 101 wards. In 2019, the estimated

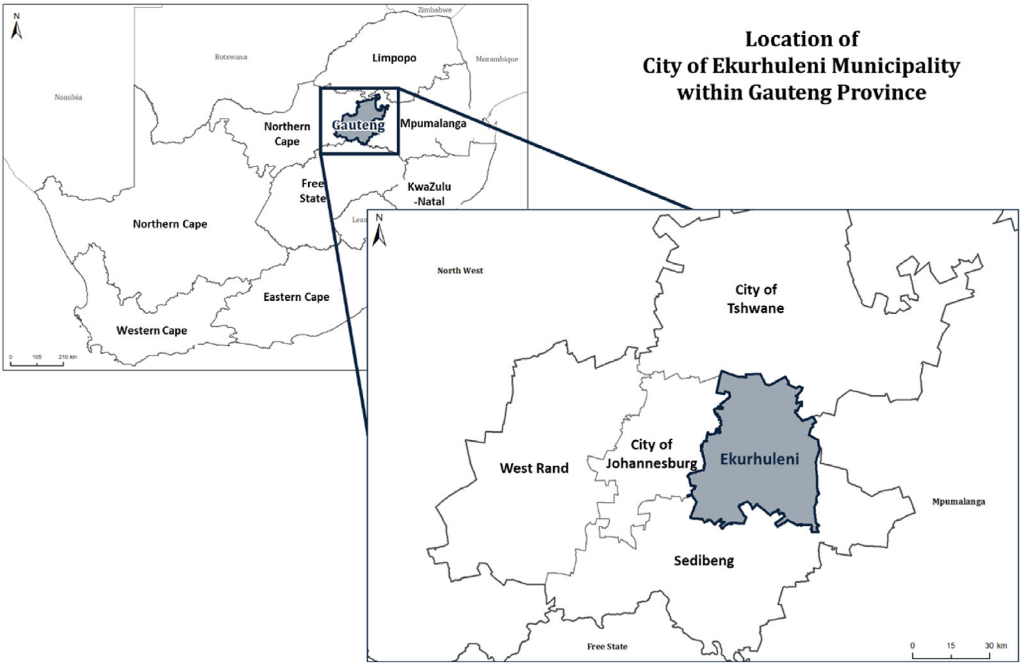


Figure 1. Location of City of Ekurhuleni within Gauteng Province. [Colour figure can be viewed at wileyonlinelibrary.com]

population was 3,649,053 people with a population density of 579 people per square kilometer and an estimated 1,299,449 people were employed (Quantec, 2020). It is projected that the population of Ekurhuleni will increase to 1.7 million people by 2050, which is a 60% projected growth making Ekurhuleni the third fastest growing municipality in Gauteng (Le Roux et al. 2019).

Ekurhuleni has a very diverse economy that contributes to almost a quarter of Gauteng's economy. It is known as *Africa's Workshop* (Cooperative Governance and Traditional Affairs 2020), since many of the country's goods and commodity factories are located here. O.R. Tambo International Airport (i.e., busiest airport in Africa) and the largest railway hub in South Africa is located in the municipality cementing its status as a transportation hub. When considering all these factors, it is clear that Ekurhuleni is a major employment hub in the province and country (Municipalities of South Africa 2019), which is one of the reasons why it was selected as the study area to test the methodology.

The many job opportunities (in various sectors) make it an ideal study area to develop and test the methodology for disaggregating employment data. Another reason for choosing Ekurhuleni as the study area was the fact that it is a metropolitan municipality, and more resources are available for the municipality as more research has been done using metropolitan municipalities rather than smaller municipalities, which leads to a range of data sets being available to use in the development of the solution methodology.

Related work and overview of data used

Five existing disaggregation methods were reviewed in order to identify possible data sets that could be used to create a base data set. These methods make use of algorithms and examine disaggregation methods that could be more applicable to the South African context. Abraham et al. (2005) identified a problem in the accuracy of data on floor space inventory in three land use transportation modeling projects. The floor space data was incomplete, and it showed inconsistencies with both population and employment data. The research included three different methods that could be used to synthesize a built form input data set at a micro level that could be used by land use transportation models. An algorithm was developed for each of the three land use transportation projects. The first algorithm was the Oregon approach, which included an analysis of the demand and supply floor space on both the macro and micro levels. From this, an algorithm was developed to assign development types as well as the quantity of space. The second algorithm was the Sacramento approach, which resulted in the assignment of the total inventory of floor space to suitable parcels. The last algorithm was the Oahu approach, which resulted in jobs assigned to certain grid cells as well as synthesized floor space in grid cells. This approach by Abraham et al. (2005) shows the advantage in using floor space when disaggregating data.

Patterson et al. (2010) considered 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 disaggregated. As with the Brussels case, there was also no buildings data available. In the Brussels case study, the data was disaggregated from zonal level to grid cell level. 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 and the process not being efficient in the end.

In the Lyon case study, similar steps to the Brussels case were 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.

Huynh, Barthelemy, and Perez (2016) developed a heuristic approach to synthesize 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 associated household. 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 specifically use employment data but some of the techniques could be used in disaggregating employment data (Huynh, Barthelemy, and Perez 2016).

Antoni, Vuidel, and Klein (2017) made use of the MobiSim Population Synthesizer to synthesize the population, as well as assign the population to buildings for three cities in France. The method is divided into two steps. First, the population that is linked to households is generated and second, this population is linked to the buildings. To create the household data set, 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, Vuidel, and Klein 2017). The capacity of the building along with the type of building is used to link a building to a household. The result is a data set where each individual is linked to a household and each household is linked to a building and each building is linked to a specific geographic area. Once again, the approach does not specifically involve employment data, but the approach 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.

Waldeck and Holloway (2016) also identified the problem that there is no data available in South Africa on work locations. Therefore, they developed a method that could be used to create a data set of where individuals are employed in South Africa. This method made use of cadaster 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 cadaster. They then disaggregated the number of jobs proportionally by using the valuation of the cadaster divided by the total valuation of the cadasters 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. The process was done manually and as was the case in Patterson et al. (2010), it was not very efficient. Some components from the study could be used, such as the validation process, which could possibly be implemented for this study. The factors considered in the disaggregation process could also be considered. The five studies are summarized in Table 1 and from this, possible data sets that could be used for disaggregating employment data were identified.

This information was taken into account in the development of the base data set that was used as the starting point for the employment allocation. The base data set comprised of employment capacities for each building, which were later used to allocate employment opportunities. From the review, the following four requirements were identified to enable employment allocation to specific buildings: (1) Location and use of the buildings; (2) Size (i.e., area) of the buildings; (3) Economic sector that the employment opportunities are linked to; and (4) Employment capacity of the buildings. The base data set was created accordingly using the data sets in Table 2. The sections that follow, discuss the process of creating the base data set in more detail.

Buildings data set

A building data set containing all buildings in Ekurhuleni was obtained. The 2012 version of the data was specifically selected as the employment data from the 2011 Census was used. The 2011 Census data was used because this research is an essential element of a larger urban modeling research project and 2011 is the base year for the urban growth simulation model. The buildings data set that was used, contained the location of the buildings and the underlying land use. Each building had a primary class that broadly indicated its use and a secondary class that provided more detail. The buildings for five secondary classes were removed and replaced by ancillary data sets as they provided more information that would allow a more accurate calculation of the employment capacity. Table 3 provides an overview of the primary classes and ancillary data sets used.

Table 1. Summary of Existing Methods

Authors	Research topic	Data used
Abraham et al. (2005)	Three Methods for Synthesizing Baseyear Built Form for Use in Integrated Land Use-Transport Models	Floor space Population totals Employment totals
Patterson et al. (2010)	Disaggregate Models With Aggregate Data: Two UrbanSim Applications	Building proxies Household totals Employment totals
Huynh, Barthelemy, and Perez (2016)	A Heuristic Combinatorial optimisation Approach to Synthesising a Population for Agent-based Modelling Purposes	Population demographic data Household demographic data
Antoni, Vuidel, and Klein (2017)	Generating a Located Synthetic Population of Individuals, Households, and Dwellings	Population demographic data Household demographic data Building data
Waldeck and Holloway (2016)	Determining the Place of Work for Urban Growth Simulation	Building data Building valuation Employment totals

Fig. 2 shows an example of the buildings data set, classified according to the primary land use class.

Building floor space

The floor space of the buildings had to be determined from the building footprints data set. The building footprints data set contained the area and approximate number of floors of a building. Fig. 3 shows a sample of the building footprint data set.

The first step of calculating the floor space was to join the footprint data to the buildings. There were situations where more than one building intersected with the footprint boundary. To resolve this, the area of the footprint was divided by the number of buildings intersecting with the footprint. The average building footprint size was calculated for each sub-place and joined to the buildings that did not intersect with any building footprints. The buildings with intersecting footprints were assigned the average size of the buildings that fell within its sub-place. The floor space was calculated by multiplying the area of the building by the number of floors of the building. The number of floors was provided as a range (e.g., 1–3). For the purpose of this article, the maximum value in the range was used to calculate the floor space of the building, resulting in the desired maximum capacity of each building.

Economic sectors

In South Africa, there are 11 main economic sectors according to the Standard Industrial Classification (SIC) of all economic activities (Statistics South Africa 2012). To determine

Table 2. Data Sets Used for the City of Ekurhuleni

Data set	Data custodian	License details
City of Ekurhuleni residential buildings	Council for Scientific and Industrial Research (CSIR)	Licensed
City of Ekurhuleni nonresidential buildings	CSIR	Licensed
City of Ekurhuleni building footprints	CSIR	Licensed
Sub place boundaries for the City of Ekurhuleni	StatsSA	Freely available
Main place boundaries for the City of Ekurhuleni	StatsSA	Freely available
Planning regions in the City of Ekurhuleni	StatsSA	Freely available
Local municipal boundary for the City of Ekurhuleni	StatsSA	Freely available
Primary and high schools in the City of Ekurhuleni	Gauteng Department of Education (GDE)	Government owned
Police stations in City of Ekurhuleni	South African Police Service	Government owned
Employment per economic sector for the City of Ekurhuleni	Quantec	Licensed

Table 3. Various Primary Classes Within the Building Data Set

Primary class	Ancillary data sets used
Mining	
Transport	
Utilities and infrastructure	
Health care	
Education	*Secondary classes removed*: Preschools, primary schools, secondary school, and other schools *Replaced with*: Gauteng Department of Education schools that included the number of classrooms and learners, and the location of the school
Commercial	
Industrial	
Recreation and leisure	
Tourism	
Institutions	*Secondary classes removed*: Police buildings *Replaced with*: The location of each of the police stations and the size of the population the station serves
Residential	

Figure 2. Sample of buildings data set. [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 3. Building footprints. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 4. Economic Sectors Linked to Building Classes

Primary class	Economic sector
Mining	Mining and quarrying [SIC2]
Transport	Transport, storage, and communication [SIC7]
Utilities and infrastructure	Community, social, and personal services [SIC10]
	Electricity, gas, and water supply [SIC4]
	Financial, insurance, real estate, and business services [SIC8]
	Transport, storage, and communication [SIC7]
Health care facilities	Community, social, and personal services [SIC10]
Education	Community, social, and personal services [SIC10]
Commercial	Financial, insurance, real estate, and business services [SIC8]
	Wholesale and retail trade, repairs, hotels, and restaurant [SIC6]
Industrial	Manufacturing [SIC3]
Recreation and leisure	Community, social, and personal services [SIC10]
Tourism	Wholesale and retail trade, repairs, hotels, and restaurant [SIC6]
Institutions	Community, social, and personal services [SIC10]
	Financial, insurance, real estate, and business services [SIC8]
	General government [SIC9]
Residential	Agriculture, hunting, forestry, and fishing [SIC1]
	Community, social, and personal services [SIC10]
	Financial, insurance, real estate, and business services [SIC8]
	Wholesale and retail trade, repairs, hotels, and restaurant [SIC6]

which sector an employment opportunity is related to, each of the primary classes was linked to a relevant economic sector (refer to Table 4).

The AGRICULTURE, HUNTING, FORESTRY, FISHERIES economic sector was linked to the residential building class as buildings classified under agriculture only include major outbuildings (e.g., barns, sheds, and other agricultural buildings) on farms and smallholdings. These buildings are commonly used for storage and were thus not linked to employment opportunities. The CONSTRUCTION economic sector was omitted from the study as this is a sector that commonly only offers temporary employment opportunities. It was also not possible to link employment opportunities for the construction sector to specific buildings, since construction workers do not work in buildings but on construction sites and the main aim of this research was to link employment opportunities to the buildings where employees work. The current methods used to calculate the employment capacity, discussed in the next section, was also not applicable for the way in which the construction sector works.

Table 5. Method Used to Calculate Employment Capacity

Method	Primary building classes
Proportionally divided economic sector employment total	Commercial, education, health care facilities, institutions, mining, recreation and leisure, residential, tourism, transport, utilities and infrastructure
Used ancillary data	Education and institutions
Used estimated area per employee	Commercial and industrial

Calculating employment capacity

The employment capacity was calculated using the combined data set (described in the section Related work and overview of data used) that contained the building footprints with usage (i.e., land use class), floor space, and the economic sector it was related to. This was the final step for creating the base data set that was used in the development of the algorithm. To calculate the employment capacity, three main methods were used as are summarized in Table 5 and discussed next.

The first method was to proportionally divide the total employment of the economic sector that a building class was linked to, between all the buildings that were associated with that economic class by making use of the building floor space. Hence, when the employment totals were divided by the floor space, larger-sized building would have a higher capacity than smaller buildings. The second method used an estimated area per employee to calculate the employment capacity. The area per employee was calculated by using plans for new developments within the City of Ekurhuleni that outlined 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). From this, an estimated area per employee was determined and the amount of floor space was divided by these values to calculate the final employment capacity.

The third method consisted if using the ancillary data added in Buildings data set section to calculate the employment capacity. This method was specifically applied for the police points where the employment capacity was calculated using a standard ratio of the number of police officials that are required for a certain population size in Gauteng (Le Grange, 2013). It was also used for some of the education points where the number of classrooms and number of learners were used to determine the capacity of the schools. Standard ratios were used for the number of educators to classroom (Department of Basic Education, 2015), as well as the number of administrative staff to learners (Department of Basic Education, 2013).

The RESIDENTIAL buildings data set already had a precalculated 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. This capacity was determined as part of the larger urban modeling project that this research forms part of and was based on 2011 Census information on home-based jobs, which was readily available.

Employment allocation algorithm

After reviewing the previous studies in Related work and overview of data used section, it could be seen that the use of an algorithm to allocate information is more efficient than to manually allocate the data. Therefore, more studies were looked at to identify types of algorithms that can be best used to solve resources allocation problems. Chen, Fu, and Lim (2002), Lee et al. (2003), Osman, Abo-Sinna, and Mousa (2005), and Kumar et al. (2010) all looked at various resource allocation problems. In its essence, the problem being solved in this study is a resource allocation problem, where the employment opportunities are the resources that needs to be allocated to the buildings. All five of the studies identified Genetic/Evolutionary Algorithms as a very effective method to solve resource allocation problems, especially where there is a multiobjective problem to solve. Therefore, in this section a short discussion on evolutionary algorithms is provided and then the components for developing the *employment allocation algorithm* are presented.

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 Evolutionary/Genetic Algorithms, Differential Evolution, Swarm Intelligence, and Cultural Algorithms, to name a few (Eiben and 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 (Eiben and Smith 2015). An EA initially begins with a population that consists 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 customized 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 and Smith 2015). Fig. 4 illustrates the process that most EAs follow.

During this process, the solutions that are best fit to solve a particular problem will survive and grow and those less-than-optimal solutions will be removed. The process continues until

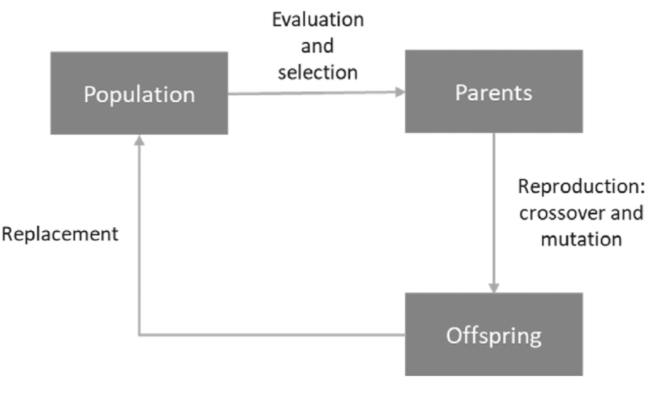


Figure 4. A generation in evolutionary algorithms (Talbi 2009).

Employment allocation algorithm

```
assign number of generations to create (t)
assign number of iteration to 0 (i = 0)
initialise individual
evaluate fitness of individuals in population
while (i < t) do
    i = i + 1
    crossover individuals
    mutate individual
    evaluate the fitness of individuals
    select the best individuals
end while
return best individual
```

Figure 5. Evolutionary algorithm template.

some termination criteria are met. Fig. 5 shows a template for the process that an evolutionary algorithm follows.

From Fig. 5, it can be seen that most evolutionary algorithms consist of four main components: (1) Initializing the population; (2) Evaluating the fitness; (3) Specifying the crossover and mutation rules; (4) Selection of the best offspring. The sections that follow explains how each of the four components were implemented for the employment allocation algorithm. The Distributed Evolutionary Algorithm in Python (DEAP) (DEAP 2018) was used as the framework to build the employment allocation algorithm and implement each of the four components. DEAP is an open-source evolutionary algorithm framework that is developed in Python and enables users to rapidly create prototypes by allowing them to use the framework to create their own custom evolutionary algorithms to suit their needs (Fortin et al. 2012; DEAP 2018). The main reasons that the DEAP framework was chosen as the basis for the employment allocation algorithm was because of the multiobjective functionality, the customization capabilities, the ease of use, and the built-in toolbox.

Initializing the population

The initial population is made up of a number of individuals, where each individual represents a tentative solution (Eiben and Smith 2015). Before the initial population can be created, the type of individuals of the population first needs to be determined. A list of integers was identified as the most appropriate format to use with the unique building ID used as the index of the list and number of allocated employment opportunities as the values at each position in the list. Fig. 6 provides an extract of the base data set created in the section Data and study area. Fig. 7 provides an example of the list format used to allocate the employment opportunities.

To create the individuals of the initial population, each building was allocated a certain proportion of its employment capacity. The process that was used to generate the initial population was updated a number of times 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.

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 6. Extract from base data set.

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

Figure 7. 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 8. Example of proportion of job capacity being assigned to create the initial population. [Colour figure can be viewed at wileyonlinelibrary.com]

A randomly generated proportion was used to allocate the initial employment opportunities by multiplying the random value with the capacity to allocate several employment opportunities to each building. A random value was generated for each of the economic sectors. For each economic sector, the generated value was then multiplied with the capacity of the buildings in the specific economic sector. Fig. 8 illustrates how a proportion of the job capacity was allocated for the initial population and how that was subsequently transferred to an individual.

Evaluate the fitness

The first step of creating the fitness functions is to decide what variables could be used to evaluate the fitness of an individual. For the employment allocation, two measures of fitness were identified, thus a multiobjective fitness function was used. The fitness functions were developed based on the data that was available to measure the fitness of a solution. The first fitness function, or *Objective 1*, considered the number of employment opportunities allocated

	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 9. Example of the first objective calculation.

per economic sector at municipal level. These values were compared to the actual number of employment opportunities per economic sector based on the 2011 national Census data. 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 and this value was then used for the first objective. Fig. 9 provides an example of the calculation of the fitness value of the first objective for one individual.

The second fitness function, or *Objective 2*, considered the difference between the number of allocated employment opportunities per economic sector and the employment capacity of each economic sector. This fitness function was used to ensure that the algorithm does not allocate more employment opportunities than a building had the capacity for. It also ensured that it did 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 were summarized to 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 these values could differ a lot based on the different capacity sizes for each economic sector, the percentage difference was rather calculated and subsequently used to assign a “penalty” score. Negative percentages received the largest penalty, since that indicated that more employment opportunities were allocated than the capacity allowed for. Percentages between 0 and 60 received a lower score of 10, as this was a tolerable difference. Percentages above 60 again received a larger penalty of 100 as this meant that the allocated employment opportunities were a lot less than the capacity. The score for each economic sector was finally summarized to provide the final value for the second objective. Fig. 10 provides an example of the calculation of the fitness value of the second objective for one individual.

In evolutionary algorithms, the fitness function can be either minimized or maximized by using a weight and the sign of the weight determines which function it is (Fortin et al. 2012). For

	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 10. Example of the first objective calculation.

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

Specify the crossover and mutation rules

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. 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 (Eiben and Smith 2015). Throughout the process of developing the algorithm, many crossover and mutation operators were tested.

For crossover, a *One-point crossover* and a *Two-point crossover* were tested with the algorithm. 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. *Two-point crossover* works the same, but instead of just one portion of the individual being swapped, two portions of the individuals are swapped (Fortin et al. 2012). The *Two-point crossover* was chosen and applied in the end as it provided the most variation between the generation and therefore provided the best results. Fig. 11 provides an example of two-point 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 that the offspring is not identical to the parent, and it is an important step to ensure that the algorithm does not become stuck in local extrema (Eiben and Smith 2015). Some of the mutation operators that were tested included a Gaussian mutation and a shuffle index mutation. The Gaussian mutation randomly selects a position within the individual and adds a Gaussian

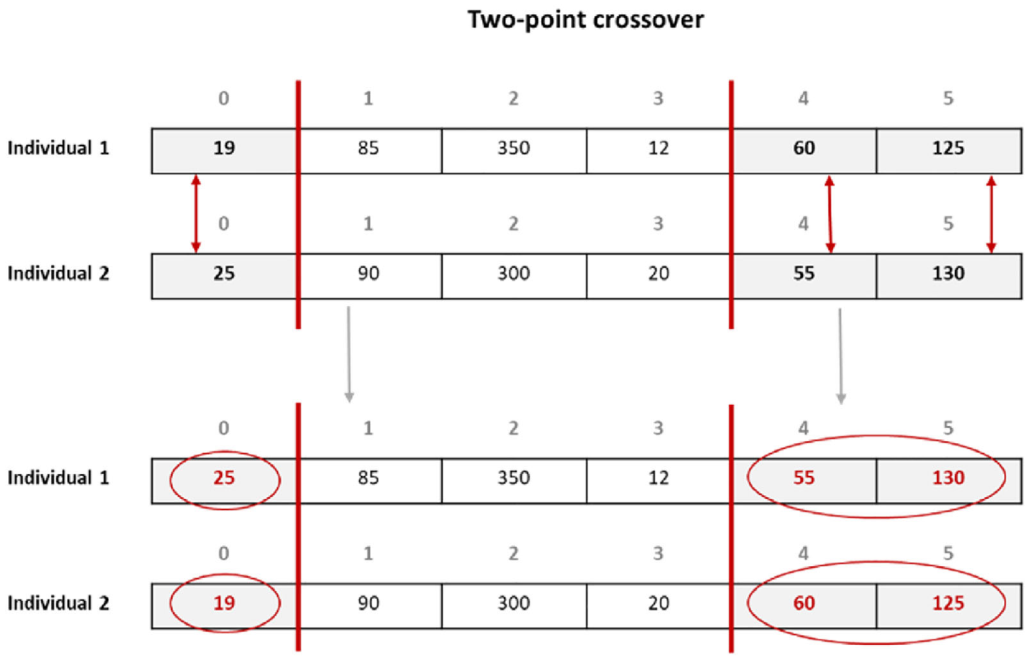


Figure 11. Example of two-point crossover. [Colour figure can be viewed at wileyonlinelibrary.com]

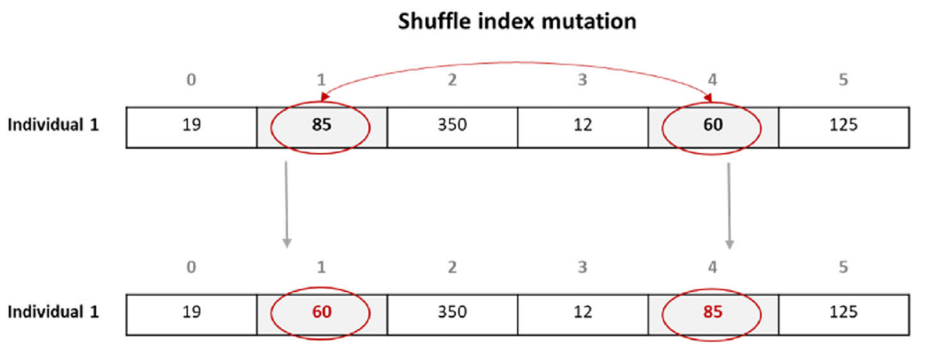


Figure 12. Example of shuffle index mutation. [Colour figure can be viewed at wileyonlinelibrary.com]

random value to the value at that position. The shuffle index mutation swaps two values within the individual (Fortin et al. 2012). The shuffle index mutation was chosen and applied because it kept the overall employment allocation aligned with the control totals and just created variation within the employment sectors. Fig. 12 gives an example of how the shuffle index mutation works.

Select the best offspring

The final step in developing the algorithm was deciding on the method to use to select the best offspring of which the next generation would consist of. In order to select the best solution when using multiobjective optimization, a compromise needs to be reached between satisfying both objectives. For the employment allocation algorithm, the best offspring was chosen based on individuals that had a combination of the best fitness scores for both Objective 1 and Objective 2.

Validation of the results

In various fields, validation often takes place in two forms, internal and external validation. Internal validation is performed by comparing the simulated results to actual, verified data that was used to develop the simulated results (Steyerberg et al. 2001, 2003; Edwards and Clarke 2009; James, Lomax, and Birkin 2019). In this article, one way of internal validation would be comparing the allocated employment opportunities to the employment information that was used as control totals during the evaluation of the fitness function at municipal level. The data used for internal validation is often at a more aggregated level, therefore the more detailed simulated data would then be aggregated to the same level at which the validation data is available in order to compare the two.

External validation is usually done by making use of an external data source to measure against the simulated data in order to see the accuracy of the simulation. This external data should not stem from the same source that is used for internal validation (Steyerberg et al. 2001, 2003; Edwards and Clarke 2009; James, Lomax, and Birkin 2019). For this specific study, external validation is problematic as there is no external data set readily available in South Africa on employment opportunities (which is also why the algorithm was created for the study). The only data that is readily available is the number of people that are employed within the municipality and the most detailed level this data is available at, is sub-place level. Therefore, in the section Discussion of results, the allocated employment opportunities are compared to the Census employment data.

Internal validation of algorithm

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 visualized using various graphs to determine whether the algorithm was performing as expected or not. One of the measures that was used, was plotting the minimum and maximum values for both Objective 1 and Objective 2 to see if the algorithm results converged to a final solution. Fig. 13 provides the minimum and maximum values for Objective 1 and Objective 2 for one of the runs.

Fig. 13 illustrates that the algorithm exhibited the behavior of a typical evolutionary algorithm. Initially, there was a lot of variation between the solutions but as the algorithm progressed, there was a convergence to a final solution. At about the 30th iteration, the marginal improvement of the algorithm decreased significantly. The algorithm stopped showing any significant improvement at about the 110th iteration and was subsequently terminated. Another measure of accuracy is shown in Table 6, which indicates the evaluation of the difference at municipal level between the allocated employment and the actual employment, based on the national Census data, for one of runs.

From Table 6 it can be seen that overall, the algorithm under allocated 2082 employment opportunities for the entire Ekurhuleni. This led to an overall error of 0.21%, which was deemed acceptable for the purposes of this study and the final allocation was accepted. There was higher variation per economic sector than for the overall allocation, however, most economic sectors had an error of less than 1%, which was favorable.

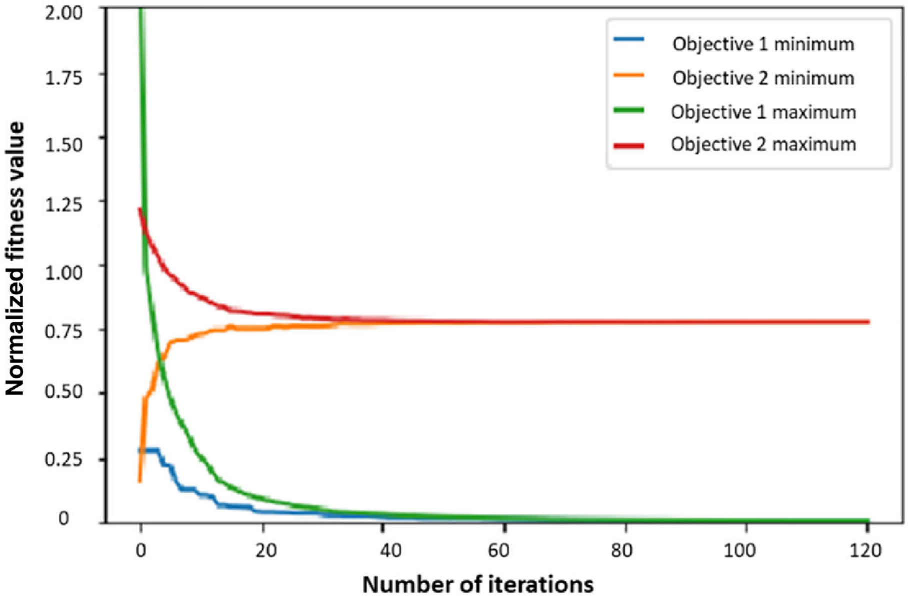


Figure 13. Evaluation criteria graph for the algorithm. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 6. 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

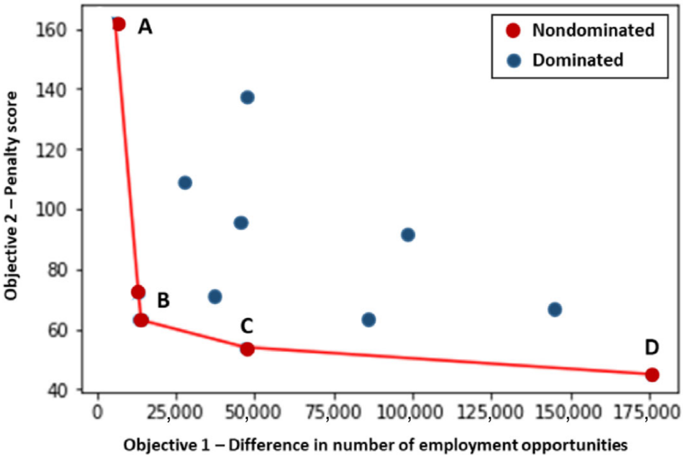
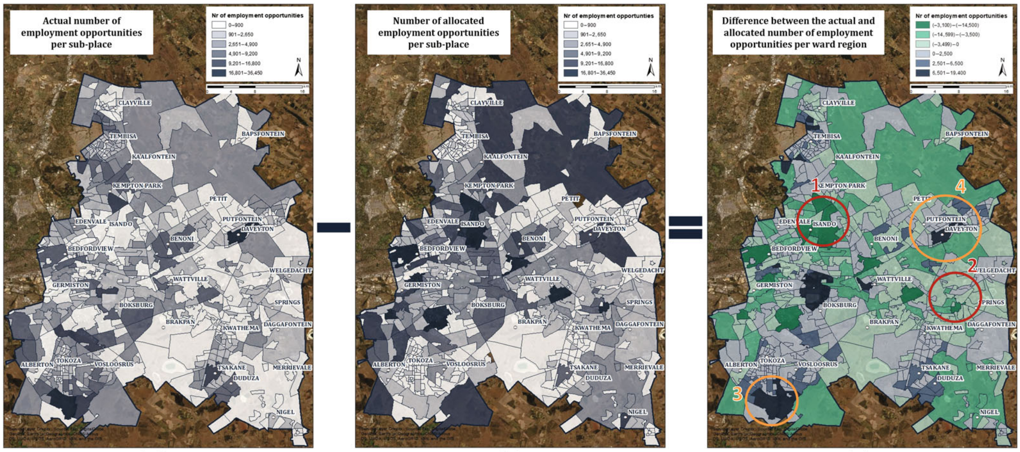


Figure 14. Pareto front graph for the final run. [Colour figure can be viewed at wileyonlinelibrary.com]

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. The RESIDENTIAL buildings data set already had a precalculated employment capacity for each building. This capacity had an exact number of employment opportunities linked to each building and therefore, the allocation of the employment opportunities was perfectly aligned.

The final validation measure was examining the algorithm's Pareto front for the final run, which is depicted in Fig. 14. 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 nondominated 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 the most important factor to consider in the final solution.

For the employment allocation, the trade-off was between the total number of employment opportunities per economic sector and the total number of employment opportunities per building. In other words, the trade-off was that the solution could either be more accurate per building and less accurate per economic sector or vice versa. Solution A in Fig. 14 represents 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. Conversely, Solution D indicates a solution in which 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 favor one objective as is the case with A and D. 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.



(a) (b) (c)

Figure 15. Number of employment opportunities per sub-place. (a) Number of employment opportunities per sub-place as determined in 2011 Census; (b) number of employment opportunities per sub-place as allocated by the algorithm; and (c) difference in number of employment opportunities between (a) and (b). [Colour figure can be viewed at wileyonlinelibrary.com]

Discussion of results

One of the benefits of having employment data at such a fine scale is the fact that the data can be reaggregated 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 different types of information than the employment data that is provided by Census, for example. In the Census data, the number of employment opportunities that are provided per sub-place or main place is based on the home location of the individuals 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 totals 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. To demonstrate this, the results are presented at various levels, which includes sub-place level and region 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. Fig. 15 shows the total number of employment opportunities per sub-place for Ekurhuleni.

The two red circles on the Fig. 15c show two areas where there are major differences between the Census data and the allocated employment opportunities (more employment opportunities allocated than there were in Census data). 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 second map, it can be seen that in the sub-place there are between 16,801 and 36,450 number of employment opportunities. This is more accurate as the O.R. Tambo airport is one of the major employment hubs in the metro. There are also many businesses surrounding the airport that provide products and services to the airport.

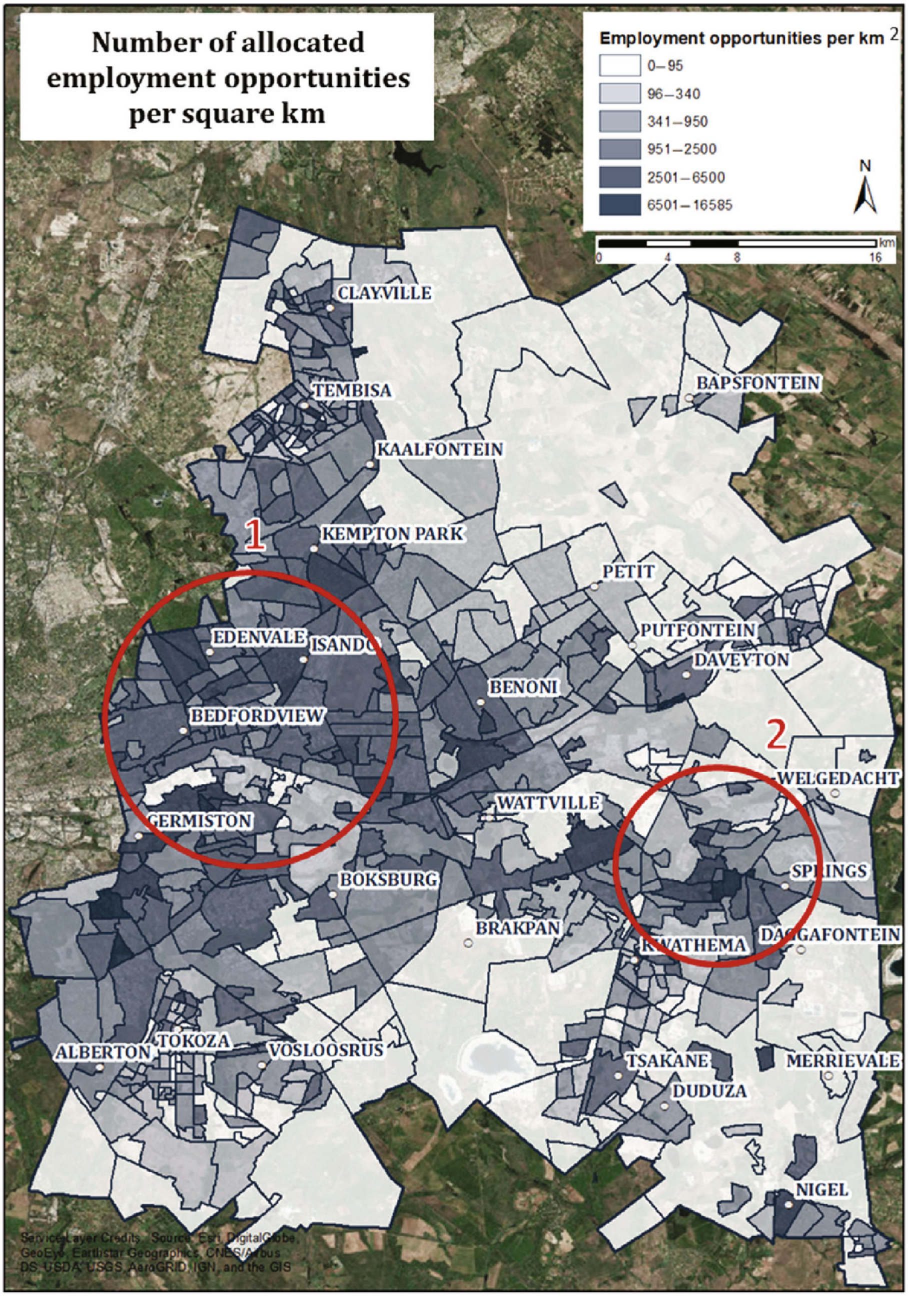


Figure 16. Number of allocated employment opportunities per square kilometer for the City of Ekurhuleni. [Colour figure can be viewed at wileyonlinelibrary.com]

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, for that reason there are no employment opportunities in this sub-place in the Census data. On the second map, it can be seen that this sub-place fall in the second classification area of between 901 and 2,650 employment opportunities. This is a significant increase in employment opportunities, which makes it 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 on the third map 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 (see Fig. 15a), this is an area that shows a significant number of employed individuals, but it is also an area that is mostly residential. In the second map, it can be seen that there are a lower number of employment opportunities, and it falls in the second classification class of 901–2,650 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 have a higher number of employed individuals. In reality, there are not as many nonresidential buildings in the area and on the second map, it can be seen that many of the sub-places within the circle have a lower number of employment opportunities.

Fig. 16 shows the total number of employment opportunities per square kilometer on sub-place level for Ekurhuleni. 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 kilometer. The figure indicates that there is a higher density of employment opportunities in the West of Ekurhuleni and the Eastern parts of Ekurhuleni.

Conclusion

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 is that they find a number of good solutions in addition to an optimal solution, and so there are other good solutions available as well. Considering the results at a more aggregated level (i.e., sub-place level), the results highlighted the high employment areas that are expected to form the employment hubs in a municipality. Examples of these expected employment hubs would be the CBD areas within a municipality, highly developed areas with higher densities, and transportation hubs (i.e., airports and harbors). The areas with lower employment numbers are also where they are expected to be, namely areas where there is a larger concentration of residential rather than nonresidential types.

The algorithm had an acceptable level of accuracy for the overall employment totals. The algorithm allocated 0.21% fewer employment opportunities than the actual amount of employment. The algorithm proved to be less accurate on sector level. A solution to this could be to improve the employment capacity calculations even further to be more specific for those sectors that have a larger difference between the allocated and actual number of employment opportunities. The results from this study prove to be more accurate at sub-place level than Census data. This is useful to, among others, researchers using place of employment in synthetic populations for transport and urban growth models and developing more accurate origin–destination (OD) matrices for transport models.

This research can be further expanded in future by applying the algorithm to years other than 2011. The algorithm can be run using base data from any year as long as there is data available. Calculating the employment opportunities for various years could be helpful to do time series analysis and see how the employment hubs of a city has changed over the years at a finer scale. In South Africa, the next Census is taking place in 2022 and once the results of the Census are available, the algorithm can be applied and a more updated data set can be created.

Acknowledgements

We would like to acknowledge Jenny Holloway, Athi Gxumisa, and Gerbrand Mans for steps in related data preparation processes that informed some of the processes in this article. We would also like to thank Dr Victoria Rautenbach for inputs to early versions of the article.

References

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

Abutaleb, K., O. J. Taiwo, F. Ahmed, and A. Ngie. (2013). *Modeling Urban Change Using Cellular Automata: The Case Study of Johannesburg, South Africa*. Johannesburg and Stellenbosch: University of Johannesburg, s.n.

Antoni, J., G. Vuidel, and O. Klein. (2017). "Generating a Located Synthetic Population of Individuals, Households, and Dwellings." Luxembourg Institute of Socio-Economic Research (LISER) Working Paper No. 2017-07.

Brelsford, C., J. Lobo, J. Hand, and L. M. A. Bettencourt. (2017). "Heterogeneity and Scale of Sustainable Development in Cities." *PNAS* 114(34), 8963–8.

Chaudhuri, G., and K. C. Clarke. (2013). "The SLEUTH Land Use Change Model: A Review." *The International Journal of Environmental Resources Research* 1(1), 88–105.

Chen, P., Z. Fu, and A. Lim. (2002). "Using Genetic Algorithms to Solve the Yard Allocation Problem." In *Proceedings of Genetic and Evolutionary Computation Conference 2002 (A recombination of the GP-2002 and the ICGA-2002), New York City, USA. San Francisco: Morgan Kaufman.* 1049–56.

Cooperative Governance & Traditional Affairs. (2020). *Profile: City of Ekurhuleni*. Pretoria, Gauteng, South Africa: COGTA.

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.

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.

Demographia. (2019). *Demographia World Urban Areas: 2019*. Belleville, IL: Wendell Cox Consultancy.

Edwards, K. L., and G. P. Clarke. (2009). "The Design and Validation of a Spatial Microsimulation Model of Obesogenic Environments for Children in Leeds, UK: SimObesity." *Social Science & Medicine* 69(7), 1127–34.

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

Environmental Geography. (2021). "CLUE Model." [Online] Available at: https://www.environmentalgeography.nl/site/data-models/data/clue-model/.

Fortin, F. A., F. M. De Rainville, M. A. G. Gardner, M. Parizeau, and C. Gagné. (2012). "DEAP: Evolutionary Algorithms Made Easy." *The Journal of Machine Learning Research* 13(1), 2171–5.

Guan, D., et al. (2011). "Modeling Urban Land Use Change by the Integration of Cellular Automaton and Markov Model." *Ecological Modelling* 222(20), 3761–72.

Huynh, N. N., J. Barthelemy, and P. Perez. (2016). "A Heuristic Combinatorial Optimisation Approach to Synthesising a Population for Agent-based Modelling Purposes." *Journal of Artificial Societies and Social Simulation* 19(4), 1–6.

James, W. H., N. Lomax, and M. Birkin. (2019). "Local Level Estimates of Food, Drink and Tobacco Expenditure for Great Britain." *Scientific Data* 6(1), 1–21.

Kumar, M., M. Husain, N. Upreti, and D. Gupta. (2010). "Genetic Algorithm: Review and Application." Available at SSRN 3529843.

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., and P. Augustijn. (2017). "Quantifying the Spatial Implications of Future Land Use Policies in South Africa." *South African Geographical Journal* 99(1), 29–51.

Le Roux, A., K. Arnold, S. Makhanya, and G. Mans. (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.

Lee, Z. J., S. F. Su, C. Y. Lee, and Y. S. Hung. (2003). "A Heuristic Genetic Algorithm for Solving Resource Allocation Problems." *Knowledge and Information Systems* 5(4), 503–11.

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

Osman, M. S., M. A. Abo-Sinna, and A. A. Mousa. (2005). "An Effective Genetic Algorithm Approach to Multiobjective Resource Allocation Problems (MORAPs)." *Applied Mathematics and Computation* 163(2), 755–68.

Patterson, Z., M. Kryvobokov, F. Marchal, and M. Bierlaire. (2010). "Disaggregate Models With Aggregate Data: Two UrbanSim Applications." *The Journal of Transport and Land Use* 3(2), 5–37.

Quantec. (2020). "EasyData by Quantec". [Online] Available at: www.easydata.co.za/.

Ramaswami, A., et al. (2016). *Meta-Principles for Developing Smart, Sustainable, and Healthy Cities*. Science. 352, 940–3.

Seto, K. C., J. S. Golden, M. Alberti, and B. L. Turner, II. (2017). "Sustainability in An Urbanizing Planet." *PNAS* 114(34), 8935–8.

Statistics South Africa. (2012). "*Standard Industrial Classification of All Economic Activities*."

Steyerberg, E. W., F. E. Harrell, Jr., G. J. Borsboom, M. J. C. Eijkemans, Y. Vergouwe, and J. D. F. Habbema. (2001). "I*nternal Validation of Predictive Models: Efficiency of Some Procedures for Logistic Regression Analysis*." *Journal of Clinical Epidemiology* 54(8), 774–81.

Steyerberg, E. W., S. E. Bleeker, H. A. Moll, D. E. Grobbee, and K. G. Moons. (2003). "Internal and External Validation of Predictive Models: A Simulation Study of Bias and Precision in Small Samples." *Journal of Clinical Epidemiology* 56(5), 441–7.

Talbi, E.-G. (2009). *Metaheuristics: From Design to Implementation*, Vol 74. Hoboken, NJ: John Wiley & Sons.

UCSB. (2021). "Project Gigalopolis." [Online] Available at: http://www.ncgia.ucsb.edu/projects/gig/index.html.

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. (2021). "UrbanSim." [Online] Available at: http://www.urbansim.com/urbansim/.

Waldeck, L., and J. Holloway. (2016). *Technical Report on Determining the Place of Work for Urban Growth Simulation*. South Africa: Council for Scientific and Industrial Research.

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."