Extracting input data for residential waste collection capacitated arc routing problems

Llewellyn J. Steyn

A dissertation in partial fulfilment of the requirements for the degree

Master of Engineering (Industrial Engineering)

in the

Faculty of Engineering, Built Environment and Information Technology

University of Pretoria

August 4, 2021

Abstract

Title: Extracting input data for residential waste collection capacitated arc routing problems
Student name: Llewellyn James Steyn
Student number: u13067215
Supervisor: Dr Wilna Bean

Residential waste collection is an essential but expensive public service provided by governments throughout the world. A key contributor to the cost of waste management is collection cost, making the potential for cost savings on waste collection an area of focus. One way to reduce collection cost is through the use of vehicle routing to improve collection routes. While various vehicle routing problem definitions exist for waste vehicle routing, the most compelling is the Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities (MCARPTIF). A challenge facing the MCARPTIF however is that the input parameters necessary to solve real world instances of the problem are difficult to estimate. These include the time taken to drop off waste, the collection and traversal time per street segment and the waste generation rate per street segment. Global Positioning System (GPS) devices and publicly available data sets offer an opportunity to provide insight into some of these parameters and to develop more realistic MCARPTIF instances and subsequently collection routes. This dissertation aims to demonstrate how these parameters can be efficiently estimated. Using GPS data and known landfill locations, landfill visit durations are estimated at a landfill in a metropolitan area. Landfill visit durations are estimated to average 16 minutes. In addition, landfill durations are shown to increase with congestion within the facility. Using GPS data and publicly available street network data from the same metropolitan area, the average vehicle velocity when collecting waste over seven case study areas was found to be 3.857 km/h. The vehicle velocity when traversing street segments within the case study areas was found to average 6.843 km/h. A synthetic population based on census data and per capita waste generation estimates was used to estimate waste generation rates per street segment for a number of case study areas. All of the above mentioned variables were compared to known parameter assumptions used in literature and differ considerably. Lastly the parameter estimates were used to solve a number of real world instances of the MCARPTIF and were compared to instances using parameters from literature. Differences between instances solved using parameters estimated in this dissertation and those based on assumptions from literature illustrate the importance of using accurate input data for waste collection routing applications.

Contents

List of Figures

List of Tables

List of Acronyms

VRP Vehicle Routing Problem

WCVRP Waste Collection Vehicle Routing Problem

CARP Capacitated Arc Routing Problem

MCARP Mixed Capacitated Arc Routing Problem

CARPIF Capacitated Arc Routing Problem with Intermediate Facilities

CARPTIF Capacitated Arc Routing Problem with Time intervals and Intermediate Facilities

CLARPIF Arc Routing Problem with Intermediate Facilities under Capacity and Length Restrictions

MCARPTIF Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities

GPS Global Positioning System

SVRPTW Stochastic Vehicle Routing Problem with Time Windows

MSW Municipal Solid Waste

GDP Gross Domestic Product

KS Kolmogorov Smirnov test

OSM *OpenStreetMap*

IQR Inter Quartile Range

API Application Programming Interface

ARN Actual Road Network

RNG Random Network Generation

PBI Public Benchmark Instance

KPI Key Performance Indicator

PUMS Public Use Micro Sample

Chapter 1

Introduction

1.1 Problem background

Municipal waste collection is a key service offering provided by municipalities and local governments to citizens all over the world. While it is an important portion of local service delivery to residents, it does not come without considerable cost. Broadly speaking the Municipal Solid Waste (MSW) value chain consists of household waste generation, waste collection, waste segmentation and finally, treatment and disposal. When considering the end to end value chain, collection presents by far the greatest cost. For middle income countries, such as South Africa, Hoornweg (2012) has found that between fifty and eighty percent of the typical waste management budget is spent on collection. It therefore follows that any initiatives aimed at reducing total expenditure on waste management should target collection activities in particular.

In this regard, vehicle routing presents a good opportunity at reducing collection costs. By developing improved collection routes residents can be serviced with fewer collection vehicles at reduced operational cost. To this end literature contains various examples of vehicle routing models aimed at improving waste collection and similar operations. The Capacitated Arc Routing Problem (CARP), first proposed by Golden and Wong (1981), is a problem definition particularly suited to waste collection vehicle routing applications.

When applying the CARP to waste collection, each road segment represents an arc or edge that must be serviced by a waste collection vehicle. Each road segment has a number of parameters, such as the amount of waste produced along the segment, the vehicle travel time along the segment and the vehicle service time for the segment. The objective of the vehicle routing problem is then to develop routes of minimum cost for a set of heterogeneous collection vehicles with a fixed capacity so that all street segments where waste is generated by residents are serviced. For the latest review of the CARP the reader is referred to the work of Corberán and Laporte (2015) and Mourão and Pinto (2017).

The basic CARP model is expanded in literature to a number of related variants. Bautista et al. (2008) define the first of these by developing the Mixed Capacitated Arc Routing Problem (MCARP), where the problem is expanded by including a mixed road network that takes into account the direction of travel along street segments (one way or bidirectional travel). Bautista et al. (2008) solve the problem by making use of ant colony optimisation. Another problem definition is Ghiani et al. (2001) where constructive heuristics are used to solve the Capacitated Arc Routing Problem with Intermediate Facilities (CARPIF) where vehicles can complete multiple sub-trips and drop waste off at landfills or transfer stations. Ghiani et al. (2004) and Ghiani et al. (2010)

use a tabu search and ant colony optimisation, respectively, to solve the Capacitated Arc Routing Problem with Time intervals and Intermediate Facilities (CARPTIF) where the route length is restricted to limit trip durations to a single working shift. However, the most comprehensive waste collection problem definition is presented in Willemse (2016); Willemse and Joubert (2016b,c) in the form of the Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities (MCARPTIF). It is the most comprehensive problem definition for curbside waste collection since it incorporates both a mixed road network as well as multiple waste drop-offs at landfills or intermediate facilities. In addition, similar to Ghiani et al. (2010), a time constraint is imposed which limits the total duration of a route from exceeding the available working hours in a shift.

A shortcoming of many of CARP variants developed in literature however, is not the problem definition itself, but the lack of accurate real world data to solve instances of the problem. The effect of this, as Ghiani et al. (2014) observe is that very few papers make use of stochastic input parameters for essential variables such as waste generation rate and travel times. For example, to produce accurate real world results the MCARPTIF requires accurate velocities across street segments, both when servicing the segment and when not servicing the segment (deadheading). In addition to this, the duration of visits to landfills or intermediate facilities and waste generation rates per street segment is also required. Although these can be estimated for small areas, efficiently collecting this data for a whole metropolitan area is impractical. This data is also not static, since various factors affect and change these variables on a continuous basis. This makes any static observation of these variables invalid for an extended period of time and these variables must therefore continuously be updated. Literature therefore presents a wide range of vehicle routing strategies and methods for solving CARP variants related to waste collection, but methodologies for estimating input data are limited to the work of Ghiani et al. (2015). In terms of estimating waste generation rates various strategies do exist, as summarised by Beigl et al. (2008). Many of these methods do however suffer from the problem of only being able to predict waste generation rates in specific temporal or spatial settings. This limits the effectiveness of any vehicle routing model on real-world applications where accurate and timely data is required to make tactical and operational decisions.

Wilson et al. (2007) reports on the use of Global Positioning System (GPS) data to analyse vehicle activity. The authors find that GPS data is reliable, accurate and that great potential for analysis of the data exists. For waste collection fleets that are fitted with GPS devices, this is a data source which can provide insight into waste collection operations.

Figure 1.1 provides a short example which will illustrate this point. The figure shows a residential road network in a metropolitan area in South Africa. Waste collection vehicles service the area once a week, and the MCARPTIF problem definition provides an opportunity to find the optimal sequence of street segments in which to service the area, such that the overall service time is minimised. However to solve an MCARPTIF instance of this area the waste generation rate per street segment must be known, so that the vehicle capacity is not exceeded. In addition an estimate is required as to the amount of time required to service each segment. This is where the GPS points scattered across the network in figure 1.1 could prove to be useful since the duration of time spent within each segment can be calculated. In addition this data could be used to calculate vehicle travel times to landfills as well as the amount of time spent within the landfills.

The application of Big Data techniques to GPS data and other public data sets, therefore presents a good opportunity to estimate the parameters required to develop

Figure 1.1: GPS points and a road network in a metropolitan area in South Africa

accurate waste collection routes. The outcome of this could be routine decision support models that develop optimal collection routes based on accurate routing parameters that are updated on a continuous basis. In the long term, the possibility exists for live routing models that automate all routing decisions for waste collection vehicles and in doing so reduce collection costs.

Given the potential benefits of waste collection vehicle routing on the cost of the waste value chain for municipalities and local governments, particularity in developing countries, this dissertation poses the following research question: Can GPS data and publicly available data sets be used to better estimate MCARPTIF routing parameters to develop realistic real world routing instances and solutions?

1.2 Research design

In this section the research question is further broken down into its sub-components. Since the MCARPTIF requires multiple input parameters the research design poses a number of sub-questions which must be addressed by the dissertation. These are namely:

1. *Can GPS data be used to estimate landfill visit durations?* Landfill visit durations are important since time spent at landfills reduces the productive time spent collecting waste.

2. *Can GPS data and publicly available street networks be used to estimate service and dead-heading times for individual street segments?* Service and deadheading costs are crucial as they directly impact the duration of any proposed collection route, for this reason accurate estimates of these parameters are integral to collection route quality.

3. *Can waste generation rates be estimated using publicly available census data and known generation rates?* An important aspect of any CARP variant that aims to

solve waste collection problems is that vehicle capacity is taken into account. Since vehicles have a finite capacity the amount of waste generated on a street segment limits the number of segments the vehicle can service. A good estimate of waste generation rate is therefore important in producing quality collection routes.

4. *Can realistic MCARPTIF instances be generated from the above input data and can these instances be solved?* The above data requirements must ultimately culminate in problem instances that are solvable by MCARPTIF solution algorithms and as such this component of the research question must be addressed.

In addressing this research question, and its components, this dissertation presents the following artefacts:

1. Algorithms for the estimation of landfill visit duration given waste collection vehicle GPS records and landfill locations.
2. Algorithms for estimating service and dead-heading times per street segment given waste collection vehicle GPS records and service area street network.
3. Distribution of service and deadheading times per street segment.
4. Algorithms for estimating waste generation rates based on a synthetic population generated from census data and known waste generation rates per population income group.
5. MCARPTIF instances with all of the above parameters estimated for a number of case study service area based on actual waste collection data sources.
6. Solutions generated using existing algorithms on the real MCARPTIF instances, and compared to solutions based on widely used MCARPTIF input parameter assumptions in literature.

1.3 Research methodology

The research methodology followed in this dissertation is primarily based on articles by Pidd (2010) and Manson (2006). Pidd (2010) presents four main categories of model use. These are namely *Decision Automation, Routine Decision Support, System Investigation and Improvement* and finally *Providing insights for debate*. This represents a spectrum of model use between almost no human interaction (Decision Automation) and a high degree of human interaction (Providing insights for debate).

The desired outcome of the research presented in this dissertation is the development of algorithms that can estimate input data for the MCARPTIF. With accurate and tangible methods for estimating these parameters, efficient collection beats can be developed for collection vehicles. Collection beats refer to the route taken by a waste collection vehicle on a particular day. A collection beat is therefore a portion of the weekly waste collection task for a metropolitan area, allocated to a collection vehicle and conforming with vehicle capacity and shift constraints.

The models, making use of accurate input data, can be used to develop more efficient collection routes on a routine basis, and in doing so will reduce the operational costs of waste collection operations.

The use of the proposed models fall within the realm of Routine Decision Support, however the models developed for this dissertation are realistically more for the purpose

of System Investigation and Improvement with the aim of prompting future work that will involve Routing Decision Support models. For this reason the modelling requirements for Routine Decision Support models are considered below.

The variables that affect routing change routinely, as traffic conditions and waste generation rates change, for example. Efficient routes must therefore be developed routinely, with support from the routing models. Pidd (2010) also explains that Routine Decision Support models are used by trained individuals with intricate knowledge of the subject matter under consideration. Only when models are able to automate decisions does the skill of the user reduce. The models proposed here are therefore for Routine Decision Support, as they will have to be used by waste collection managers with strong knowledge on local waste collection activities.

In developing a theory of model use, Pidd (2010) proposes three aspects of model use, these are namely the importance of model validation, data requirements for the type of model use and the value added by model use. Validation is discussed in more detail in section 1.3.4.

In terms of data, Pidd (2010) considers the important point of input data requirements as a function of model use. Pidd (2010) specifically stresses the importance of input data for Routine Decision Support models, that large amounts of accurate data are required and that these sources must be updated continuously. The author's emphasis on the importance of input data for Routine Decision Support models serves to further justify the importance of work on input data for the MCARPTIF problem variants.

The third consideration presented by Pidd (2010) with regard to model use is the value added through the use of the model. The author explains that models for Routine Decision Support add value to problems where neither the model, nor the model user, can alone add enough value to overcome the problem. This is true for vehicle routing problems, where the problems are NP-hard and not solvable by humans. However the models are at the same time just abstractions of reality and they do not incorporate the practical and operational realities of daily waste collection activities. The optimal solution is therefore Routine Decision Support where the model and the decision maker work in conjunction.

Pidd (2010) provides a good framework for defining model use and for understanding the advantages, challenges and requirements associated with the use of different model types.

To improve the solution quality the modelling framework proposed by Manson (2006) for the field of Operations Research was used. The framework proposed byManson (2006) is appropriate as it requires rigorous and iterative understanding of the problem, solution development and evaluation of potential solutions. The process has five components, these are *Awareness of the problem*, *Suggestion*, *Development*, *Evaluation* and *Conclusion* which guides the researcher towards the development of a model where the model outcome is evaluated against the stated objective. In following this research approach the following five steps were undertaken.

1.3.1 Problem awareness and identification

Awareness of the problem came about through a thorough literature review into waste vehicle routing and the MCARTPIF problem definition proposed by Willemse (2016); Willemse and Joubert (2016b,c). Literature demonstrates that the current gap relates to the generation of input data to make current solution algorithms feasible for real world applications. Awareness of the problem also involved research into Municipal Solid Waste Management, the costs associated with collection and the potential improvement opportunities represented by accurate vehicle routing models.

1.3.2 Evaluation of solution strategies

As part of the suggestion phase potential data sources were considered. GPS data was identified as a feasible source of routing data, with a number of preliminary models and studies supporting the use of GPS data to generate MCARPTIF parameters. While GPS data is not universally used by waste collection fleets, it is relatively common and a test data set could be obtained for preliminary evaluation.Preliminary research was conducted during an undergraduate thesis by Steyn (2016) and as part of conference proceedings by Steyn and Willemse (2018). The work authored by Ghiani et al. (2015) also served as a proof of concept.

1.3.3 Model development

For the developmental phase, landfill visit duration was estimated. This determines how long vehicles spend at landfills while dropping waste off, and before continuing with the rest of the days route. The next step was to determine traversal times per street segment, for vehicles servicing and deadheading segments. Once this was complete the next step was to determine waste generation rates using a synthetic population developed from census data. Finally all parameters were combined and a test case was solved using known solution algorithms developed by Willemse (2016); Willemse and Joubert (2016b,c).

1.3.4 Model evaluation

Model evaluation is a crucial step in the developmental phase. As part of the development phase models were continuously evaluated both quantitatively and qualitatively against known MCARPTIF routing parameters. In particular statistical methods were used to evaluate whether parameter estimates were statistically relevant and appropriate for use in the solution algorithms. More detail on this will follow in the next chapters. To further improve model evaluation the models were also tested against a number of additional areas to estimate MCARPTIF parameters and solve instances of the routing problem and demonstrate model scalability over multiple service areas within the metropolitan area.

1.4 Expected contributions

As a result of this dissertation the following contributions are expected. These are namely efficient methods for the evaluation of waste vehicle drop-off durations at landfills or transfer facilities using GPS data and geofences. A methodology for the use of a synthetic population based on census data to estimate residential waste generation rates. Methods for estimating vehicle velocity over street segments, as well as methods of separating vehicle velocity samples into service and deadheading velocities for the MCARPTIF algorithms. Finally this dissertation also presents comparisons between standard parameter estimates for waste collection routing problems and those developed here and the impact of these estimates on collection routes. Practically these contributions are expected to lead to better Routine Decision Support models, based on GPS Data and publicly available data sources where waste collection vehicle routing problems can be routinely solved using accurate data to reduce collection costs.

1.5 Document structure

Following this chapter is a literature review where the above topics are discussed in detail and critically evaluated. Alternative solution strategies are assessed and the strategies with the highest probabilities of success are selected.

Following the literature review the analysis chapters are presented. The first of these is Chapter 3 and looks at using GPS data to estimate landfill drop off durations as well as investigating the effects of congestion within a landfill on drop off durations. Following this is Chapter 4, which focuses on estimating traversal and service costs for service areas using detailed road networks and GPS data. Chapter 5 estimates waste generation rates for service areas based on per capita waste generation rates and a synthetic population developed from census data.

Finally in Chapter 6 a number of test instances are solved, based on the actual data collected in the previous Chapters, and compared to input data assumptions from literature on MCARPTIF algorithms to evaluate the effect of using actual data estimates on feasible routes.

Chapter 2

Literature review

In reviewing literature relevant to this dissertation, Capacitated Arc Routing Problem (CARP) variants, with a focus on the input data required for different variants, are discussed. Following that, stochastic routing models and their input variables are discussed. The specific input parameters for the Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities (MCARPTIF) is then discussed, along with potential methods for determining these parameters.

2.1 Capacitated arc routing problems

The CARP refers to an optimisation problem, first proposed by Golden and Wong (1981), where edges have to be serviced by a fleet of vehicles. CARP problems differ from traditional Vehicle Routing Problems (VRPs) where a number of nodes must be serviced by a vehicle, or fleet of vehicles, as opposed to edges in the case of CARP. An *edge* refers to a street segment which can be traversed in both directions. When an edge can only be traversed in a single direction it is termed an *arc*, however this will be dealt with in more detail at a later stage. Each edge can be traversed, while some edges have demand for a service. The particular definition of *demand* and *servicing* changes depending on the variant of the problem. Gendreau et al. (2015) identifies waste collection, distribution planning, mail delivery and salt gritting as popular applications of CARP instances.

During residential waste collection, vehicles move from house to house, collecting waste placed on the sidewalk by residents. Each street segment represents an edge which must be serviced by the vehicle. Vehicles have a particular capacity and therefore the demand on each route (or subset of edges serviced by a particular vehicle) cannot exceed the vehicle capacity. Vehicles are considered homogeneous, meaning vehicles in the fleet have the same speed and capacity characteristics. Vehicle routes start and end at the same vehicle depot. This then describes the basic CARP model. For the latest review of the CARP the reader is referred to Corberán and Laporte (2015) and Mourão and Pinto (2017).

The classic CARP does not completely cater for problem characteristics specific to waste collection. Waste collection operations require mixed road networks. This refers to the properties of particular road segments. Some streets are one-ways and can therefore only be serviced or traversed in one direction while other streets are two-ways but might be very busy, meaning that each side of the street must be serviced separately for example. By taking this into consideration Belenguer et al. (2006) turn the problem into the Mixed Capacitated Arc Routing Problem (MCARP). The MCARP therefore takes into account both edges and arcs. Bautista et al. (2008); Belenguer et al. (2006) both solve MCARP variants for waste collection. The addition of mixed road networks in itself presents

an input data problem, as accurate, up to date road network data is required. This is addressed in section 2.3.

Ghiani et al. (2001) further develop the CARP by adding intermediate facilities to the problem . During normal operations, vehicles stop collecting once the vehicle is full and dump waste at a landfill or transfer station. The vehicle then resumes collection operations after that. This means that each sub-trip between landfill visits cannot exceed the vehicle capacity. The problem is also addressed by Polacek et al. (2008). The authors present a solution strategy for this problem variant, based on a variable neighbourhood search, but do not provide information on how intermediate facility visit durations can be determined. The addition of this assumption requires an estimate of intermediate facility visit durations, as the length of time spent at the intermediate facility will have a bearing on the feasibility of the route. Literature with regards to estimating this parameter is addressed in section 2.4.

The next crucial assumption and addition to the problem definition is that of length or time restrictions. Vehicle routes must be constrained by time since collection operations can only occur within a certain amount of time, which is usually determined by the length of the collection crew's shift. The Capacitated Arc Routing Problem with Time intervals and Intermediate Facilities (CARPTIF), also called the Arc Routing Problem with Intermediate Facilities under Capacity and Length Restrictions (CLARPIF) was first proposed by Ghiani et al. (2004). The problem is additionally solved by Ghiani et al. (2010).

The addition of time into the problem definition is a crucial improvement but also poses further challenges. By imposing time restrictions on vehicle activity the implication is that activity times must now be estimated. These would include collection time and travel time between collection points as well as the drop-off duration for any visits to landfills or intermediate facilities.

The combination of all of the above into the MCARPTIF by Willemse (2016); Willemse and Joubert (2016b,c) then defines the most comprehensive problem definition for curbside waste collection.

Another problem definition is the Waste Collection Vehicle Routing Problem (WCVRP) group of problems. The reader is referred to the most recent article on the topic by Aliahmadi et al. (2021) where a bi-objective problem is solved to minimise both cost and collection time. The difference between the MCARPTIF and the WCVRP is that the MCARPTIF is servicing arcs, or street segments, while the WCVRP is servicing known demand locations, much like the travelling salesman problem. For instance, Aliahmadi et al. (2021) solve a case study where bin locations are known and waste generation rates per bin are sampled from triangular distributions. The MCARPTIF is therefore more suitable for problems with unknown bin locations, which is the case for the residential curbside waste collection problem solved in this dissertation.

Having identified the correct type of vehicle routing problem, consideration can be given to the input data required to solve the problem. This includes the street network of the area to be serviced. For each segment in the network, the waste generated for that segment (or the street demand), the service and travel times of the vehicle through that segment, and the direction of travel and service through that segment are required. Additional input data required includes vehicle capacity, location of intermediate facilities and the vehicle depot and shift durations of the crews. Different approaches to obtain the input variables into the MCARPTIF problem are considered below.

2.2 Stochastic routing models

Routing models based on accurate input data from real-world sources are likely to be stochastic in nature. To this end literature on stochastic routing models, both for standard VRPs and CARPs variants are discussed. The aim of including literature on stochastic problem variants is to evaluate how model parameters for stochastic problems are estimated and whether actual data sets are used to estimate parameters.

Laporte et al. (2010) solve a stochastic CARP variant where route demand is stochastic. The authors develop solution algorithms for a first-stage solution, after which the stochastic demand is revealed. A second set of algorithms then provide recourse should the route exceed capacity after stochasticity has been applied. The paper solves stochastic instances of existing deterministic benchmark instances and derives the stochastic demand variables from Poisson distributions with a mean value equal to the deterministic counterparts. The stochastic variable is therefore not based on any real world demand observation, although it does provide a problem formulation for problems with stochastic demand.

Christiansen et al. (2009) also propose a CARP variant by introducing demand as a stochastic variable. In the paper, a Possion distribution is once again used to model demand on an edge. It is unclear however what the assumption of the Poission distribution is based on or how the distribution parameters are determined per edge. A branch and price algorithm is used to solve the problem.

Literature also looks at service and travel times as stochastic or random variables. Kenyon and Morton (2003) route vehicles through a network where travel times are stochastic.

Chen et al. (2014) consider an interesting CARP variant where daily maintenance operations are scheduled on a network with stochastic service and travel times. Input data is reported to be from real data sources, however no information is provided on the data sources or any pre-processing to produce the service and traversal times. Further optimisation strategies are presented on the same problem variant by Chen et al. (2016).

In addition to arc routing problems where routing parameters are assumed to be stochastic there is a large body of work on VRPs where stochastic variables are used. Ehmke et al. (2015) solve the Stochastic Vehicle Routing Problem with Time Windows (SVRPTW). Similar variants are solved by Taş et al. (2014a,b). All of the variants discussed above have pre-determined routes. The stochastic variable is revealed upon execution of the route, and the model has recourse to adjust the route should constraints be violated.

While problem variants and solution strategies do exist for stochastic routing models, in particular stochastic Arc Routing Problems (ARPs), there remains a gap in literature with regard to data sources and processing methods for generating accurate input data for these problems, in particular for waste collection applications. The following sections deal with estimating these parameters.

2.3 Road network and benchmark instances

Any CARP variant requires a road network to be an accurate representation of the real world. To develop efficient collection routes CARP problem instances must therefore be constructed from existing road networks. The process of constructing CARP instances from road networks is therefore an important part of the solution process.

For the purpose of testing solution strategies, most authors either produce random networks or make use of existing benchmark instances available in the public domain.

Belenguer et al. (2006) develop a random network for testing, while Bautista et al. (2008), Willemse and Joubert (2016b) and Kiilerich and Wøhlk (2018) produce networks using actual street network data. The network graphs are constructed from data containing the coordinates of endpoints of each street segment as well as the segment length. Household data is also known, with the exact locations and composition of each household. Households are then allocated to street segments. This is similar to the process undertaken in Chapter 5, with the exception that the network data in this case is based on open source map data and the household data is synthesised, due to the frequent lack of rich data in developing nations. Most authors who solve CARP variants however either use existing benchmark instances, or synethically create instances on which to test algorithms.

Lum et al. (2018) provide an efficient tool for developing benchmark instances directly from *OpenStreetMap*. *OpenStreetMap* is an open source software platform, pioneered by Haklay and Weber (2008) that allows users to contribute to the mapping of street networks for public use. Various authors report using *OpenStreetMap* for the generation of benchmark instances.

Kiilerich and Wøhlk (2018) generate large CARP instances for waste collection in Denmark. The authors have access to detailed household data and waste generation estimates per household. This data is allocated to a road network to generate demand estimates per arc or edge. Traversal costs for the benchmark instances are not addressed in detail. While this approach is detailed, it requires detailed municipal waste data, which is not always available for all municipal areas, particularly in developing countries.

To address this later chapters detail the process of extracting road networks from open source maps, estimating routing parameters along these networks and constructing benchmark instances based on the road network and routing parameters.

2.4 Intermediate facility visit times

A key input parameter for the MCARPTIF problem variant is the amount of time required for a vehicle to visit a transfer station and drop off waste. At the point where the vehicle has reached capacity it travels to the nearest transfer station and spends time dropping waste off before continuing collection activities. In Belenguer et al. (2006); Willemse (2016); Willemse and Joubert (2016b,c) a 5 minute drop-off duration is used, however how this parameter is determined is not discussed in depth. Benjamin and Beasley (2010) solve a waste collection routing problem on public benchmark instances but also assume a deterministic disposal time at intermediate facilities.

The problem of estimating landfill visit durations is addressed in literature by Wilson and Vincent (2008) by using Global Positioning System (GPS) data and geofences to estimate this variable. Geofences are spatial boundaries drawn around landfills. This allows GPS points within the boundaries to be isolated and visit durations to be estimated. Wilson and Vincent (2008) find landfill visits for five vehicles, at 11 landfills, over a period of one year to average 16.4 minutes, with a standard deviation of 14.3 minutes. When comparing this with the previous assumptions of 5 minutes for example, it illustrates that time estimates based on actual data could differ significantly from those assumed in literature.

Analysis using GPS data also allows the authors to compare landfill visit durations between facilities, as this variable is a function of facility layout and traffic volume. For the MCARPTIF, landfill visit durations can therefore be different depending on the facility, which will likely affect solutions. A vehicle might opt to travel a slightly longer distance to visit a landfill with a faster service time than to visit the nearest landfill with a longer

service time.

GPS data therefore presents a good opportunity to estimate landfill visit durations accurately. This method is also repeatable, scalable and can be regularly updated. Given this opportunity to estimate landfill visit durations, in Chapter 3 this strategy is pursued to estimate visit durations for a metropolitan landfill in South Africa.

2.5 Segment traversal times

Segment traversal times, when vehicles are both servicing and deadheading segments is a crucial piece of information for the MCARPTIF, or for any CARP variant used for vehicle routing, regardless of the application. As discussed earlier, stochastic formulations for this problem are available but input data generation for these parameters are limited to the work of Ghiani et al. (2015).

To estimate traversal costs, most authors make use of a velocity estimate. For instance Belenguer et al. (2006) estimate vehicle velocity at 20 km/h, when a vehicle is traversing or deadheading a segment. When the vehicle is servicing the segment, 10 s per 10 kg bin is added to the cost in order to estimate a service cost per segment. A similar approach is followed by Willemse and Joubert (2016b) where vehicle velocity is estimated at $28km/h$ while deadheading and 14 km/h when servicing, with the addition of 1 s per kg of waste. In Willemse and Joubert (2016a) the same assumptions as Belenguer et al. (2006) are used.

Literature does however contain work on estimating travel times for non-waste applications. Jiménez-Meza et al. (2013) propose a methodology for extracting travel time, distance and speed per street segment for taxis using only GPS data. Given the fact that the input data for the descriptive model will be GPS traces from waste collection vehicles the methods used by Jiménez-Meza et al. (2013) can be used to identify both service and travel times per segment. This implementation is for a taxi, so while it remains relevant a better option is Ghiani et al. (2015) that describe detailed map matching with specific application to waste collection. Ghiani et al. (2015) use the width of the street to experimentally determine the size of the area within which GPS points are assigned to that particular street. While this yields good results for street segments, intersections are more complex. Here Ghiani et al. (2015) compare the performance of different geofence shapes, namely rectangular and circular geofences. The authors also introduce a procedure for classifying points to street segments using the points that follow and precede the particular point. Once points can accurately be linked to a particular street segment the deadheading time can be calculated by taking the time difference between the first and last point in that segment.

In terms of identifying service times, Ghiani et al. (2015) search for points within a known waste collection area. This is done by first clustering points, by grouping points that are below a certain distance from other points together. These clusters are considered service clusters if they are in close proximity to known collection areas. The service time for that cluster is the difference between the first and last point in that particular cluster.

This is different from the residential service time calculated in this paper as collection in this case is curbside waste collection, as opposed to waste collection from centralised points. This poses a unique challenge as it is not immediately clear whether a vehicle is collecting waste or simply traversing the segment, while in the case of Ghiani et al. (2015) this distinction is clear as collection occurs at predetermined locations. While this does present a very good attempt at estimating service and traversal times, determining whether a vehicle is traversing or deadheading is not addressed in literature.

For this reason Chapter 4 addresses this gap by both estimating vehicle velocity over

street segments and separating these velocities into statistically distinct service velocity and deadheading velocity populations.

2.6 Waste generation rates

Waste generation rates refer to the rate that waste is generated by a population and is a crucial input to the MCARPTIF as it represents the demand for the collection service. In reviewing municipal solid waste generation Beigl et al. (2008) find that at settlement level (or suburb level) data sources include census data, market research based on geo-demographic classification packages and questionnaires. Some of the more significant studies that use these data sources are discussed below. The focus in citing these sources is on the results they have produced. Greater detail on the methods used is given in the next section.

In reviewing Municipal Solid Waste (MSW) generation and management Karak et al. (2012) find that in developed countries waste is generated at between $521.95 - 759.2$ kg per person per year. In developing countries this is around $109.5 - 525.6$ kg per person per year.

Qdais et al. (1997) used source sampling at 40 different households to determine a waste generation distribution. The study found that waste is produced at a rate of 1.76 kg per person per day in Abu Dhabi city. Another study by Minghua et al. (2009) in the city of Shanghai, China, produce a result of 1.11 kg per person per day. A study in Kuala Lampur, Malaysia by Saeed et al. (2009) estimate a generation rate of 1.62 kg per person per day while Troschinetz and Mihelcic (2009) reports a generation rate of 0.77 kg per person per day in a study on MSW.

It is clear that solid waste generation rates differ significantly from city to city, and estimates must therefore be obtained for the area of study. For the purposes of this thesis a waste generation rate reported by Solid Waste Management (2016) for the City of Cape Town is used at 580kg per person per annum. This is discussed in more detail in Chapter 5.

Many factors contribute to solid waste generation rates, for example Liu and Wu (2010) find economic growth, household income and urban development to be some of the major factors. Kinnaman (2009) look specifically at the economics of MSW generation and states that the relationship between Gross Domestic Product (GDP), a measure of income, and waste generation rates is positive and linear. Any solid waste generation rate estimate must therefore be based on regional estimates of the aforementioned variables. Determining accurate waste generation rates for the MCARPTIF will directly affect the nature of the routes produced, this is therefore an important input variable.

2.6.1 Source sampling techniques

A challenge with estimating MSW generation rates is that generation rates must be estimated at street segment level for the MCARPTIF. A potential solution is source sampling, or measuring generation rates directly. An example of this is the use of stratified cluster sampling by Dangi et al. (2008). The study makes use of a team of student scientists to physically visit households in Kathmandu Municipality, Nepal to weigh and sort solid waste. By directly measuring the waste produced in an area an accurate estimate of generation rates for the area can be determined. The study estimated waste generation in the city of Kathmandu at 0.1612 kg per person per day. This approach has its limitations though. The reported generation rate can only realistically be a reflection of the sample

period, no indication of seasonal generation rates or variance is therefore present. The study makes use of a number of students equipped with scales and bio-hazard gear to perform the sampling. It is therefore unlikely that this method is a feasible estimation method for a metropolitan area with millions of residents. The last drawback is that waste generation rates are not static, as the underlying variables—such as household income for example—change so does the generation rate. Source sampling therefore does not accommodate continuous updates of these variables.

2.6.2 Synthetic populations from census data

A potential solution to the problem of estimating waste generation rates is the use of synthetic populations, the hypothesis being that if population characteristics predict waste generation rates then waste generation rates can be estimated if population data is available. Traditionally synthetic populations would be used for agent based transport modelling whereby population characteristics are required to model transport patterns, as used by Huynh et al. (2013) to model transportation systems in Sydney, Australia.

Typically census data contains detailed information down to a suburban level, but not down to household level. However since census data commonly, as is the case in South Africa, contains detailed information about a certain portion of the population, more detailed population parameters can be synthesised down to a suburban level.

Harland et al. (2012) compare the various tried and tested methods of producing synthetic populations at various spacial scales. These are deterministic re-weighting, conditional probability and simulated annealing. The basic modelling approach is that a sample of individuals, at a higher level region such as a city or country, is disaggregated to lower level regions such as suburbs by applying weights to individual sample members such that known constrains on the lower level region is satisfied. For instance, one might know that a particular region contains 1000 individuals, where age, ethnicity, gender and income for each individual is known. At the subregion level one might know the constraints, or the characteristics of the particular sub region, for example that there are 50 males in a sub region. The aim of generating the synthetic population is then to allocate the right sample of males to the sub region such that the gender constraint as well the other constraints (age, ethnicity and income) are as close to the sub region as possible.

Deterministic Reweighting achieves this by assigning each individual in the sample a weight, for instance using the above example a male individual would be assigned a weight of $50/1000 = 0.05$. Harland et al. (2012) however note that deterministic reweighting runs into problems where characteristics for geographic areas are very different to the overall population. This could be a problem in South Africa where there are large disparities in population characteristics such as income between different areas.

The Conditional Probabilities method is similar to Deterministic Reweighting except that weights are assigned stochastically and are sampled from a probability distribution. Harland et al. (2012) notes for both Deterministic Reweighting and Conditional Probabilities that the order in which attributes are weighted affects the outcome of the synthetic population and that the most important attribute should be weighted first.

Of the three methods discussed so far, Simulated Annealing performed best. With this method, a random sample from the population is allocated to each sub region and the sub region constraints are then evaluated. The algorithm then swaps members out between regions and tests whether the move improves or deteriorates the sub regions proximity to the constraint.

However, the most promising opportunity for population synthesis for the purposes of this dissertation is presented by Müller and Axhausen (2012). This is because the

algorithm has already been successfully implemented on 2011 Census Data for a number of South African cities by Joubert (2014). The two components that were used by Joubert (2014) to perform multi-level fitting are *Community Profile Data* and the *10% Public Use Data* released by *Statistics SA*. More information on this process is provided in Chapter 5.

2.7 Conclusion

The MCARPTIF requires input data to produce feasible collection routes that reduce the significant collection cost of Municipal Solid Waste. Currently literature presents an abundance of research on solution strategies for MCARPTIF and other CARP waste collection variants. Estimating crucial input variables remains largely unadressed in literature though. In particular, a need exists for estimating waste generation rates per street segment, vehicle traversal and deadheading times per street segment and transfer station visit times. All of the aforementioned variables are likely to be stochastic and will directly impact the quality of collection routes produced by MCARPTIF algorithms.

In Table 2.1 a summary of the CARP variants discussed in this chapter, and the input parameters used for each variant is shown. Road networks are split up into Actual Road Network (ARN), Random Network Generation (RNG) and Public Benchmark Instance (PBI). In most cases where PBI are used, other parameters are not explicitly discussed, as they are included in the PBI. Furthermore the table contains summaries of waste generation rates, deadheading and service velocities, dumping cost, vehicle capacity and time durations used, where applicable.

Literature shows that estimating traversal and deadheading times can possibly be achieved by using public street networks and waste vehicle GPS data. Synthetic Populations from publicly available census data, along with existing per capita waste generation estimates also present an opportunity to estimate waste generation rates per street segment. Finally, GPS data can be used to estimate landfill and transfer station visit durations. It is therefore likely that all the MCARPTIF variables can be successfully estimated and realistic collection routes be produced to reduce waste collection costs.

Table 2.1: Input parameters reported for CARP variants in literature

Article	PT[a]	RN[b]	PBI[c]	WGR[d]	DHV[e]	SV[f]	DC[g]	VC[h]	TD[i]
Belenguer et al. (2006)	MCARP	RNG	-	10-400kg per segment	20km/h	Traversal cost + 10s per 10kg bin loading time	300	NA	NA
Bautista et al. (2008)	MCARPTC	ARN	-	Population Estimate	Not clear	Not clear	Not clear	$25m^3$	-
Ghiani et al. (2001)	CARPIF	PBI, RNG	Benavent et al. (1992)	Not clear	Not clear	Not clear	Not clear	Not clear	NA
Polacek et al. (2008)	CARPIF	PBI	Ghiani et al. (2001), Ghiani et al. (2004)	Not clear	Not clear	Not clear	Not clear	Not clear	NA
Ghiani et al. (2004)	CARPTIF	PBI	Benavent et al. (1992)	Not clear	Not clear	Not clear	Not clear	Not clear	-
Ghiani et al. (2010)	CARPTIF	PBI	Benavent et al. (1992)	Not clear	Not clear	Not clear	Not clear	Not clear	-
Willemse and Joubert (2016b)	MCARPTIF	ARN, PBI	Ghiani et al. (2004)	Household Estimate	28km/h	14km/h + 1s per kg of waste	300s	10 000kg	8hrs
Willemse and Joubert (2016a)	MCARPTIF	ARN	-	10kg per household, one household every 20m	20km/h	20km/h + 10s per 10kg bin loading time	300s	10 000kg	8hrs
Laporte et al. (2010)	CARPSD	PBI	Ghiani and Laporte (2000)	Poisson Distributions	Not clear	Not clear	NA	NA	NA
Christiansen et al. (2009)	CARPSD	PBI	Belenguer and Benavent (2003)	Poisson Distributions	Not clear	Not clear	NA	NA	NA
Kiilerich and Wøhlk (2018)	CARP	ARN	-	Actual Waste Data	Not clear	Not clear	Not clear	Various Capacities	Various Time Durations

[a]Problem Type
[b]Road Network
[c]Public Benchmark Instance
[d]Waste Generation Rate
[e]Deadheading Velocity
[f]Service Velocity
[g]Dumping Cost
[h]Vehicle Capacity
[i]Time Duration

Chapter 3

Landfill visit analysis

This chapter contains work presented by Steyn and Willemse (2018). Landfill visit durations are an important part of the Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities (MCARPTIF) input parameters and represent the cost, in time, of dropping waste off at a landfill or transfer station. The more time spent at a landfill or transfer station, the less time there is to perform value adding collection activities. The aim of this chapter is to show that landfill visit durations can be determined using GPS data and publicly available geographical data and that these durations are stochastic in nature.

3.1 Input data

The GPS data used in the analysis of vehicle transfer station behaviour is over a period of nine months during the year 2014. The data was extracted from 571 vehicles in a waste collection fleet in a South African Metropolitan area and consists of 48 million GPS records. Each GPS record has the following attributes: time, longitude, latitude, vehicle ignition status and vehicle registration. Using this raw data, a SQLite database was constructed to expand, analyse and store this data. By extracting a single vehicle's data on a particular day from the database, vehicle activities can be visually identified. Figure 3.1 shows the GPS traces of a single vehicle collecting waste from a particular area and disposing waste at a transfer station on multiple occasions throughout a day.

When more than one vehicle is plotted simultaneously the behaviour of a typical waste collection fleet becomes more apparent. Figure 3.2 shows three vehicles that all collect waste and then head to the transfer station at more or less the same time to dispose waste. The purpose of the figures is to illustrate how the GPS data can be used to characterise vehicle activity visually.

Landfill and transfer station locations are generally public knowledge, and all of the landfill and transfer station locations used in the analysis were publicly listed on the Metropolitan area's website. However, only a GPS location and address for each landfill and transfer station is provided. To effectively analyse vehicle behaviour in and around transfer stations a geofence is required. This is a polygon that denotes the boundaries of the transfer station. To generate the polygons *Google Earth Pro* was used to survey each transfer station. On satellite view the boundaries of the transfer stations can be identified and drawn as polygons with ease. The vertices of these polygons, as shown in Figure 3.2 then represent each landfill or transfer station during analysis.

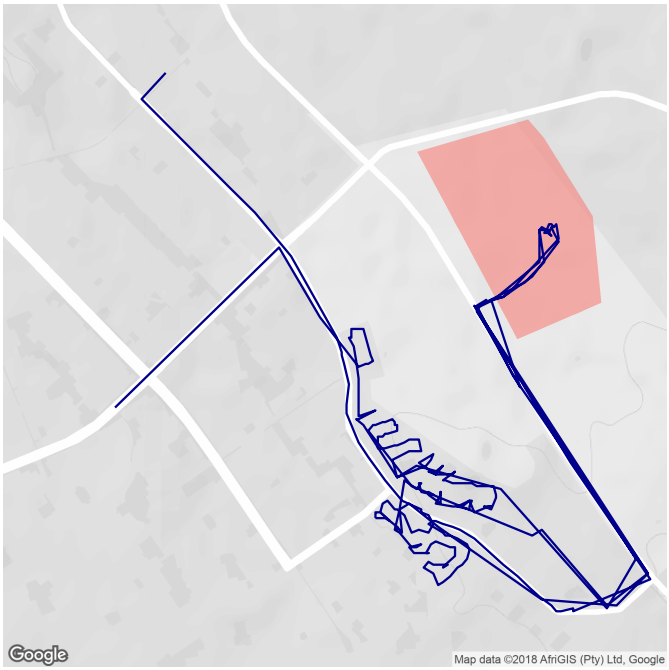

Figure 3.1: GPS traces of a single collection over the course of a day

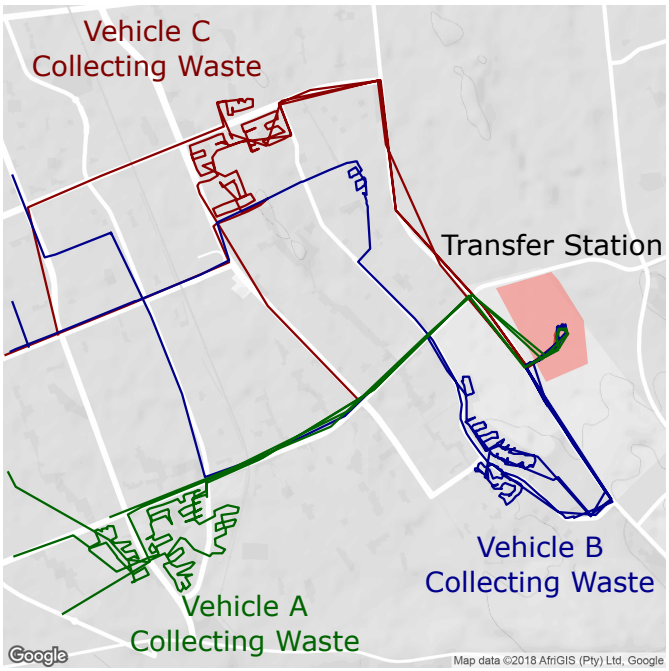

Figure 3.2: GPS traces of multiple vehicles over the course of a day, with the transfer station area shown in light-red.

3.2 Analysis

The analysis discussed in this chapter delves into identifying transfer station or landfill visits and subsequently determining offload durations per visit as well as looking at congestion as a potential variable in extending transfer station visit durations.

3.2.1 Identifying landfill or transfer station visits

Given the above-mentioned input data, the next step was to accurately identify when a vehicle visits a particular landfill or transfer station. For the purposes of the analysis, a single transfer station in a densely populated area was selected. To identify visits to the transfer station all Global Positioning System (GPS) points on a particular day are analysed per vehicle. The times of the points that fall within the geofence facility are then compared against the times of those that fall outside the geofence. The GPS points of a vehicle are scanned in order of their time-stamps. Starting with $i = 1$, when a point outside the facility is followed by a point inside the facility, the time of the inside-point is taken as the start of facility visit, i. The rest of the points are then scanned and when a point inside the facility is followed by a point outside the facility, the time of the outside-point is taken as the end time, e_i, of facility visit i. The index i is then incremented by one and the process repeats for all points of the vehicle, for all vehicles on the day, and for all days in the study period. The duration of a visit, d_i, is then calculated as:

$$d_i = e_i - a_i \tag{3.1}$$

3.2.2 Number of vehicles within transfer stations

The number of vehicles within a transfer station could affect the duration of waste drop-offs as vehicles might have to compete for resources such as weigh bridges. To measure congestion in a facility, the number of vehicles already within the transfer station was calculated for each vehicle arrival. Let vehicle visit i have an arrival time a_i and exit time e_i at a facility and let all the visits to the facility on the same day be $j \in \boldsymbol{V}$. All the visits j to the facility that do not overlap with i are those where $a_i > e_j$ or $e_i < a_j$. All other visits overlap, and the number of these visits represent the number vehicles in the facility when visit i starts.

3.3 Results

With vehicle visits to the transfer station identified, visit durations calculated and the number of vehicles within the transfer station upon a vehicles arrival determined, results are presented and interpreted in the next sections.

3.3.1 Drop-off durations

The most important piece of information missing from routing optimisation studies that is addressed in this chapter is drop-off durations at landfills or transfer stations. In literature waste collection drop-off times are generally considered constant. However, this may not be a fair assumption. Using GPS data and geofences this can be tested.

The distribution for drop-off durations for 31 vehicles, resulting in 3053 unique visits at a single transfer station from the GPS dataset over a nine month period is displayed in Figure 3.3. The mean duration that vehicles spend within the facility is 16 minutes.

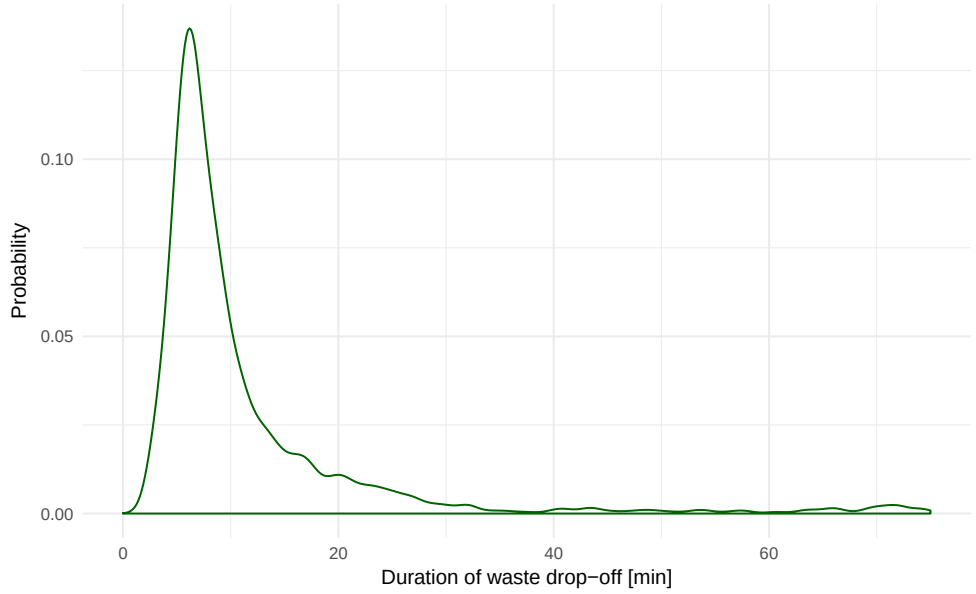

Figure 3.3: Probability distribution of drop-off durations at a single transfer station

Drop-off time is right skewed with a median duration of 8 minutes and interquartile range of 6 minutes indicating that visit duration is variable. However, observing that drop-off duration is variable is not sufficient as future routing applications will require a distribution of drop-off values from which to sample observations. For this reason an attempt is made to fit drop-off durations to a distribution and to estimate its parameters. A good starting point is the Cullen and Frey graph from the *fitdistrplus* package in **R** by Delignette-Muller and Dutang (2015). The Cullen-Frey graph for seven common distributions compared against drop-off duration is shown in Figure 3.4.

The interpretation of the graph is that the closer the bootstrapped values are to the theoretical distribution, the better the distribution fits the data. Except for the beta distribution, none produce an adequate fit.

Next, a power-law distribution was considered and the approach based on Clauset et al. (2009) were implemented. First, the parameters for the power-law distribution α and $x_{\min}$ were estimated using the Maximum Likelihood Estimation method. The parameters were estimated as $\alpha = 5.16$ and $x_{\min} = 74.35$. To test the goodness of fit for the power law distribution, the Kolmogorov Smirnov test (KS) was performed to see if generated data from the distribution, with the above parameters, and the observed data come from the same distribution. To do so a large number of synthetic data sets of the power law distribution with $\alpha = 5.16$ and $x_{\min} = 74.35$ are generated. For each data set, the KS test was then performed between the synthetic and observed data. A significance level of 0.10 was used in the tests, meaning that if the p value from the KS test was less than 0.10, we would reject the null hypothesis and conclude that the two datasets do not come from the same distribution. If the p value is greater than 0.10 we fail to reject the null hypothesis and conclude that the data may come from the same distribution. With the significance level of 0.10, we expect to reject the null hypothesis for about 10% of our synthetic data sets, should the data sets come from the same distribution. If this is the case, the power law distribution can be considered a good fit. A total of 2500 synthetic data sets were generated, on which we failed to reject null hypothesis 1882 times, or approximately 76% of the time.

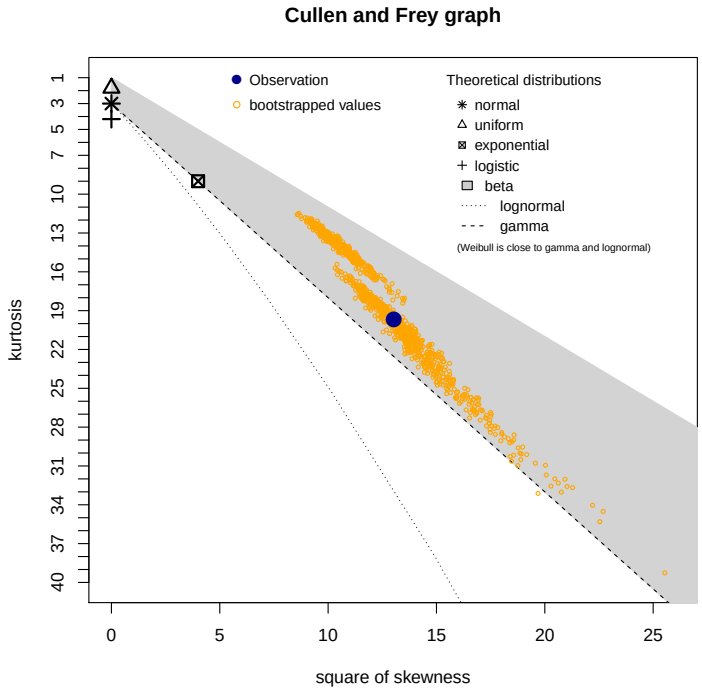

Figure 3.4: Cullen and Frey graph for drop-off durations

It is important to note at this point that a power law distribution, with $\alpha = 5.16$ and $x_{\min} = 74.35$, excludes the bulk of the data. Figure 3.3 shows that the data set has a median of 8 minutes. With the power law distribution only producing a good fit in drop-off durations exceeding 74 minutes, it does not adequately describe the behaviour of the entire data set and can therefore not be considered a good fit of drop-off durations. Attempts at fitting various other distributions for observations below 74 minutes also failed.

Although fitting the data set to a distribution would have been a good outcome, it would likely have involved identifying and isolating confounding variables that affect drop-off duration that are not visible in the data. An example would be the number of staff available at the facility on a particular day to facilitate drop-offs. In addition, distributions are likely to differ between facilities with different layouts and equipment configurations, the aim is therefore simply to demonstrate that a drop duration observations can be collected using GPS data.

3.3.2 Number of vehicles within transfer station

The next piece of information that can be extracted from the GPS data is vehicle arrivals. Are arrivals constant (meaning they are evenly spread through the day) or do they vary with time, and if so why? Figure 3.5 shows when vehicles typically arrive at the transfer station over the sample period by categorising arrival times into 30 minute intervals. For example, approximately 210 visits were recorded between 09:00 and 09:30. Arrivals are not evenly spread throughout the day and instead there is a peak arrival rate between 10:00 and 11:00 in the morning. Since vehicles leave the depot at more or less the same time every morning it is possible that vehicles reach capacity at a similar time and return to landfills and transfer stations on mass, leading to congestion and delays during the peak visit times seen in Figure 3.5.

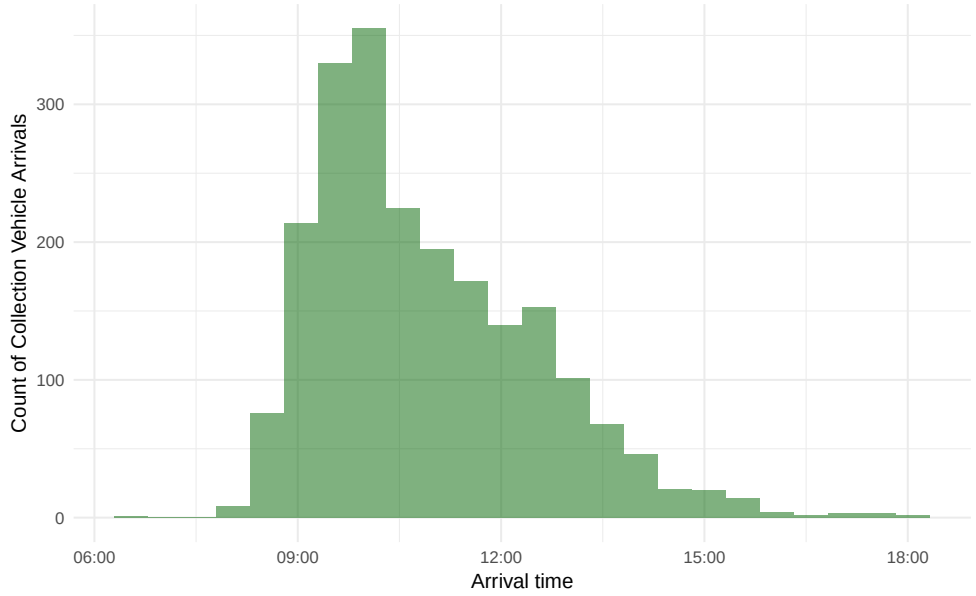

Figure 3.5: Histogram of the number of waste vehicle arrivals per 30 minute interval over the course of a day

Since a peak arrival time has been identified, is it possible that congestion within transfer stations caused by many vehicles arriving in short succession could lead to lengthy drop-off times?

To test this the next variable extracted from the data is the number of vehicles within the transfer station. With each vehicle's arrival, the number of vehicles already present within the facility is measured using the method described in section 3.2.2. Figure 3.6 shows the probability distribution for the number of vehicles already at the station upon a vehicle's arrival. For example, there is a 0.115 probability that when a vehicle arrives at the station that there will be no other vehicles in the station; there is a 0.185 probability that there will be one other vehicle at the station, etc. The figure shows that the highest probability, at 0.185, is for a vehicle to encounter one other vehicle at the station.

To test whether a relationship between transfer station congestion and drop-off duration exists, Figure 3.7 shows the number of vehicles already in the transfer station against the duration of the visit for each visit at the station. Upon inspection of the figure there seems to be two different duration behaviours. To the left, durations seem to be randomly distributed when there are less than 13 vehicles within the facility. However above 13 vehicles drop-off durations are consistently longer. It would therefore appear as if the transfer station reaches capacity in terms of servicing vehicles when around 13 vehicles are present. The same pattern emerges of a sharp increase in drop-off duration at around 26 minutes, although there are far fewer samples to substantiate this with confidence. When these two populations are isolated (drop-off durations below and above 13 vehicles), Figures 3.8 and 3.9 show that below 13 vehicles present at arrival the number of vehicles has little or no effect on drop-off duration. However, if there are above 13 vehicles present at arrival the duration becomes significantly longer. Of the 3053 transfer station visits initially observed, only 135 had vehicles enter the facility with more than 13 vehicles already present. This would indicate that it is an abnormal event and not part of daily operations. For this reason it is unlikely that there is congestion leading to extended drop-off durations at the facility under consideration.

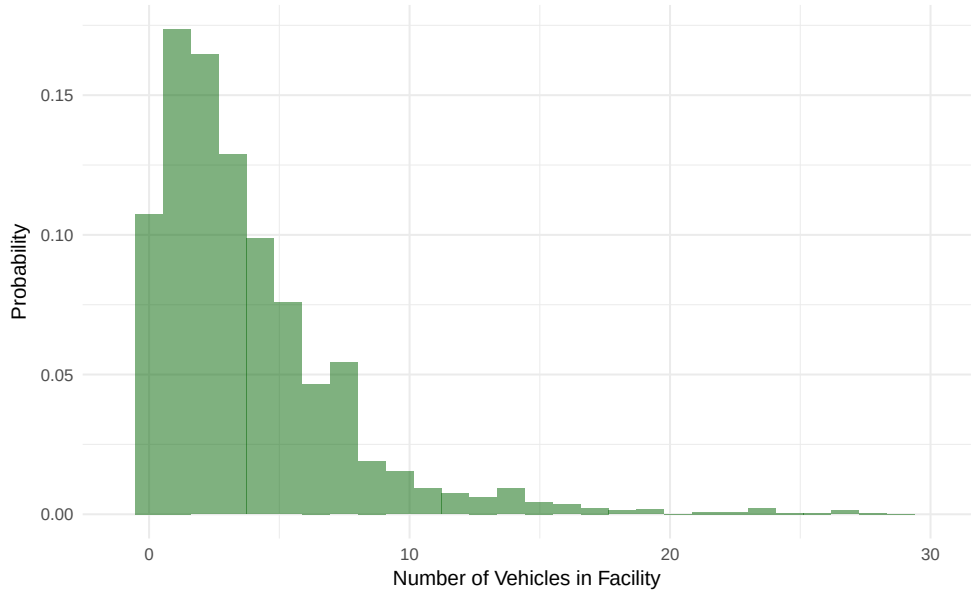

Figure 3.6: Distribution of number of vehicles in transfer station

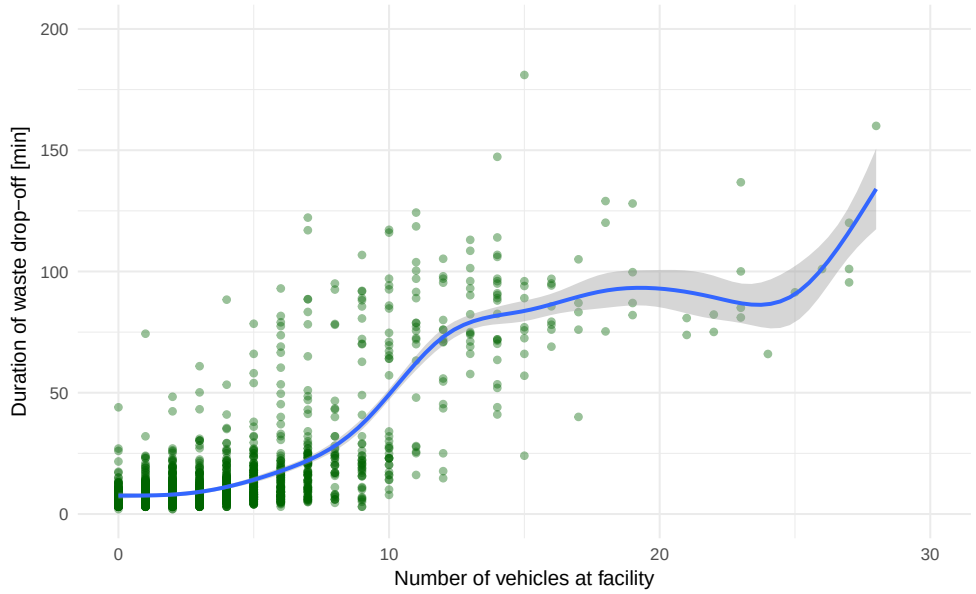

Figure 3.7: Vehicle waste drop-off durations vs number of vehicles in the transfer station

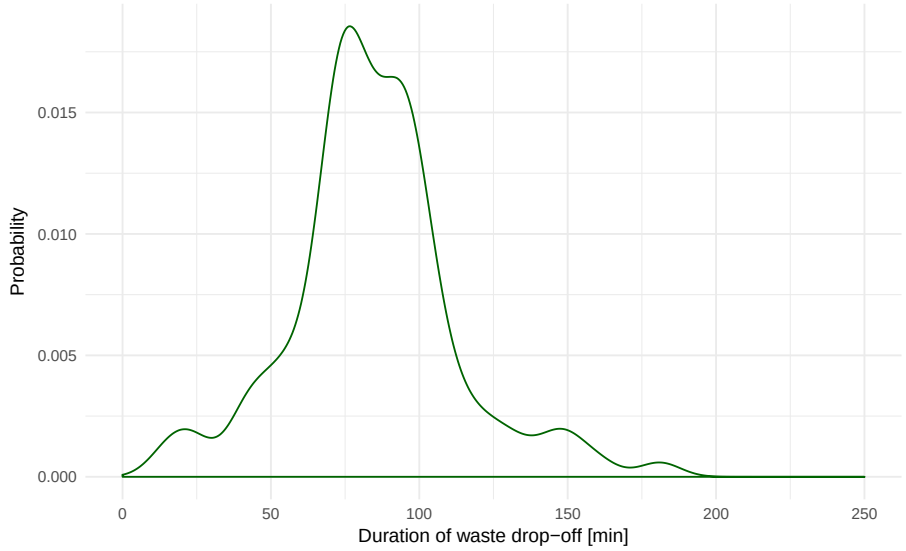

Figure 3.8: Distribution of drop-off durations when there are more than 13 vehicles within the transfer station

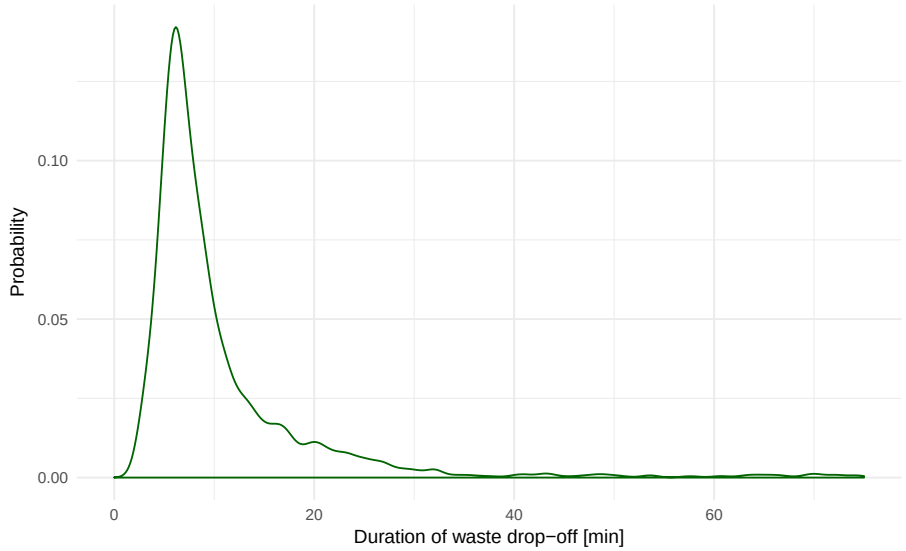

Figure 3.9: Distribution of drop-off durations when there are less than 13 vehicles within the transfer station

3.4 Conclusion

Using GPS data to conduct analysis on waste vehicle behaviour is quick, simple and reproducible. The only requirement is a fleet fitted with GPS devices, which is becoming more common. In this chapter it is demonstrated that waste collection vehicle offloading times at landfills are variable and can be estimated using vehicle GPS data. Furthermore, the potential impact of congestion at the facility on duration was demonstrated, where is was shown that with more than thirteen vehicles in the facility, drop-off times increased. The benefit of the result is that drop-off durations for all facilities within a metropolitan area can be estimated and incorporated into a new Capacitated Arc Routing Problem (CARP) variant where drop-off durations differ depending on the selected facility. It might, for example, be beneficial to the overall route length to travel a slightly longer distance to a facility with a lower drop-off duration than to visit the nearest facility where drop-off durations might be longer. By incorporating this more nuanced approach to drop-off durations overall route quality might be improved.

Chapter 4

Segment traversal analysis

A key data input for the Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities (MCARPTIF) algorithms are service and traversal costs. To accurately determine the cost of servicing a particular area, service costs and traversal costs, in terms of time, must be accurately established at a street segment level. The reason this is important is that the MCARPTIF requires a time estimate for the traversal of each segment. The time it takes to traverse a segment is a function of the vehicle velocity over the segment and the segment length. However there is additional complexity. When traversing a segment *and* collecting waste the vehicle will traverse the segment at a lower velocity, since the vehicle is stopping periodically to collect waste bins left on the curb. The question then is how much slower will the vehicle travel? Assuming that the vehicle will traverse all segments at the same velocity is therefore also not a fair assumption. What if a particular stretch of road has more speed bumps or intersections, resulting in the vehicle travelling at a lower velocity. When taking these factors into account it is clear that for accurate MCARPTIF routing solutions vehicle velocity estimates are required per segment both when traversing and deadheading street segments.This chapter aims to address that data requirement.

4.1 Input data

The two important input data sources for analysis in this chapter are the street network, which is publicly available from a number of sources such as *Google Maps* or *OpenStreetMap*, as well as the Global Positioning System (GPS) data, which is not publicly available. These two input data sources, as well as any preprocessing applied to them, is described in more detail below.

4.1.1 GPS data

GPS data consists of a record ID, unique to each point, a time stamp, longitude and latitude fields, ignition status, service day and a vehicle ID. See Table 4.1 for an extract of the GPS data. The time stamp is the time that the GPS point was produced, with a corresponding longitude and latitude. The ignition status is whether the vehicle engine was switched on that point in time, this is not used in any analysis in this dissertation. The DateID and vehicleID's simply identify the date that the GPS point was produced on, and the vehicle where it originated.

The total GPS data set consists of 48 million records, made up of 521 days and 787 vehicles. The data was collected over a nine month period in 2014 from a waste collection

Table 4.1: GPS Data extract

RecordID	time	long	lat	ignitionStatus	ServiceDateID	WasteVehiclesID
24555	10:00:03.0000	18.67863	-33.81490	T	44	377
24561	10:00:07.0000	18.67860	-33.81470	T	44	377
24740	10:00:35.0000	18.67860	-33.81453	T	44	377
24983	10:00:36.0000	18.67860	-33.81453	T	44	377
24984	10:00:40.0000	18.67860	-33.81437	T	44	377
25992	10:02:08.0000	18.67867	-33.81410	T	44	377
25997	10:02:10.0000	18.67867	-33.81400	T	44	377
26159	10:02:51.0000	18.67867	-33.81383	T	44	377
26207	10:02:55.0000	18.67850	-33.81373	T	44	377
29907	10:05:30.0000	18.67750	-33.81383	T	44	377

fleet in a metropolitan area in South Africa. The data set contains 144059 unique day and vehicle combinations. The mean number of records per vehicle per day is 332 records. The metropolitan area is made up of 960 beats, or service areas, which are serviced once a week by a waste collection vehicle. Figure 4.1 shows the metropolitan area, with beats drawn as grey polygons. Data on the collection beats are publicly available from the metropolitan area's website, and are imported as shape files.

Figure 4.1: Metropolitan service areas

The frequency with which GPS points is recorded is crucial to data accuracy. In the full data set the time interval between consecutive GPS points has median of 59 s and mean of 156 s. The first and third quartiles are at 2 s and 60 s. Since the mean is above the 3rd quartile and the distribution contains extreme values the median is the most appropriate summary statistic according to Ross (2017) and therefore it seems that GPS points are produced every 59 s by vehicles.

To efficiently store and access data, the GPS data as well as beat data is stored in a *SQLite* database. For the full entity relationship diagram, see Figure 4.2. The database is broadly divided into tables containing information on the collection fleet, the service areas, and the GPS points themselves.

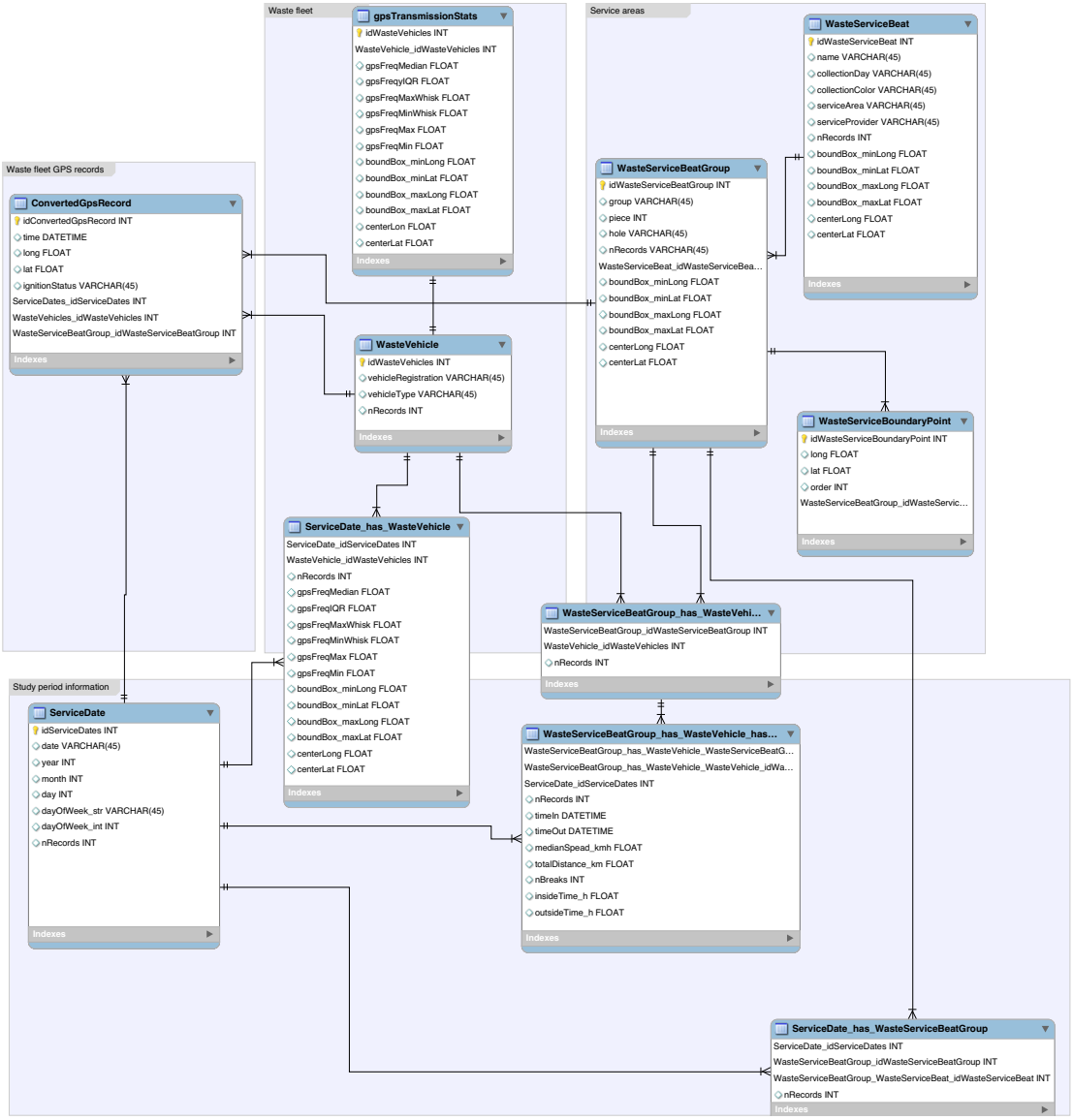

Figure 4.2: Entity Relationship Diagram of SQLite database containing GPS and beat data

For the purposes of demonstrating the techniques applied in this chapter, a case study service area was selected. The techniques presented can be replicated on any service area within the metropolitan area, and results on other service areas are discussed in the results chapter, however for the moment we will focus on a single area. There are 20923 GPS points that fall within this case study area, over a period of 224 days. The area consists of 331 street segments. Figure 4.3 shows the sample of GPS records within the area. Subsequent analysis will be based on the GPS points within this service area, before being expanded to include other areas.

Figure 4.3: Case study service area with vehicle GPS points

4.2 Network construction

The first step to analysing how vehicles traverse street segments is to build a road network. To this end *OpenStreetMap* (OSM) provides open source street maps for public use, and developed through public contribution. This is because Capacitated Arc Routing Problem (CARP) models require a network on which to perform routing. A network consists of nodes, edges and arcs. When importing an OSM network a network similar to the one in Figure 4.4 is produced. The figure shows a series of nodes and links, however for the purposes of vehicle routing and the subsequent analysis the network must be simplified. What is required for vehicle routing applications is a network consisting of intersections, with street segments that link intersections. We then calculate the routing parameters (demand, service and deadheading cost etc.) per segment.

Careful consideration of Figure 4.4 reveals nodes on segments, as well as on intersections. The nodes between intersections are called interstitial nodes, and show the topography of the line segment or street segment. For the most part the interstitial nodes are not relevant to routing, since all we need to know is which segments are connected to which intersection (i.e. where the vehicle can travel next from the current intersection). The actual topography of the segment is not important except for the length of the segment, which determines the servicing and deadheading cost. For this reason we remove interstitial nodes when constructing the routing network. We calculate the segment lengths at a later stage. In addition, what appears to be interstitial nodes can cause computational errors when solving CARP instances since they are not necessarily connected to the network and might represent topography other than road segments. The result being that there are nodes in the network which are impossible to access. To solve this the OSMnx package developed by Boeing (2017) is used to identify only required nodes (at intersections), and to then simplify the network into graph form for further analysis.

Figure 4.5 shows the interstitial nodes, in red, which are to be removed. Once these nodes have been removed Figure 4.6 is produced, which shows the simplified network. Note that the original network topography has been maintained in the network file, but that interstitial nodes are no longer recognised as nodes for the purposes of analysis.

Figure 4.4: *OpenStreetMap* network

Figure 4.5: *OpenStreetMap* network, non required nodes are displayed in red.

Figure 4.6: *OpenStreetMap* network, simplified.

4.3 Point snapping

Once a viable network has been created from the OSM topography, the next step is to snap GPS points to street segments in the network. GPS points contain natural jitter, and as a result coordinates do no correspond precisely to the line segments in the OSM network. For this reason GPS points need to be snapped to the network, so that each point can be associated with a particular segment in the network. The purpose of this is to associate the time stamp on a particular GPS record to a street segment, and in doing that estimate entry and exit times and segment traversal durations.

We perform the actual point snapping using the Maptools Package in R, developed by Bivand et al. (2019). Practically, what this entails is that both the road network as well as GPS points are converted to the same coordinate reference system. The point snapping algorithm then determines the nearest line segment to each point. Following this the GPS point is shifted to the nearest point on the line segment. The result is that all points in the vicinity of the line segment are snapped to the line segment and assigned to that particular segment. It is therefore critical to assess the distance with which points are shifted. Points that are not close to any line segments, and are therefore shifted a large distance, are likely to be the result of GPS jitter and could be disregarded.

However, Figure 4.7 shows that the large majority of points are shifted relatively small distances. The mean shift in GPS point is 5.09 metres, with a median of 4.17 metres and a maximum of 38.07 metres. In South Africa, single carriageway lanes are between 3.5 metres and 3.7 metres in width, according to SANRAL (2009). A single carriageway with lanes in both directions, as is the norm in residential areas, will therefore be between 7 and 7.4 metres in width. Given this context, and the fact that GPS devices do produce jitter, the distances that GPS points are shifted to correspond with line segments is reasonable and therefore no outliers are removed at this stage.

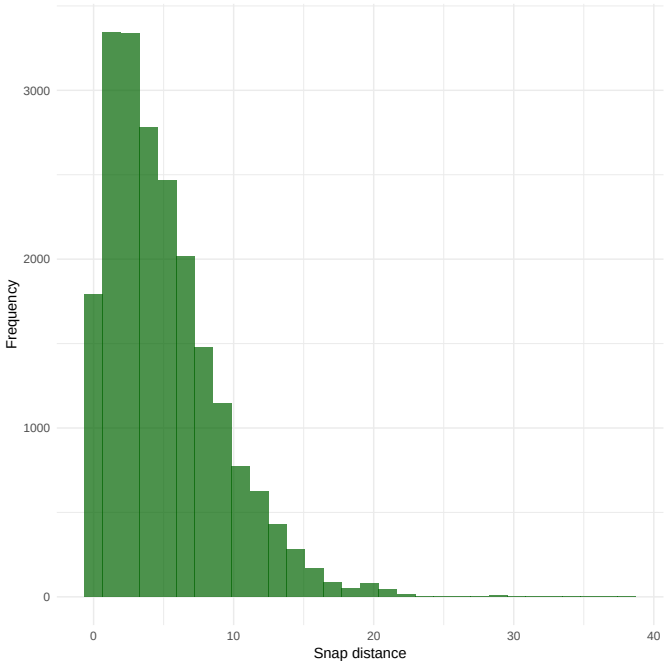

Figure 4.7: Histogram of snap distance, in metres.

4.4 Outlier detection

Having snapped GPS points to the network, and validated the snap distance, the next step is to detect outliers which might skew analysis down the line. Velocity provides a good opportunity for outlier detection as realistic vehicle velocities are known. By calculating velocity, anomalous points can be clearly identified. Anomalies are produced by GPS devices either by recording incorrect time stamps or incorrect locations. Either of these errors can be detected by considering velocity since any point in a time ordered series of consecutive GPS points can only be within a realistic distance from the previous point (i.e travel to the next point at a realistic velocity). If either the geographic location or time stamp is significantly incorrect it will appear as a velocity spike.

For this reason instantaneous velocity is calculated at consecutive points by calculating the change in euclidean distance over time. For the purposes of outlier detection euclidean distance was used. The reason euclidean distance is used for outlier detection is because it provides a more conservative estimate of velocity, since the velocity between two points will be lower *as the crow flies* than on the actual road topography between the two points. A more conservative velocity estimate at this point in the analysis means less points are disregarded. At a later stage, when traversal velocities are calculated the distance along the line segment is used for improved accuracy.

To remove velocity outliers the Inter Quartile Range (IQR) method is a potential option. An outlier is defined as an observation above the 75th or below the 25th percentile, by a factor 1.5 times the inter quartile range. Prior to outlier detection the mean vehicle velocity was 52 km/h with a minimum of 0 km/h and a maximum of 1761 km/h. Since the maximum velocity is a physical impossibility outliers are undoubtedly present in the data. Using the IQR outlier detection the mean is reduced to 5 km/h with a median of 2 km/h and a max of 43 km/h. The outlier detection using velocity reduces the total number of GPS points within the service area from 20923 to 17300. By using the IQR method however the velocity distribution is significantly altered and the maximum velocity

remains within the realistic range of a waste collection vehicle ($0 - 100$ km/h).

A more practical option therefore is to exclude velocities that are physically improbable, since the IQR method might exclude velocities that are statistically outliers but still practically possible. For this reason points with velocities above 100 km/h were excluded. The resulting data set then has a median velocity of 1.619 km/h and a mean of 7.66 km/h. Using this approach the total sample size is reduced from 20923 points to 18350 points. Outlier detection using the velocity cut-off is automatically applied to all service areas.

Following this we now have a series of vehicles GPS points, associated with a line segment and with outliers removed. This is described visually in Figure 4.8, which shows the GPS data, overlayed onto the street network and clearly shows the effect of the snapping process on the GPS data.

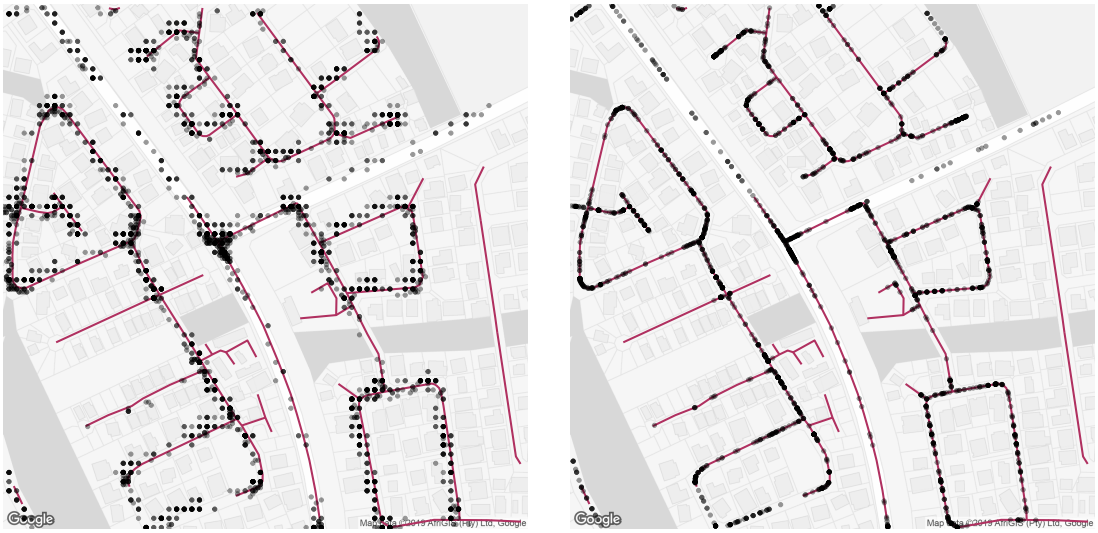

Figure 4.8: GPS points before (left) and after (right) point snapping

4.5 Segment visit identification

Once points have been snapped to the road network, and each point is associated with a particular road segment, segment visit data can be extracted for further analysis. The key variables required for each segment are:

1. When did a vehicle arrive at a particular segment?
2. Where on the segment was the vehicle arrival detected?
3. When did a vehicle depart from a particular segment?
4. Where on the segment was the vehicle departure detected?
5. How long was the vehicle within the segment?
6. How much distance did the vehicle traverse while on the segment?
7. In which direction did the vehicle traverse the segment?

By calculating these variables for each segment the MCARPTIF variables such as traversal velocity can be calculated. The first step of extracting segment visit data is arranging the GPS data according to day, vehicle and time. This means GPS points are first grouped by day, then by vehicle, and finally they are ordered on the GPS time stamp from earliest to latest within the day. Since GPS points are snapped to a particular line segment, a group of consecutive points, all assigned to the same line segment, indicate a visit to that particular line segment. This process is described in detail in Algorithm 1.

It follows that the first and the last of that group of points can be considered *entry* and *exit* points on the line segment. By looking at these points, a lot of information about that particular segment visit can be inferred. The first is that the vehicle likely travelled from the entry to the exit point, thus direction of travel is established. While it is possible that the vehicle could have made a u-turn within the segment, this would have likely occurred infrequently since the vehicles are travelling in residential areas where the roads are narrow and u-turns are difficult. Therefore the assumption is made that travel direction is always from entry to exit point (or first to last point).

Input: GPS data frame
Output: Data frame of segment visits
days = unique days;
arrivalTime ← vector;
departureTime ← vector;
vehicleID ← vector;
segmentID ← vector;
for $d \in days$ **do**
  vehicles = unique vehicles on day d;
  for $v \in vehicles$ **do**
    points = points on day d from vehicle v;
    if $length\ of\ points \leq 30$ **then**
      BREAK
    end
    else
      order points by time ;
      calculate time difference between points;
      remove points with time difference $\leq 1s$;
      changes = vector of the lengths of runs of equal values in segment ID vector;
      CumalativeChanges = vector of cumulative sum of changes;
      Append arrivalTime, vehicleID,segmentID with point[1] timestamp, vehicle ID, Segment ID;
      for $y\ in\ CumalativeChanges$ **do**
        Append departureTimewith point[y] timestamp;
        Append arrivalTime, vehicleID,segmentID with point[y+1] timestamp, vehicle ID, Segment ID;
      end
      Append departureTimewith point[length(point)] timestamp;
    end
  end
end

Algorithm 1: Segment Visit Identification Algorithm

The amount of time spent within that particular segment can also be determined, as the difference in time between the exit and entry point. To calculate the vehicle velocity within the segment visit, the only missing variable is distance. This is slightly more intricate, the partial distance across the segment is required. In other words, the distance from the first to the last point on the segment must be calculated. Since the street segment is not necessarily a straight line it is not as simple as taking the euclidean distance between the points. To do this a short algorithm for calculating the partial distance over the segment was created. See Algorithm 2 for more information.

In essence the algorithm splits streets segments into its component parts and determines whether a point intersects with each sub-segment. If it does, the algorithm calculates the distance between the point and the node at the end of the sub-segment. If it does not, the algorithm calculates the distance across the sub-segment. The sum of these distances is the partial distance of the segment. The algorithm does this for both entry and exit points and subsequently calculates the partial distance between the entry and the exit point.

Input: Entry Point, Exit Point, LineSegment
Output: Partial distance on segment
Split LineSegment into component sub-segments
for $i \in$ *sub-segment* **do**
  if *Entry Point intersects i* **then**
    EntryPartialDistance = EntryPartialDistance + length of i start to point
  end
  else
    EntryPartialDistance = EntryPartialDistance + length of i
  end
end
for $i \in$ *sub-segment* **do**
  if *Exit Point intersects i* **then**
    ExitPartialDistance = ExitPartialDistance + length of i start to point
  end
  else
    ExitPartialDistance = ExitPartialDistance + length of i
  end
end
PartialDistance = max(EntryPartialDistance, ExitPartialDistance) -
 min(EntryPartialDistance, ExitPartialDistance)

Algorithm 2: Partial distance on segment algorithm

Having calculated the partial distance travelled over the segment, as well as the duration of time that the vehicle spent within the segment, the velocity can be calculated simply as the distance traversed divided by the time spent within the segment.

4.5.1 Traversal speed for short segments

A notable challenge at this point is the time interval between GPS points. In section 4.1.1 the median time interval of 59 s is discussed. Consider for a moment that if a vehicle travels at 20 km/h, or 5.5 m/s, that it will travel a distance of 324 m between consecutive GPS points. At this velocity any segments less than 324 m in length will not have more than two consecutive points, with which to calculate velocity.

The challenge this poses is that it reduces resolution in instances where vehicles only generate a single GPS point per segment. This can be due to segment length, where the segment is very short and the vehicle traverses the segment in less than sixty seconds, resulting in only one GPS point on the segment. It can also be due to vehicle velocity, where the segment is not necessarily short, but the vehicle moves at a velocity such that it passes through the segment in less than sixty seconds. Both cases pose a challenge as a traversal velocity cannot be calculated using only a single point.

In the case study area only 45 percent of segment visits had more than 1 consecutive GPS point, which greatly reduces the sample space. In addition this problem is likely to skew the results, since the sample of GPS points will be skewed towards instances where the vehicle is moving at a lower velocity.

To overcome this a separate algorithm was developed to deal with these cases. The algorithm aims to calculate a traversal velocity for single point segments by making use of the GPS points directly before and after the point in question. The assumption being that the instantaneous velocity through the segment is the same as the velocity of the vehicle just before and after entering and exiting the segment.

The partial distance to the end of the previous section is calculated, as well as the partial distance between the next segment and the next point on the GPS trace. An average velocity over these points is then calculated and assumed to be the instantaneous velocity at the single point segment.

The result of the above steps to extract segment visits from GPS data and street network data is that a total of 5839 unique segment traversals could be extracted from the data, for the service area.

4.5.2 Inferring segment activity

Having snapped points to the road network, identified and listed segment visits, and calculated traversal velocity the next key research question is whether we can identify if a segment is being serviced or traversed. The assumption with this problem is that a waste collection vehicle, travelling within a service area is either collecting waste from a segment, or travelling through the segment (dead-heading) without collecting waste. Since the street segment is the lowest level of analysis the assumption is that a vehicle is either servicing a segment, or deadheading a segment, not both. No allowance is made for partial collection or dead-heading on a single segment. The implications of this assumption might be that there are observations within the data where the vehicle has traversed a portion of the segment before collection activities start (or vice versa) and that the velocity estimate is therefore a hybrid of both collection and traversal. While this is possible, further seperation of vehicle activity would require much less granular data to detect changes in velocity across the segment. These cases are likely also infrequent and the effect of this minimal on the velocity samples. For this reason the street segment as the lowest level of analysis is therefore presumed to be sufficient.

The key research question is then whether an inference can be made as to whether a vehicle is servicing or dead-heading a particular segment when it is passing through that segment, given the available data. The foundational assumption behind inferring whether a vehicle is servicing or deadheading a segment is that the expected service velocity for a segment will always be lower than the expected deadheading velocity. This is practically obvious since a vehicle must both traverse a segment *and* perform the collection activity when servicing. Which would mean that the velocity through the segment is always lower than just traversing the segment. The aim therefore is to identify potential variables that could indicate whether a vehicle is servicing or deadheading a segment, and test whether

the vehicle velocity differs depending on the variable.

A number of potential variables are proposed. These are:

1. When a vehicle visits a particular segment
2. Which vehicle visits a particular segment
3. The sequence in which segments are visited
4. Whether a segment has a residential population to be serviced

By looking at these variables it is likely that a good estimate can be made on whether or not a vehicle is dead-heading or servicing a particular segment. The aim then is to separate the sample of segments visits which were just extracted into servicing and deadheading visits and then subsequently to produce service and deadheading velocity estimates. The reason this is important is to extract both a service and traversal cost for the MCARPTIF model.

4.5.3 Service day

Waste collection within the study area happens once a week, on the same day every week. Since vehicles will only enter a residential area to collect waste on the collection day, on non-service days, all traversals will be deadheading, whereas on a service day, traversals will be a mix of deadheading and service.

The service day is determined by finding the day of the week with the most segment visits. See Figure 4.9 for a count of segment visits per day of week. The figure shows that Tuesday is the service day, with a total of 5016 segment visits. All other days of the week showed less than 300 segment visits. It is therefore highly probable that segments are more likely to be serviced on Tuesdays than any other day of the week.

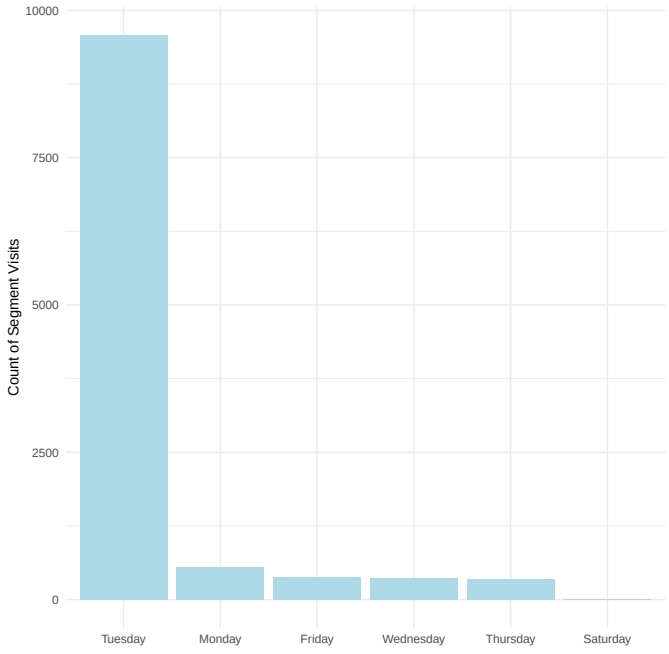

Figure 4.9: Count of segment visits per week day

4.5.4 Service vehicle

The next potential predictor is which vehicle visits a segment. Vehicles are typically assigned to specific beats or service areas. If that vehicle enters the service area it is more likely that segments will be serviced. Figure 4.10 shows that a single vehicle visits the service area a lot more frequently than other vehicles. It is therefore more likely that segments will be serviced by this particular vehicle. This is also practically visualised in Figure 4.11 where GPS points are colour coded by vehicle ID. Each colour is therefore a different vehicle passing through the areas. It is clear from the figure that most of the activity within the service area comes from a single vehicle.

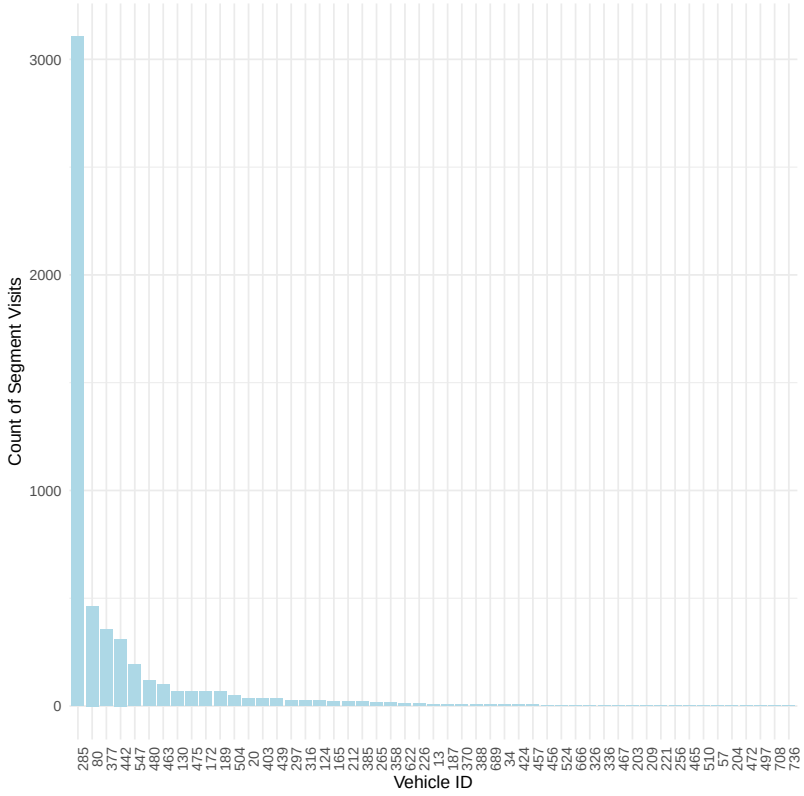

Figure 4.10: Count of segment visits per vehicle ID

Figure 4.11: GPS points grouped by vehicle ID for the service area

4.5.5 Traversal sequence

The next potential variable for determining whether a segment is being serviced or deadheaded is the number of times a segment is traversed on a service day. Here the assumption is that if a segment has been traversed multiple times on a single day, that the earlier traversals are more likely to be as a result of the vehicle servicing the segment. Furthermore if a vehicle only traverses a segment once on a service day, the likelihood is high that the segment was serviced during that traversal. Subsequent traversals are therefore more likely to be the vehicle deadheading the segment to get to areas of the beat not yet serviced. Figure 4.12 shows the count of segment visits against the number of visits per day. As expected the bulk of segments are only traversed once on a service day.

4.5.6 Segment population

Lastly, we can consider whether the fact that certain segments have residential populations on them affect the vehicle velocity. The premise of this hypothesis is that the presence of households means that the segment is more likely to be serviced than not. Figure 4.13 shows that the large majority of segment visits are to segments with populations on them (4229 visits as opposed to 281). This is to be expected since vehicles are performing residential waste collection and will try to minimise travelling through segments where there aren't residential populations to be serviced. The synthesis of the population data used here is described in more detail in Chapter 5.

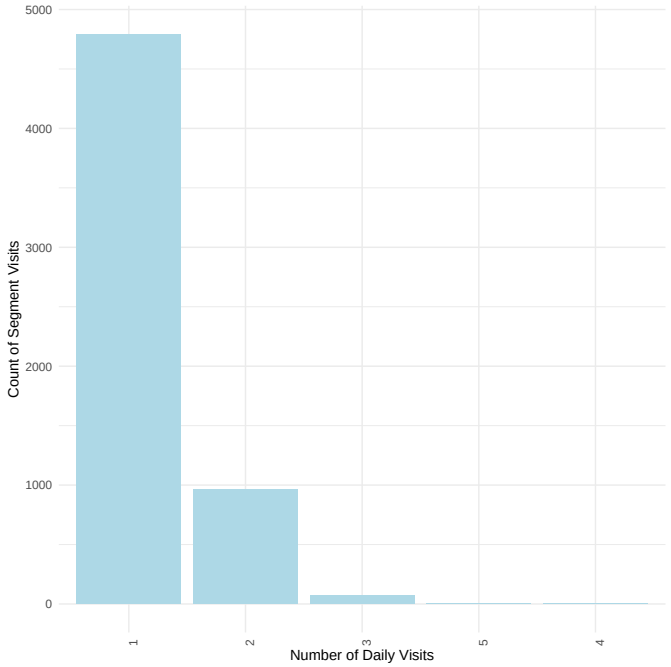

Figure 4.12: Count of number of segment visits

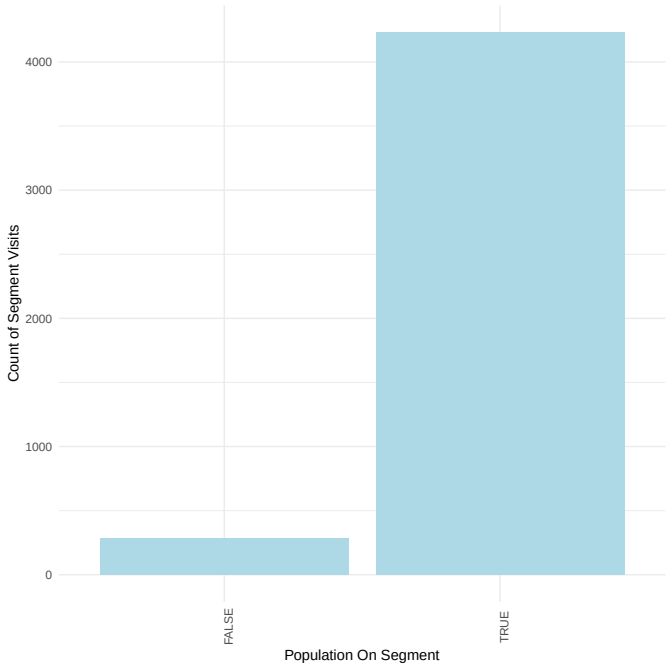

Figure 4.13: Count of segment visits for segments with and without population.

4.6 Results

4.6.1 Overall velocity

Given the above analysis, results can be presented and a number of conclusions reached. The mean velocity for all segment visits over the sample period is 4.67 km/h, with a median of 2.19 km/h. This velocity is consistent with what we know about residential waste collection, that vehicles move at low velocity from residence to residence collecting waste. Even before service and deadheading velocities are extracted in the subsequent section, this is a significant piece of information, since this estimate is lower than velocity assumptions in literature. Figure 4.14 shows the overall velocity distribution within the segment, the blue line represents the median velocity and the red line shows the mean velocity.

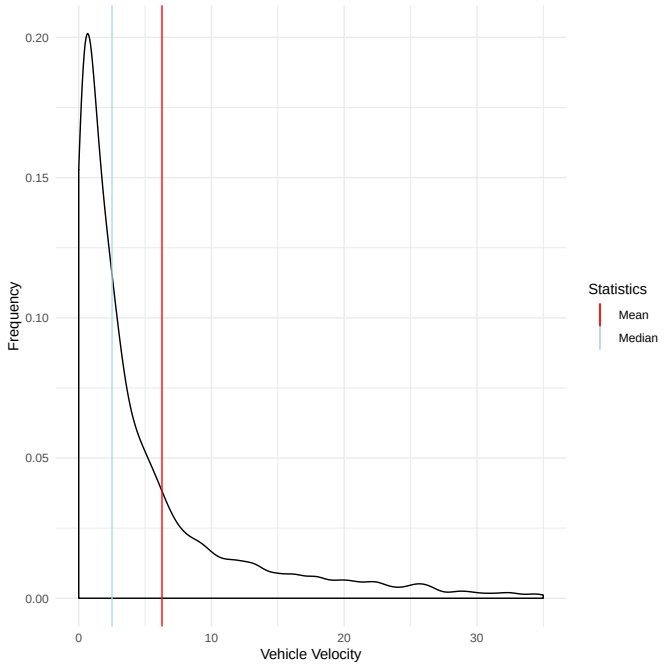

Figure 4.14: Overall vehicle traversal velocity

4.6.2 Service day as predictor

The first variable discussed earlier, which is likely to predict vehicle activity, is the service day. The theory being that when a collection vehicle enters the service area on a collection day that the probability of the vehicle servicing segments will be higher. Figures 4.15 and 4.16 show the distribution of vehicle velocities between service and non-service days. The figures show similar right skewed distributions, but with slightly more observations at higher velocities for non service days. The mean velocity when vehicles enter the area on service days is 4.186 km/h with a median of 2.157 km/h. On non service days vehicles have a mean velocity of 7.619 km/h and median velocity of 2.827 km/h. While the distributions do overlap, there appears to be a difference in velocity between the two samples. This is to be expected as it is unlikely that vehicles are performing collection operations outside of collection days, and therefore traversals outside collection days will be deadheading traversals at higher velocities. Table 4.2 shows summary statistics for vehicle velocity for service days and non-service days.

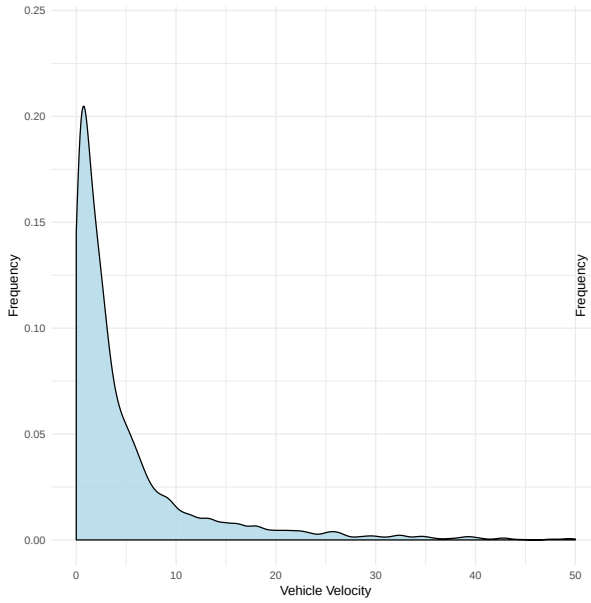

Figure 4.15: Service day vehicle velocity

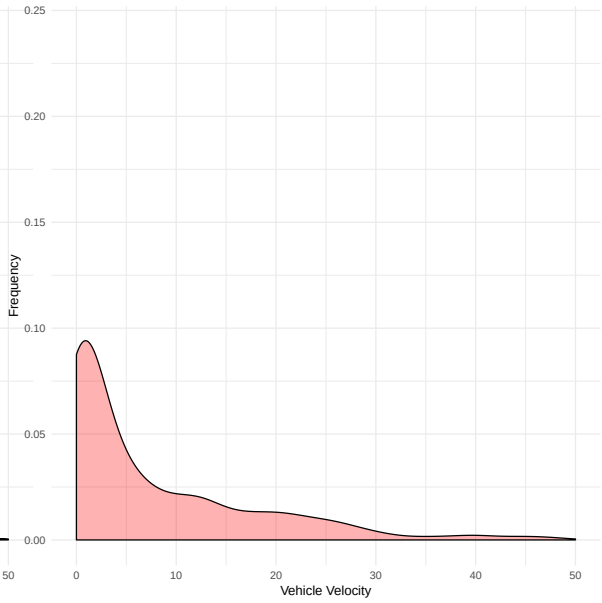

Figure 4.16: Non-service day vehicle velocity

Table 4.2: Summary statistics of vehicle velocity for service and non-service days

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Service Day	0.00066	0.92271	2.15703	4.18673	4.40328	96.39137
Non-Service Day	0.00042	0.51207	2.82695	7.61918	10.78796	92.58369

To formally test this the Wilcoxon Rank Sum Test by Wilcoxon et al. (1970) is used to determine whether the two samples are from two different distributions. The reason the Wilcoxon Rank Sum Test is used is because it is non-parametric, meaning that it does not assume a distribution. Since the distributions are not normally distributed this test is appropriate. The test allows the comparison of two distributions that are not normal, but do have similar shapes, as Figures 4.15 and 4.16 illustrate.

The Wilcoxon Rank Sum Test uses the ranks of samples, as opposed to the underlying values, to test for differences in population median. The hypothesis test is set out as below, where $\tilde{x}_{SD}$ is the median velocity on service days, and $\tilde{x}_{NSD}$ is the median velocity on non service days:

$\mathbf{H_0}$:$\tilde{x}_{SD} - \tilde{x}_{NSD} \geq 0$
$\mathbf{H_a}$:$\tilde{x}_{SD} - \tilde{x}_{NSD} < 0$

The hypothesis test is conducted at a significance level, α, of 0.05. For the purposes of demonstrating how the Wilcoxon Rank Sum Test works it will be explained in detail here, however in subsequent sections only results will be reported. The first step is to rank samples from both populations together, service days and non service days, from smallest to largest. Each observation is assigned a rank. The sum of the ranks is then calculated for the two samples, in this case the sum of the ranks of service day velocities, W_{SD} , is 11982388. The sum of the ranks of the non-service days, W_{NSD}, is 2352948. Thereafter we calculate the mean rank as:

$$\mu_W = \frac{n_1(n_1 + n_2 + 1)}{2} \tag{4.1}$$

Where n_1 and n_2 are the sample sizes for the two samples. The mean rank for this sample is 12131752. The standard deviation of the rank is, 40793.04, and is calculated as:

$$\sigma_W = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1}{12}} \tag{4.2}$$

We then calculate the Z value, using both the mean and standard deviation as:

$$Z = \frac{W_{SD} - \mu_W}{\sigma_W} \tag{4.3}$$

Calculating the Z value using the above we find a Z value of -3.66, which corresponds with a p-value of 0.000125. At a significance level of 0.05 we can reject the Null Hypothesis and accept the alternative hypothesis, that $\tilde{x}_{SD} - \tilde{x}_{NSD} < 0$, or that the median velocity on service days is less than the median velocity on non-service days. We can therefore also conclude that the two samples come from different distributions and that the service day produces different vehicle velocity behaviour.

4.6.3 Service vehicle as predictor

As discussed earlier it is likely that by identifying the service vehicle a velocity sample can be extracted where the probability of a segment being serviced is higher. Figures 4.17 and 4.18 show the velocity distributions for service vehicles and non-service vehicles in the area. While there is a difference it is not apparent that the velocity behaviour differ that significantly. The Service vehicles had an average velocity of 4.209 km/h with a median of 2.160 km/h while non-service vehicles averaged 6.493 km/h and had a median of 2.397 km/h. Table 4.3 shows the summary statistics between the service vehicle and non-service vehicle samples. While the distributions clearly overlap there seems to be a difference in vehicle velocity between the two samples.

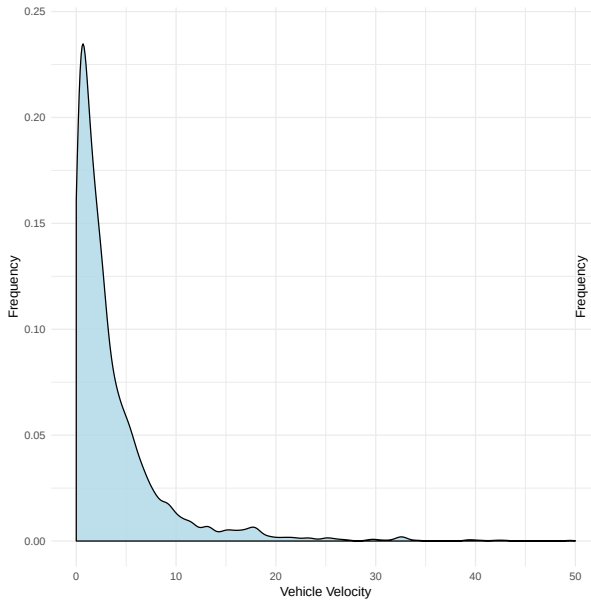

Figure 4.17: Service vehicle velocity

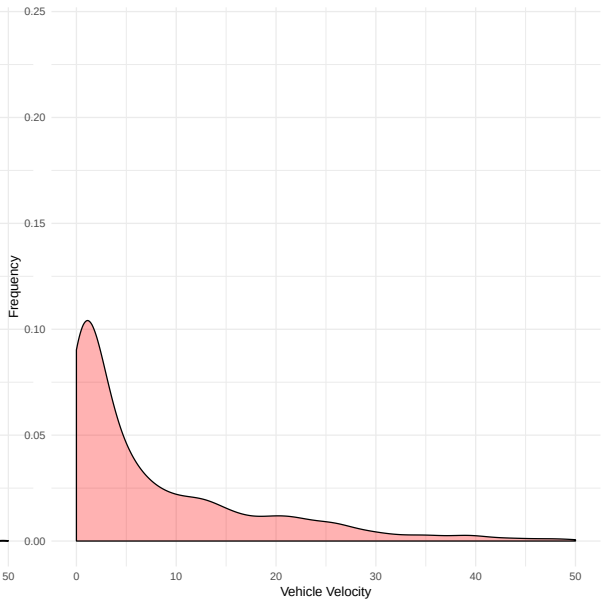

Figure 4.18: Non-service vehicle velocity

Table 4.3: Summary statistics of vehicle velocity for service and non-service vehicles

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Service Vehicle	0.00066	0.92697	2.16025	4.20934	4.40995	96.39137
Non-Service Vehicle	0.00042	0.66592	2.39735	6.49339	7.55055	92.58369

Again, we apply the Wilcoxon Rank Sum Test to test whether the velocities for service vehicles and non-service vehicles come from different distributions. The hypothesis test is set out as below, where $\tilde{x}_{SV}$ is the median velocity for service vehicles travelling through the area, and $\tilde{x}_{NSV}$ is the median velocity for non service vehicles travelling through the area:

$\mathbf{H_0}$:$\tilde{x}_{SV} - \tilde{x}_{NSV} \geq 0$
$\mathbf{H_a}$:$\tilde{x}_{SV} - \tilde{x}_{NSV} < 0$

Applying the Wilcoxon Rank Sum Test in this case yields a p-value of 0.0185, at a significance level of 0.05 we reject the null hypothesis and accept the alternative hypothesis, meaning that there is evidence that the median velocity for service vehicles is lower than the median velocity of non-service vehicles.

4.6.4 Traversal sequence as predictor

Visit frequency could also potentially help predict whether a vehicle is servicing or dead heading a segment. Table 4.4 shows summary statistics of vehicle velocity when a segment is visited for the first time on a collection day, as opposed to the second or third, etc. time. Figures 4.19 and 4.20 show the velocity distributions for first vehicle arrival as opposed to subsequent arrivals. Again there appears to be a difference in median velocity. To formally test this the Wilcoxon Rank Sum Test is again used. The hypothesis test is set out as below, where $\tilde{x}_{1st}$ is the median velocity through segments the first time that segment is traversed on a day, and $\tilde{x}_{>1st}$ is the median velocity through segments for all subsequent traversals on a day:

$\mathbf{H_0}$:$\tilde{x}_{1st} - \tilde{x}_{>1st} \geq 0$
$\mathbf{H_a}$:$\tilde{x}_{1st} - \tilde{x}_{>1st} < 0$

At a significance level of 0.05 we reject the null hypothesis, since the calculated p-value in this case is 0.01. We can therefore conclude that the vehicle velocity, when it arrives at segment for the first time on a day, is from a different distribution than subsequent arrivals at the same segment throughout the day. Arrival sequence is therefore an acceptable predictor of whether a vehicle is servicing a segment.

Table 4.4: Summary statistics of number of visits per segment per day

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Single Visit per day	0.00042	0.75935	1.97303	4.27118	4.24084	96.39137
Multiple Visits per day	0.00066	1.67446	3.55771	6.49830	7.93218	63.45971

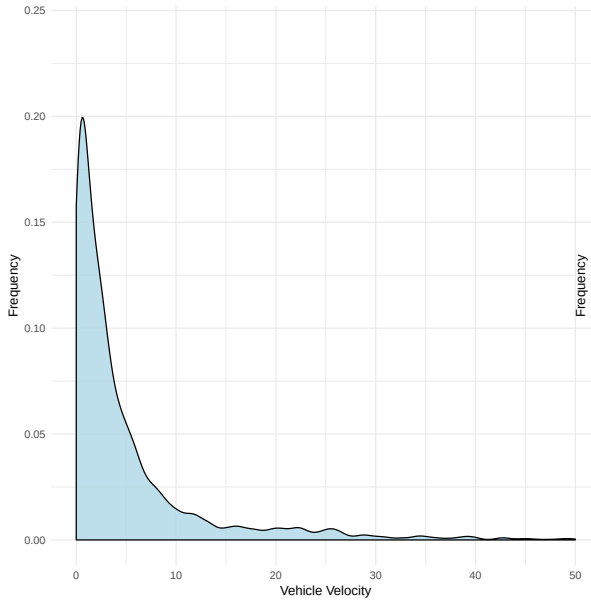

Figure 4.19: Velocity at first arrival

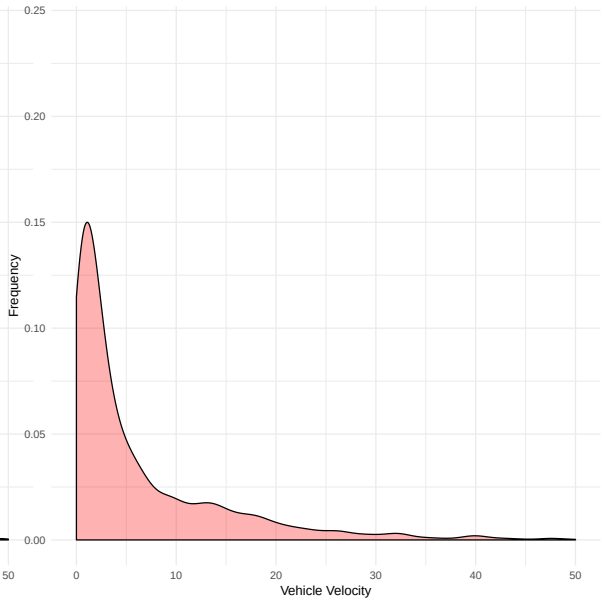

Figure 4.20: Velocity after first arrival

4.6.5 Segment population as predictor

An intuitive predictor of vehicle activity is whether there is a residential population. Table 4.5 shows the summary statistics for vehicle velocity on segments with and without residential populations. The mean and median velocities are much closer in this case. To test for separate distributions the hypothesis test in this case is set out as below, with $\tilde{x}_P$ the median velocity on segments with population and $\tilde{x}_{NP}$ the median velocity for segments without population.

$\mathbf{H_0}$:$\tilde{x}_P - \tilde{x}_{NP} \geq 0$
$\mathbf{H_a}$:$\tilde{x}_P - \tilde{x}_{NP} < 0$

In this case however at a significance level of 0.05 we fail to reject the Null hypothesis since the p-value is 0.7848871. We can therefore not conclude that population is a good predictor of whether a vehicle is servicing or deadheading a segment. While the exact reason for this is not evident there are a number of possible reasons. The first is the small sample size for segment visits without population, as mentioned earlier only 281 traversals were on segments without residential population. Since the vast majority of segments have residential population one can infer that the vehicle must frequently deadhead segments with residential population while travelling to other segments. The fact that a segment has a residential population on it therefore does not shed light on whether a particular traversal is a *service* or *deadheading* traversal. For this reason segment population is not used in further analyses to determine vehicle activity.

Table 4.5: Summary statistics of vehicle velocity for segments with and without population

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Population on Segment >0	0.00066	0.96993	2.26307	4.56350	4.74122	96.39137
Population on Segment = 0	0.00042	0.41078	1.39296	5.40236	4.68665	88.96328

4.6.6 Combined predictor

The final step in extracting service and traversal costs is to combine the variables discussed above to label segment traversals as either *servicing* or *deadheading*. Traversals are separated into *servicing* when they occur on a service day, by a service vehicle, and where the segment is only traversed once. This is based on the fact that the preceding sections all showed a statistically significant velocity difference for each of these variables which indicates that they are appropriate as predictors of vehicle activity. This represents the clearest and most credible sample of velocities associated with servicing. Samples that don't conform to these criteria are therefore labelled as *deadheading*. For the case study area a total 3299 samples are labelled as *servicing* and 2540 are considered *deadheading*. Again we need to statistically verify that these two samples, which are separated using the predictor variables just discussed, are from two different populations. For that reason we again apply with Wilcoxon Rank Sum Test as before, with a significance level of 0.05.

$\mathbf{H_0}$:$\tilde{x}_S - \tilde{x}_D \geq 0$
$\mathbf{H_a}$:$\tilde{x}_S - \tilde{x}_D < 0$

The p-value for the combined predictor based on the variables discussed is $1.043522E-31$. Therefore even at a lower significance level of 0.01 we reject the null hypothesis and can conclude that service velocity is lower than deadheading velocity. Table 4.6 summarises vehicle velocity for servicing and deadheading traversals while Figures 4.21 and 4.22 show the velocity distributions for the two velocity samples.

Table 4.6: Service and deadheading velocities

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Service	0.00127	0.86056	1.86953	3.29221	3.63904	96.39137
Deadheading	0.00042	0.89292	2.78632	6.46072	7.64930	92.58369

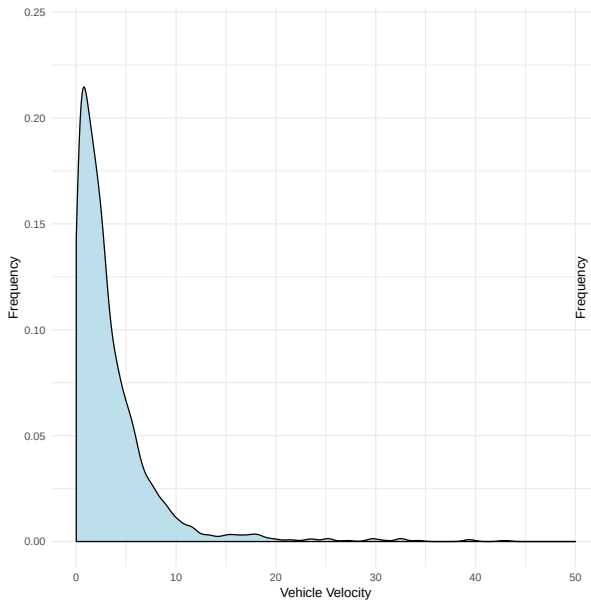

Figure 4.21: Service Velocity

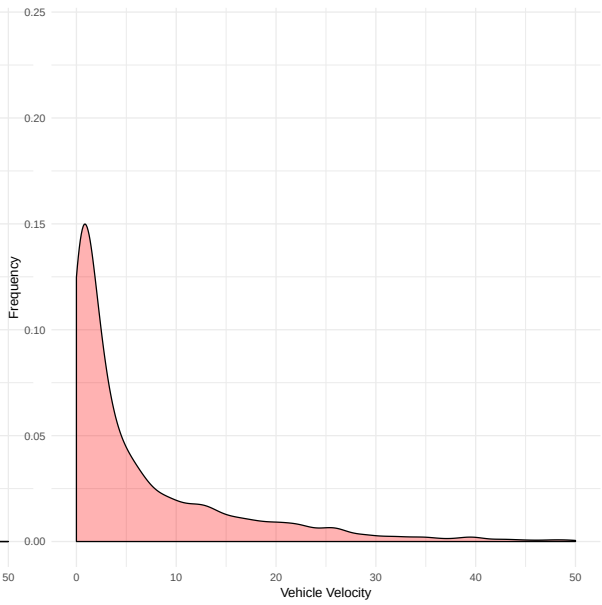

Figure 4.22: Deadhead Velocity

4.7 Results for additional case study areas

It is important at this point to compare results between different service areas. While detailed results for a single beat were presented to illustrate the analytical process, results from multiple service areas provide insights into the effectiveness of the analysis and highlights the stochastic nature of the underlying data.

4.7.1 Service days over multiple service areas

Figure 4.23 shows a selection of seven beats. The boxes represent the IQR, or the area between the first and third quartile. The line in the box is the median and the lines extending from the box is the area 1.5 times the IQR above and below the first and third quartiles, respectively. The dots represent outliers. What the plot aims to illustrate is the velocity distributions for different service areas, on service and non service days.

Table 4.7 summarises the same information as Figure 4.23, but in tabular form. It also allows a comparison between mean velocities per service area for service and non service days. Finally, the results of the Wilcoxon Rank Sum Test on different service areas are also displayed, at a significance level of 0.05. What can be concluded from the table is that:

1. Mean vehicle velocities per service area are in all cases lower on service days than on non service days
2. Median vehicle velocities are not in all cases lower on service days than on non service days
3. Velocities at the 3rd Quartile are in all cases lower on service days than on non service days
4. The Null hypothesis that $\mathbf{H_0}$:$\tilde{x}_{SD} - \tilde{x}_{NSD} \geq 0$ is rejected in 4 of 7 cases, meaning that the velocity distributions for service days compared to non service days are statistically distinct in 4 of 7 cases.

The implications of comparing velocities for multiple areas between service and non service days is that we can evaluate the effectiveness of this metric. In 4 of the 7 cases the velocity samples from service days are statistically different from velocity samples from non service days. This indicates that the variable is not a perfect predictor of vehicle activity, though it does appear to work in some cases.

Table 4.7: Summary statistics and results of hypothesis test using Wilcoxon Rank Sum Test on multiple service areas, when comparing service days

BeatID	Service Day	1st Quartile	Mean	Median	3rd Quartile	P-Value	Hypothesis Test
1	No	0.251	7.452	3.422	10.722	0.0000009	Reject H_0
	Yes	0.805	3.836	1.807	3.565		
10	No	0.436	9.874	3.840	13.583	0.8224140	Fail to reject H_0
	Yes	1.236	8.417	3.227	9.813		
212	No	0.312	6.030	1.854	7.335	0.0000001	Reject H_0
	Yes	0.471	3.485	1.475	3.181		
342	No	0.194	14.435	2.813	14.663	0.8248478	Fail to reject H_0
	Yes	1.705	5.550	3.162	5.286		
368	No	0.219	4.848	1.456	4.734	1.0000000	Fail to reject H_0
	Yes	0.756	5.265	2.095	4.971		
484	No	0.096	15.371	2.076	28.222	0.0138486	Reject H_0
	Yes	0.649	5.178	1.883	3.682		
679	No	0.420	9.014	3.447	12.766	0.0060906	Reject H_0
	Yes	0.842	5.673	2.445	5.917		

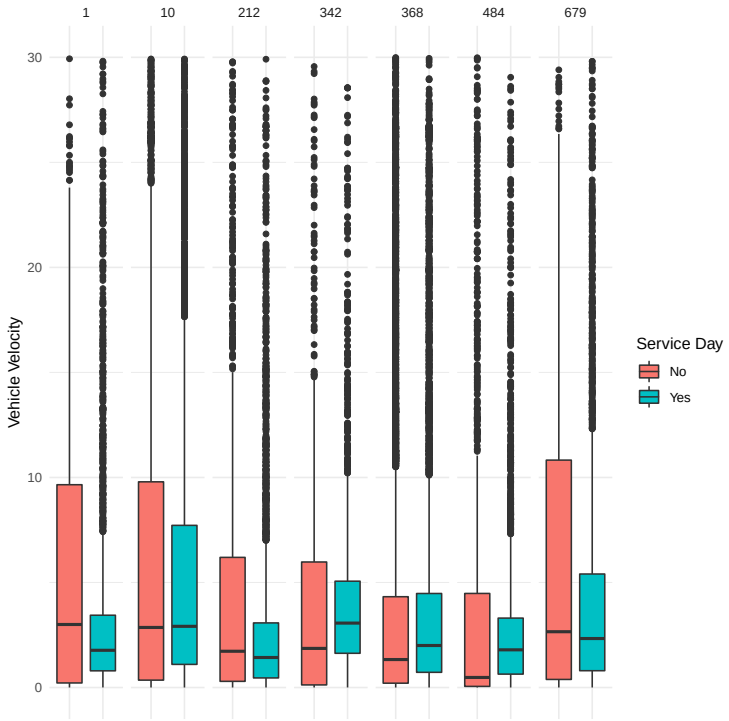

Figure 4.23: Box and whisker plot of vehicle velocity on service and non service days for a selection of seven beats.

4.7.2 Service vehicles over multiple service areas

Next we evaluate the effectiveness of using service vehicles to separate traversals into servicing and deadheading, by considering the parameter over multiple beats or service areas.

Upon considering Figure 4.24 and Table 4.8 the following observations and conclusions can be made:

1. Mean velocities in all cases are lower for service vehicles than non service vehicles.
2. Median velocities are not lower in all cases for service vehicles than non service vehicles.
3. 3rd Quartile velocities are lower for all cases for service vehicles compared to non service vehicles.
4. The Null hypothesis that $\mathbf{H_0}$:$\tilde{x}_{SV} - \tilde{x}_{NSV} \geq 0$ is rejected in 3 of 7 cases.

Service vehicle as a predictor of vehicle activity only produces statistically different velocity results in 3 of 7 cases. This can likely be explained by the fact that allocating different vehicles to a service area is relatively easy. A vehicle might break down or be in for maintenance and be replaced with another vehicle. The vehicle itself is therefore not a perfect indicator of whether a segment is being serviced, though it does give some indication.

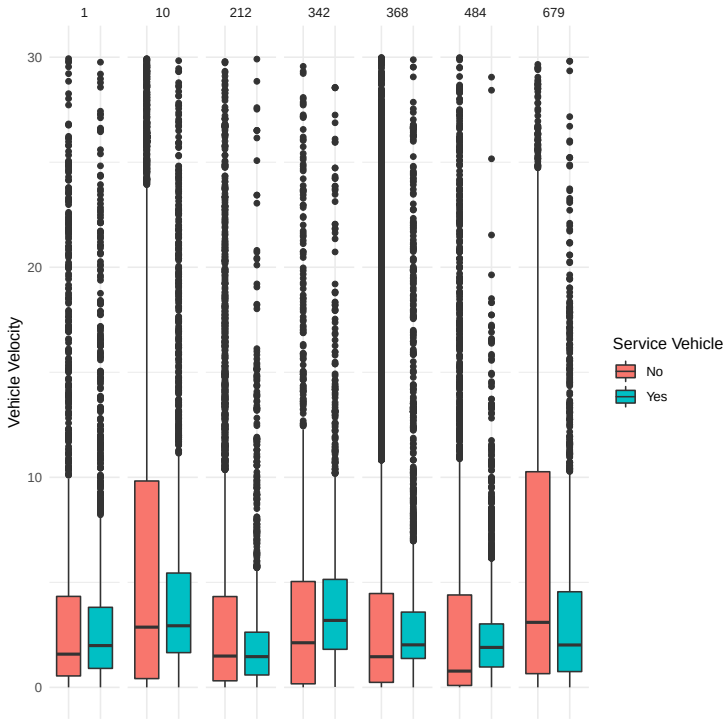

Figure 4.24: Box and whisker plot of vehicle velocity for service and non service vehicles for a selection of seven beats.

Table 4.8: Summary statistics and results of hypothesis test using Wilcoxon Rank Sum Test on multiple service areas, when comparing service vehicles

BeatID	Service Vehicle	1st Quartile	Mean	Median	3rd Quartile	P-Value	Hypothesis Test
1	No	0.564	4.871	1.632	4.732	0.9999656	Fail to reject H_0
	Yes	0.921	4.011	2.027	3.931		
10	No	0.518	10.039	3.833	13.713	0.0038390	Reject H_0
	Yes	1.695	5.797	2.981	5.760		
212	No	0.326	4.736	1.586	4.886	0.0434541	Reject H_0
	Yes	0.599	2.882	1.488	2.671		
342	No	0.271	11.273	2.659	8.972	0.9999947	Fail to reject H_0
	Yes	1.850	5.262	3.246	5.359		
368	No	0.254	5.080	1.644	5.123	1.0000000	Fail to reject H_0
	Yes	1.376	4.300	2.051	3.702		
484	No	0.137	12.628	1.868	14.350	0.2887929	Fail to reject H_0
	Yes	0.983	3.025	1.920	3.066		
679	No	0.740	8.922	3.662	12.444	0.0000000	Reject H_0
	Yes	0.774	4.437	2.093	4.802		

Table 4.9: Summary statistics and results of hypothesis test using Wilcoxon Rank Sum Test on multiple service areas, when comparing a segments first traversal to subsequent traversals

BeatID	First Traversal	1st Quartile	Mean	Median	3rd Quartile	P-Value	Hypothesis Test
1	No	1.327	5.856	2.859	6.266	1.11e-27	Reject H_0
	Yes	0.661	4.017	1.701	3.616		
10	No	1.123	11.066	4.881	16.997	2.50e-20	Reject H_0
	Yes	0.698	8.667	3.182	10.397		
212	No	0.900	5.686	2.564	7.116	2.57e-42	Reject H_0
	Yes	0.360	3.804	1.327	3.217		
342	No	2.073	6.652	3.969	7.589	8.95e-12	Reject H_0
	Yes	1.278	7.289	3.005	5.423		
368	No	0.333	5.803	2.118	6.292	1.69e-18	Reject H_0
	Yes	0.310	4.677	1.637	4.347		
484	No	0.794	10.385	2.748	8.381	1.05e-13	Reject H_0
	Yes	0.327	8.636	1.776	5.352		
679	No	0.888	7.366	2.672	9.132	1.12e-07	Reject H_0
	Yes	0.691	5.649	2.414	5.899		

4.7.3 First traversal over multiple service areas

The first traversal proved to be a good indicator of vehicle activity when considering the case study area earlier in the chapter. For this reason the first traversal is evaluated as predictor of vehicle activity by again considering multiple areas. Upon considering Table 4.9 and Figure 4.25 the following observations and conclusions can be made:

1. Mean velocities in all cases are lower when a vehicle traverses a segment for the first time, compared to subsequent traversals, except in the case of one service area.
2. Median velocities are lower in all cases for a vehicle traversing a segment for the first time, compared to subsequent traversals.
3. 3rd Quartile velocities are lower for all cases for a vehicle traversing a segment for the first time, compared to subsequent traversals.
4. The null hypothesis that $\mathbf{H_0}$:$\tilde{x}_{1st} - \tilde{x}_{>1st} \geq 0$ is rejected in 7 of 7 cases.

The traversal sequence variable performed best of all variables tested, and found statistically different velocities in all seven service areas. Indicating that traversal sequence is the best estimator of vehicle activity.

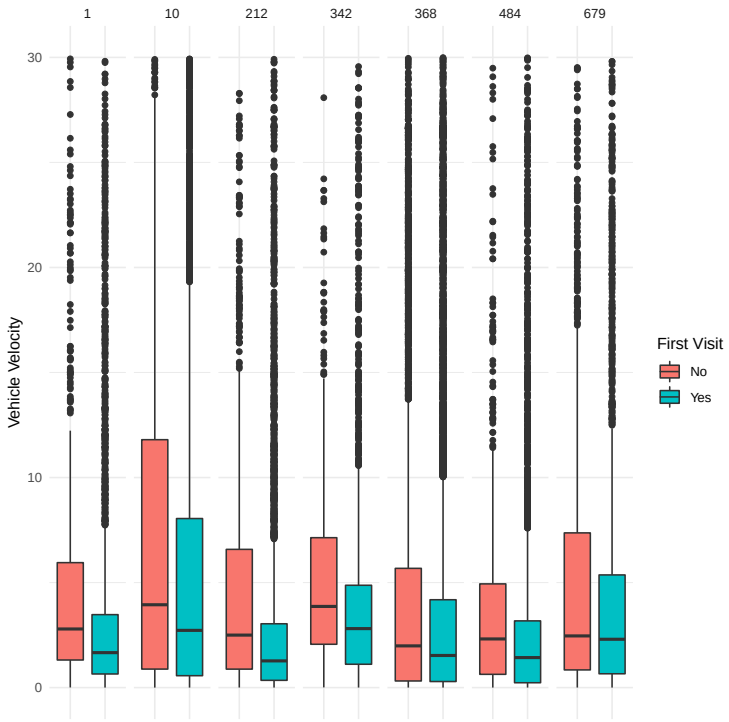

Figure 4.25: Box and whisker plot of vehicle velocity for the first segment traversal compared to subsequent traversals

4.7.4 Combined activity predictor over multiple service areas

As before we can now group observations that occur on a service day, with a service vehicle, and where the observation is the first time a segment is traversed as *servicing* and group all other observations as *deadheading*. Figure 4.26 shows the velocity box plots of the seven areas, following the classification of segment visits into servicing and deadheading. Table 4.10 summarises the velocity behaviour for the service and deadheading groups. Overall, over all seven beats, the mean service velocity is estimated at 3.857 km/h, while the deadheading velocity is estimated at 6.843 km/h. Both these estimates are significantly lower than the estimates discussed in literature. Furthermore the results of the activity classification show that:

1. Mean velocities in all cases are lower for vehicle servicing segments than for vehicle deadheading segments.
2. Median velocities are lower in all but two cases for vehicles servicing segments than for vehicles deadheading segments.
3. 3rd Quartile velocities are lower for all cases where vehicles are servicing segments compared to when they are deadheading segments.
4. The null hypothesis that $\mathbf{H_0}$:$\tilde{x}_{1st} - \tilde{x}_{>1st} \geq 0$ is rejected in 6 of 7 cases.

This indicates that the combined classification works well and that using the three variables in conjunction allows us to produce separate velocity estimates for when a vehicle is servicing and when a vehicle is deadheading a segment.

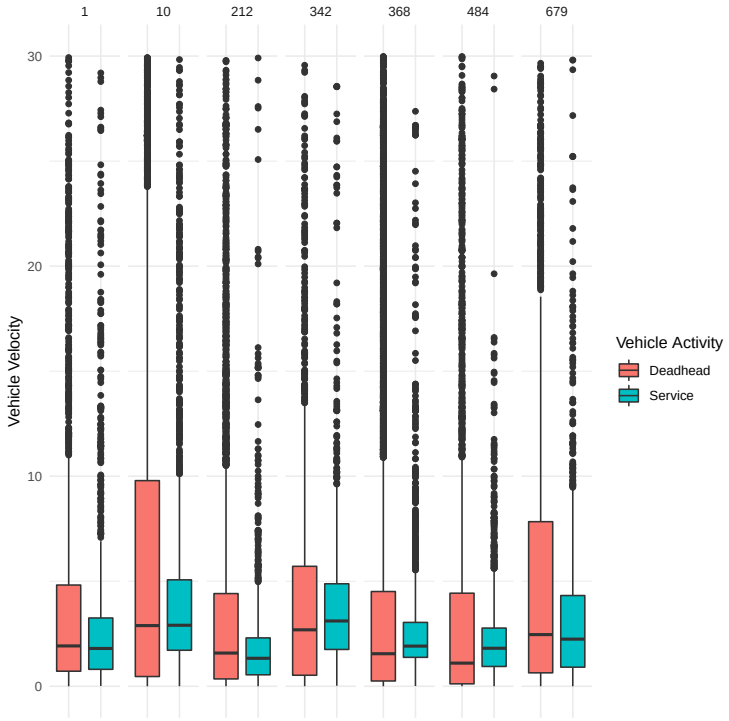

Figure 4.26: Box and whisker plot of vehicle velocity for service and deadheading traversals

Table 4.10: Vehicle service and deadheading velocities for different beats

BeatID	Service Day	1st Quartile	Mean	Median	3rd Quartile	P-Value	Hypothesis Test
1	Deadhead	0.7279942	5.068204	1.967690	5.203469	1.168486e-02	Reject
	Service	0.8112920	3.420071	1.826846	3.319047		
10	Deadhead	0.5735867	9.862515	3.745269	13.279220	7.443663e-04	Reject
	Service	1.7379910	5.428756	2.941914	5.262168		
212	Deadhead	0.3691018	4.758224	1.668291	4.963015	5.417244e-09	Reject
	Service	0.5470558	2.421892	1.352779	2.363217		
342	Deadhead	0.6321404	9.446291	3.032630	7.819317	8.392583e-01	Reject
	Service	1.7783977	5.201919	3.191482	5.083286		
368	Deadhead	0.2680600	5.139255	1.706855	5.214722	1.000000e+00	Fail to Reject
	Service	1.3775235	3.281168	1.914394	3.103874		
484	Deadhead	0.1786977	11.689058	2.029320	11.516995	5.086352e-03	Reject
	Service	0.9509374	2.855576	1.826745	2.794795		
679	Deadhead	0.6724058	7.467968	2.679237	9.387952	1.236209e-08	Reject
	Service	0.9259160	4.133057	2.306338	4.516156		

4.8 Conclusion

Service and deadheading costs are crucial parts of the MCARPTIF models. Vehicles must service all arcs and traverse arcs to access arcs to be serviced. To fully utilise the MCARPTIF models it is crucial that these costs are estimated accurately, using real data. By using GPS data and open source street network data, accurate service and deadheading velocities, and by implication, costs can be estimated using the techniques presented in this chapter. Unique service and deadheading velocities were estimated for seven service areas within the metropolitan region. In all seven cases both mean and median velocity estimates were significantly lower than those used in literature and will no doubt have a significant impact on the routing solutions presented later in this document, where routing solutions that use common velocity estimates in literature are compared to those estimated in this dissertation.

Chapter 5

Waste generation estimation

One of the crucial input data challenges with the Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities (MCARPTIF) problem variant for waste collection is the estimation of waste generation rates. Waste generation rates represent the demand for the collection service, and are likely to have a significant impact on feasible collection routes.

Municipal solid waste differs from other municipal services in that exact waste generation rates are difficult to measure at household level. Water and electricity, for example, are measured with relative ease, and consumers are billed directly for consumption. Waste generation rates are however more difficult to measure directly.

Typically, municipal waste collection service providers will measure vehicles entering landfills or transfer stations, using weigh bridges. By collecting this data service providers have a broad understanding of generation rates, at a suburb or beat level. The basic inference being that vehicles assigned to certain beats can be weighed upon return to landfills and generation rates for those beats measured over time. While this data is beneficial to strategic planning in terms of waste management infrastructure such as landfills, it is not detailed enough for the optimisation of collection routes.

For the purposes of the MCARPTIF problem variant, detailed generation rates per street segment are required. One potential solution to this problem is to use weigh bridge data at an aggregate level for service areas or for an entire metropolitan area and to disaggregate it to street segments. Since waste generation rates are primarily a function of population density, i.e. the more people in a particular area the higher the waste generation rate is likely to be, population data provides a good opportunity to estimate generation rates at street segment level.

In South Africa, census data is publicly available and of a relatively high quality. This provides the opportunity to use population data, combined with aggregate waste generation rates to estimate waste generation rates per street segment and solve more accurate and useful MCARPTIF instances. To achieve this a synthetic population developed by Joubert (2014) was used to infer waste generation rates. The development of the synthetic population was according to methods presented by Müller and Axhausen (2012) and is described in more detail below.

5.1 Synthetic population development

As briefly described in Chapter 2, census data typically contains detailed samples of a portion of the population. In South Africa this would be the Public Use Micro Sample (PUMS). The sample contains detailed information on individuals, but only contains

coarse geographic data. The Community Profile Data on the other hand contains information on population demographics down to the sub-place level, but no information on actual individuals within the population. The reason for this separation is to protect the privacy of the individuals within the population. To synthesize a population individuals from the PUMS are then allocated to a sub-place, such that sub-place demographics are preserved as accurately as possible.

The synethetic population developed by Joubert (2014) for the metropolitan area in question is based on a multi-level fitting algorithm by Müller and Axhausen (2012). We refer the reader to Müller and Axhausen (2012) for an exhaustive explanation but will briefly discuss the basics here. The synthetic reconstruction process consists of fitting and generation. In the fitting stage the PUMS (the disaggregate sample of members) is weighted, with the reweighted sample corresponding to the sub-place (from the Community Profile Data, or in this case the service area. The reweighted sample is then used to construct the set of population members for the service area, in the second phase of the synthetic reconstruction process. Each member is part of a household, and each household has a geographic location which can be plotted.

Figure 5.1 shows the result of the process and shows the household locations and sizes for a case study area. Figure 5.2 shows the same households in the case study area, separated by household size. As before the process is repeated for a number of other service areas, these can be found in Figures 5.3 to 5.5. The figures show varying household sizes within the service area as well as where the synthetic households can be found. Following the earlier assumption that population size can be used to determine weekly generation rates, the figures give an indication of population density which can be used to estimate waste generation rates per street segment or for the service area as a whole.

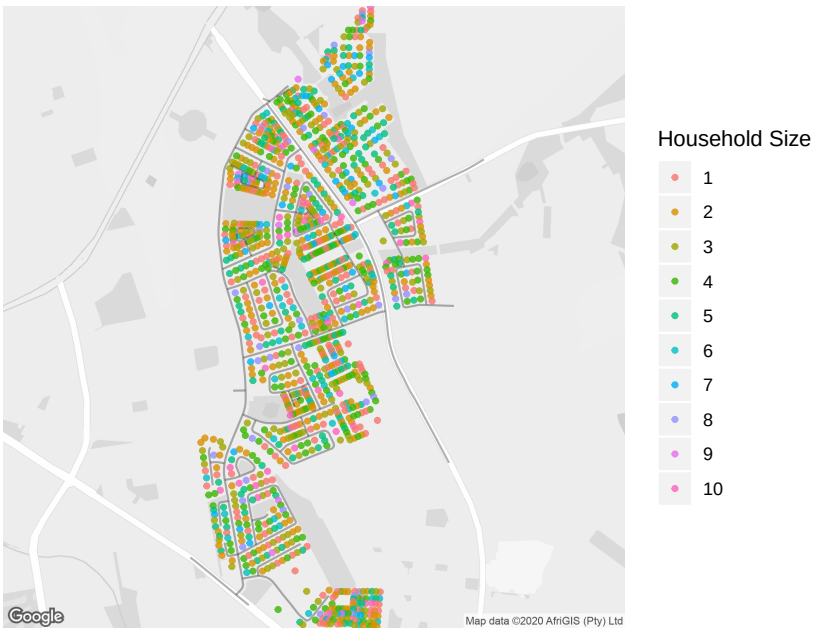

Figure 5.1: Synthetic population household locations for case study area

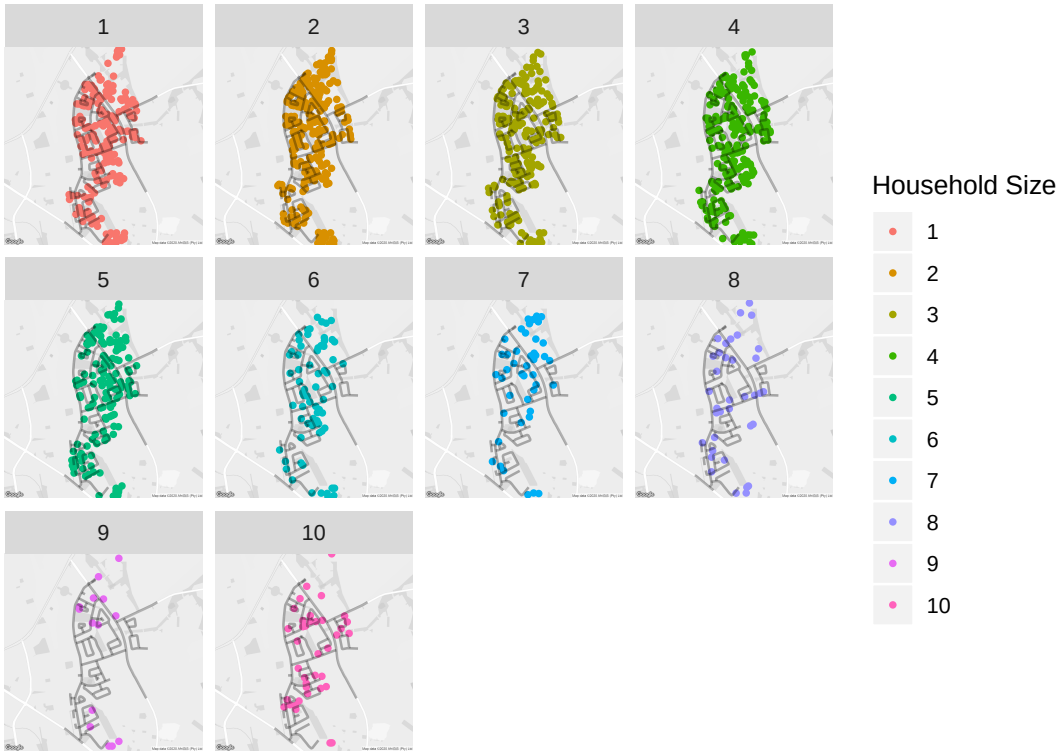

Figure 5.2: Synthetic population household locations by household size for case study area

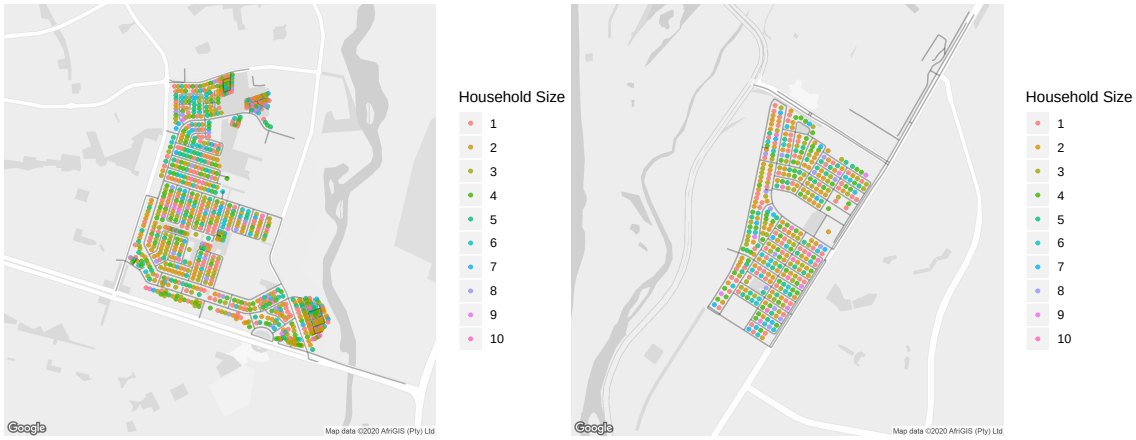

Figure 5.3: Synthetic Populations for beats 1 and 10.

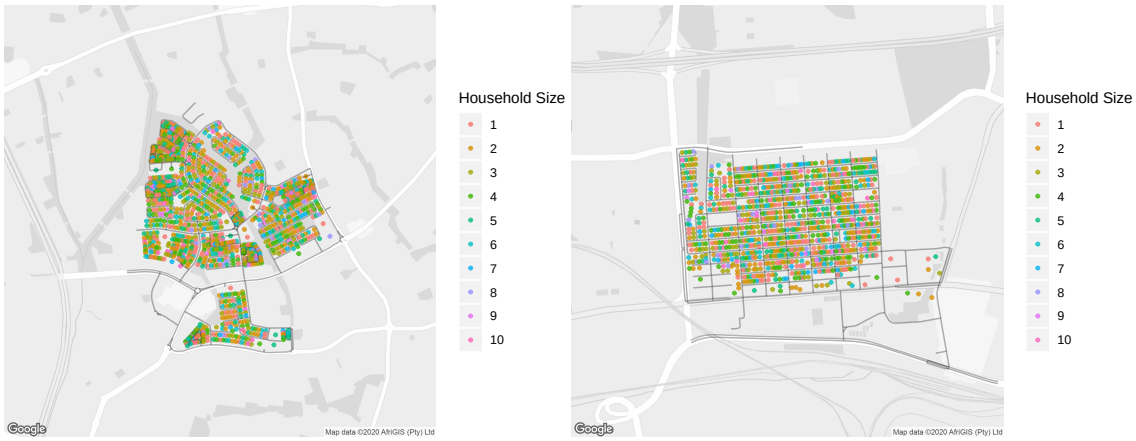

Figure 5.4: Synthetic Populations for beats 212 and 368.

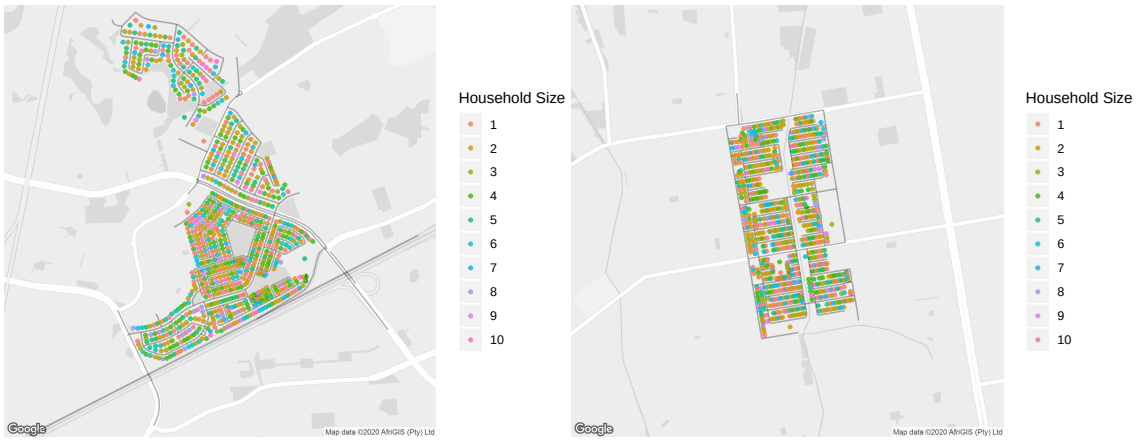

Figure 5.5: Synthetic Populations for beats 342 and 484.

5.2 Matching households to street segments

Following the development of the synthetic population the next step is the assign the population members to street segments, because the MCARPTIF algorithm requires a demand, or weekly waste generation rate per street segment. Members of the synthetic population therefore must be associated with a street segment, such that the total population and subsequently the total waste generation rate per segment can be calculated.

Each household has coordinates to show the approximate location of the household, as well as the household size. To assign the households to street segments the same Point snapping methodology described in section 4.3 was used. The distance from each household is compared to the distance to each of the street segments in the network and the lowest distance segment is selected. The household location is then moved onto the line segment and assigned to the line segment. The result is visualised in Figure 5.6. The figure shows the households snapped to the nearest street segment. The purpose of this is simply to visually illustrate that each household is now associated with a specific street segment within the service area, allowing the calculation of the total population and demand per street segment.

Figure 5.6: Synthetic population households snapped to street network for case study area

Since members of the synthetic population are now linked with a street segment, the population per street segment for the case study are can now be plotted. Figure 5.7 shows a bar graph of segment population estimates for the case study area based on the synthetic population. The figure shows the segment ID with its population estimate which ranges from 155 people on the most populous segment, all the way down to 3 people on the least populous segment.

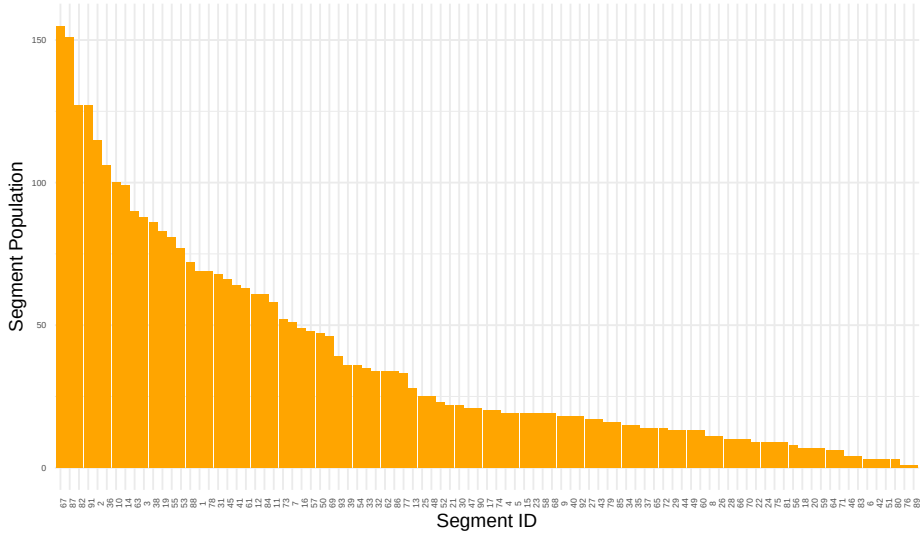

Figure 5.7: Synthetic population for street segments in case study area

5.3 Estimating segment waste demand

Following the allocation of synthetic households to street segments it follows that the total population for each segment can then be calculated. Since the problem deals with residential waste collection, the assumption is made that people are the primary source of residential municipal solid waste and that waste generation rates are primarily a function of population. This, along with the fact that living standards are a big contributor to waste generation rates, is shown by Wertz (1976) and Medina (1997) amongst others.

5.4 Waste generation rates

In section 2.6 a number of reported waste generation rates were discussed. For the purpose of this study an aggregate rate of 580 kg per person per annum was selected, as reported in an Integrated Waste Plan by the metropolitan area (Solid Waste Management, 2016). Even though the number is reported as a per capita estimate, it is based on the annual measured waste generation rate of the entire metropolitan area divided by the total population estimate. It therefore represents a good estimate of the aggregate waste generation rate. Another approach, if the data can be collected, would be to use waste generation data from weighbridges at landfill sites, as opposed to the entire city. This would help capture differences in waste generation rates within the city itself. However since this data was not available the aggregate generation rate for the city is sufficient. Of the 580 kg per person per annum an estimated 45 percent is residential municipal solid waste, the balance is made up of building rubble and commercial waste. The per capita waste generation rate per collection week is therefore calculated as $(580 \times 0.45)/52 = 5.01$ kg per person per week or 0.715 kg per person per day. This compares well with the generation rates reported in literature, 1.76 kg per person per day by Qdais et al. (1997), 1.11 kg per person per day by Minghua et al. (2009), 0.77 kg per person per day by Troschinetz and Mihelcic (2009) and 1.62 kg per person per day by Saeed et al. (2009). The small discrepancies are likely due to the inclusion of building rubble and commercial waste. If we use the entire 580 kg per person per annum estimate the average per capita

rate would be 1.59 kg per person per week. Using this estimate the weekly generation rate, or the MCARPTIF demand, per segment can be calculated. Figure 5.8 shows the generation rate per street segment. Since the generation rate is a product of the population it follows the same shape as Figure 5.7. The maximim generation rate per segment is 777 kg per week. With the minimum being 15 kg per week. The analysis is repeated for seven other service areas, Table 5.1 shows the population and waste generation estimates for the seven segments

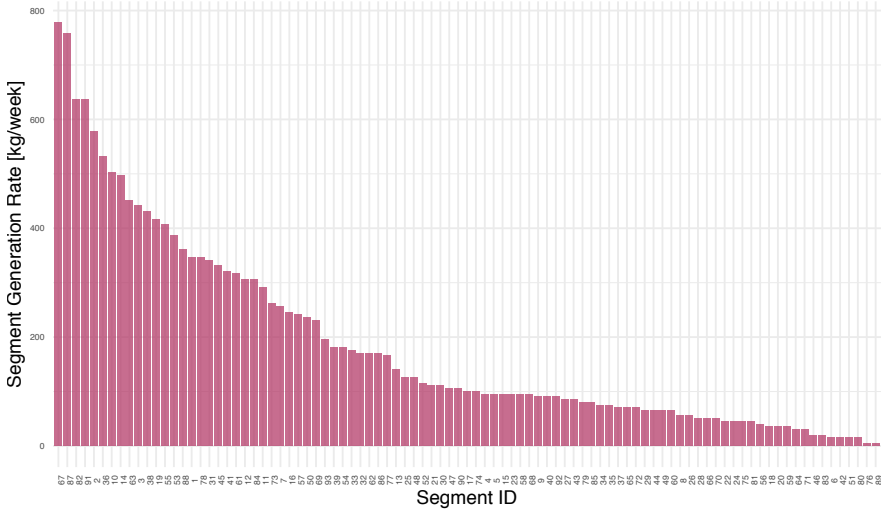

Figure 5.8: Estimated generation rate for street segments in case study area

Table 5.1: Estimated beat population and weekly waste generation rate.

Beat ID	Population Estimate	Estimated Weekly Generation Rate [kg]
1	3012	15117.92
10	1405	7052.01
212	4304	21602.76
342	2889	14500.55
368	3438	17256.11
484	2536	12728.76
679	4612	23148.69

5.5 Conclusion

Municipal waste management differs from other municipal utilities in that waste generation rates are difficult to estimate at the household or street level. This is because there are no cost effective ways of measuring waste generation at the source (household), in contrast to how you can measure water consumption for a household using a flow meter. However, to solve the MCARPTIF and improve collection routes, the demand per street segment is crucial. In this chapter a synthetic population and known aggregate waste generation rates for a metropolitan area was used to estimate waste generation rates per street segment. This then represents the demand required to solve MCARPTIF instances. Future work on

waste generation rates could look at using population income data to improve generation estimates.

Chapter 6

Comparing the effects of input parameter estimates on waste collection routes

The culmination of all the analysis described up to this point is a number of reproducible case study test instances which will be used to demonstrate the effect on routing solutions when actual data estimates are used. For the purposes of this dissertation, seven service areas were randomly selected to demonstrate the model functionality, although any service area from the metropolitan area can be used. Each beat's individual characteristics and calculated Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities (MCARPTIF) parameters will be briefly discussed. Following that the process to produce test files for the MCARPTIF algorithm will be described. Finally, the results from solving the test instances will be discussed. The goal of the chapter is to demonstrate the magnitude of the differences in routing outcomes when taking differing approaches to MCARPTIF input parameters.

6.1 MCARPTIF test files

To solve the test instances, a file is produced for each instance, which serves as input into the solution algorithm. The file contains all the test parameters relevant to that specific beat. Figure 6.1 shows an extract from a test file. The file contains the number of nodes in the street network, the number of edges and arcs, the vehicle capacity, the shift length, the maximum number of vehicles available, the offloading cost as well the street segment data such as traversal and servicing cost.

For each beat two separate test files are exported, the first based on the data estimation techniques discussed thus far (this well be referred to as the *refined approach*) and the second based on assumptions made in literature for the MCARPTIF parameters (referred to as the *standard approach*).

6.1.1 Network construction

The first step in producing the test instances is to extract the network nodes from the *OpenStreetMap* (OSM) street network. This is done by isolating the nodes at the ends of each street segment. Since the network is already in graph form from the analysis presented in Chapter 4 the nodes can simply be extracted from the network. This can be seen in Figure 6.2. Once a list of nodes is compiled a short algorithm, Algorithm 3,

```
NAME : 1_Actual
NODES : 120
REQ_EDGES : 141
NOREQ_EDGES : 0
REQ_ARCS : 0
NOREQ_ARCS : 0
VEHICLES : 10
CAPACITY : 10000
DUMPING_COST : 960
MAX_TRIP : 28800
LIST_REQ_EDGES :
(11,98)   serv_cost 123   trav_cost 83   demand 55
(13,98)   serv_cost 80   trav_cost 54   demand 35
(18,101)   serv_cost 170   trav_cost 114   demand 115
(19,101)   serv_cost 61   trav_cost 41   demand 40
(13,18)   serv_cost 81   trav_cost 55   demand 75
(11,34)   serv_cost 335   trav_cost 226   demand 236
(44,19)   serv_cost 81   trav_cost 55   demand 40
(54,44)   serv_cost 369   trav_cost 249   demand 346
(58,54)   serv_cost 116   trav_cost 78   demand 60
(33,30)   serv_cost 127   trav_cost 86   demand 50
(34,30)   serv_cost 102   trav_cost 68   demand 20
(33,14)   serv_cost 152   trav_cost 103   demand 146
(36,110)   serv_cost 385   trav_cost 259   demand 45
```

Figure 6.1: Extract from a MCARPTIF test file.

ensures that the network is complete and that all segments are attached to the network. The algorithm ensures that each segment is connected to at least one other segment by sharing a common node with another segment. This prevents errors from occurring when a feasible solution is sought using the solution algorithms.

Figure 6.2: Network nodes for beat 679.

The next step involves adding arcs and edges that connect the test instance to the landfill sites and vehicle depot discussed in Chapter 3. Connecting the service area to landfills and to the vehicle depot involves identifying nodes where vehicles can potentially enter or exit the service area. The travel cost between these nodes and the landfill sites and vehicle depot is then calculated using the Google Maps Distance Application Programming Interface (API) through the *ggmap* package in **R** by Kahle and Wickham (2013). This gives the best route between each entry or exit node and the sites in question. For each beat the nearest landfill site is selected as the designated landfill site and the selected site

becomes a node within the network. The same applies for the vehicle depot. The traversal cost between the landfill and depot sites and the beat's entry and exit nodes are calculated using the Google Maps API again.

At this point the network construction is complete and the test parameters can be added. It is here where the two files, for the *standard approach* and *refined approach* diverge in terms of how they are constructed. The first significant difference is in determining required and non required edges. Required edges refer to segments that must be serviced by the vehicle, while non required edges don't need to be serviced but can be traversed as part of the vehicle route. For the standard approach, waste generation rates are based on an estimate of waste produced per meter of street segment. For this reason all segments within the network, with the exception of the segments linking the landfills and depots, are assumed to be required edges.

For the refined approach, the synthetic population allows one to determine which segments are required and which aren't, since only segments with population present require servicing. Table 6.1 provides a summary of some of the variables pertaining to the test instances.

Input: network segments
Output: list of segments not connected to network
for $i \in segments$ **do**
  1stTo1stNodeMatch = segments[segments that share first node with segment i start node]

  1stTo2ndNodeMatch = segments[segments that share second node with segment i start node]

  2ndTo1stNodeMatch = segments[segments that share first node with segment i second node]

  2ndTo2ndNodeMatch= segments[segments that share second node with segment i second node]

  if $length(1stTo1stNodeMatch) == 1 AND$
   $length(1stTo2ndNodeMatch) == 0\ AND$
   $length(2ndTo1stNodeMatch) == 0\ AND$
   $length(2ndTo2ndNodeMatch) == 1$ **then**
    print("Segment i is not connected to the rest of the network"
  end
end

Algorithm 3: Network Integrity Algorithm

6.1.2 Waste demand

Since required and non required edges are determined, the segment demand can be determined. For the standard approach the segment demand is a function of the segment length and the estimated generation rate per metre, $0.5kg/m/week$. This, and the following estimates, are those presented in benchmark instances by Willemse and Joubert (2016a). The reason estimates by Willemse and Joubert (2016a) were selected for comparison is that the authors provide good detail on their assumptions, which allows the same

Table 6.1: Beat network variables

Beat ID	Edges	Required Edges	Non-Required Edges	Nearest Landfill [m]
1	141	121	20	1341
10	164	74	90	4229
212	261	134	127	3094
342	148	109	39	6926
368	246	170	72	3412
484	106	91	15	1678
679	103	95	8	6071

assumptions to be applied to the areas discussed in this dissertation and compared to the estimates developed here. For the refined approach the demand is the population on the segment multiplied by the estimated weekly per capita generation rate of $5.01 kg/person/week$.

6.1.3 Service and traversal cost

The next step in constructing the test instances is to determine the service and traversal times of each segment. For the standard approach the assumed traversal velocity is 20 km/h. The traversal time in seconds is then simply the product of the segment length and the traversal velocity. For the service time the assumption is made that an additional 10 s is incurred for each kilogram of waste. The service time in seconds, S_i, for segment i can therefore be expressed as:

$$S_i = (L_i \times 20) + (D_i \times 10) \tag{6.1}$$

where L_i is the length of segment i in metres and D_i is the calculated demand for segment i in kilograms.

For the refined approach the service and deadheading times are calculated using the servicing and deadheading velocities estimated for each beat in Chapter 4, divided by the length of the segment. Of course it would also be possible to use a unique velocity estimate for each segment, as opposed to for each beat, but this would require a larger sample size of Global Positioning System (GPS) points, since the number of velocity observations per street segment are low. In addition, not all segments in the beat necessarily have observed traversals, meaning that not all segments have velocity estimates. For this reason using an aggregate velocity estimate per beat is the best option to both capture the velocity behaviour unique to each beat (i.e. velocity differences per beat), as well as not having to exclude any segments due to lack of usable observations.

The impact of this on the results is that the velocity behaviour is less granular than initially planned (i.e velocities are not unique to each segment but rather unique to each beat) though it remains an improvement on velocity estimates discussed in literature. Future work, with larger sample sizes, could easily achieve the desired level of detail by using the methods described in Chapter 4.

6.1.4 General parameters

Finally, both test instances for the standard and refined approach contain general parameters. The first of these are the vehicle capacity, which for both is assumed to be $10000 kg$. The dumping cost for the standard approach is assumed to be 300 s, as reported by Willemse

and Joubert (2016a), while for the refined approach it is 960 s, which is the estimate produced in Chapter 3.

6.2 Results comparison

Once the test instances have been generated the final step is to solve the instances using the MCARPTIF algorithms. The MCARPTIF algorithms are described in detail in Willemse and Joubert (2016b). The algorithms developed by Willemse and Joubert (2016b) are constructive heuristics aimed at finding quick initial solutions. Practically this is very useful for MCARPTIF problems employed for routine decision support as new solutions are required at relatively short notice. The heuristics are based on Path-Scanning where the algorithm iteratively constructs the route by adding the nearest unserviced street segment to the end of the route, while ensuring that capacity and time constraints are not violated.

The algorithms are Python based and import the constructed test instances as text files. Once the algorithm has found a feasible solution a text file of the solution is produced. An extract of a typical solution is shown in Table 6.2. The solution includes a route ID, this represents a separate route for each vehicle. The subroute is each collection cycle, or the route between each transfer station visit. Each activity is assigned an ID and has a start and end node within the network. Activity types include depot start, collection, traversal, arrival at transfer stations, offload and depot end.

Furthermore the solution includes a column on the traversal time to the activity, as well as the time taken to complete each activity, the activity demand, the remaining capacity for waste in the vehicle and the remaining time before the end of the shift. Given these pieces of information on the feasible solutions for each beat, we can now compare the solutions from the two input parameter approaches and determine the effect of the input data on the solutions.

6.2.1 Generation rates

We begin by comparing generation rates between the estimates produced in Chapter 5, and the assumptions used by Willemse and Joubert (2016a). All beats, with the exception of beat 10, have higher actual generation rates based on the synthetic population than on the assumptions used in literature. Figure 6.3 shows the comparison of resulting waste generation rates per beat using actual data estimation methods compared to those used in literature.

The reason for beat 10's lower generation rate is because beat 10 has a low number of required edges, when the synthetic population is used to determine which edges are required. Table 6.1 shows that beat 10 has 74 required edges, of the 164 edges in total. This means that only 45 percent of edges require servicing in the network. Since the generation rate estimate method in literature does not have a mechanism for determining which edges are to be serviced, 100 percent of the edges in beat 10 are required, resulting in a larger generation rate estimate for the whole beat. This is illustrated in Figure 6.4 where the street network and synthetic population is imposed on a satellite view of the beat. Visible on the bottom left of the image is the synthetic population in blue, in what appears to be a residential area. On the top and to the right street segments are shown where there is no synthetic population present, and where the buildings appear to be commercial properties. The incorporation of the synthetic population in the refined approach therefore excludes these edges, where they are included as demand points in the

Table 6.2: Extract from routing solution

Route	Subroute	Activity ID	Arc Start Node	Arc End Node	Activity Type	Traversal Time To Activity	Activity Time	Activity Demand	Remaining Capacity	Remaining Time	Cumulative Demand	Cumulative Time
0	0	0	1	1	depot_start	0	0	0	10000	28800	0	0
0	0	21	1	59	traversal	0	1392	0	10000	27408	0	1392
0	0	193	59	7	collect	1392	157	100	9900	27251	100	1549
0	0	44	7	78	collect	0	327	291	9609	26924	391	1876
0	0	205	78	21	collect	0	318	1712	7897	26606	2103	2194
0	0	186	21	3	collect	0	196	828	7069	26410	2931	2390
0	0	184	3	11	collect	0	119	45	7024	26291	2976	2509
0	0	36	11	72	collect	0	42	256	6768	26249	3232	2551
0	0	37	72	11	traversal	0	26	0	6768	26223	3232	2577
0	0	182	11	24	collect	26	73	15	6753	26150	3247	2650
0	0	102	24	52	collect	0	105	90	6663	26045	3337	2755
0	0	150	52	51	collect	0	61	30	6633	25984	3367	2816
0	0	146	51	16	collect	0	44	171	6462	25940	3538	2860
0	0	180	16	74	collect	0	46	80	6382	25894	3618	2906
0	0	181	74	16	traversal	0	29	0	6382	25865	3618	2935
0	0	149	16	51	collect	29	179	452	5930	25686	4070	3114
0	0	151	51	52	traversal	0	38	0	5930	25648	4070	3152
0	0	101	52	26	collect	38	87	181	5749	25561	4251	3239
0	0	88	26	21	collect	0	103	176	5573	25458	4427	3342
0	0	89	21	26	traversal	0	63	0	5573	25395	4427	3405
0	0	82	26	81	collect	63	136	110	5463	25259	4537	3541
0	0	85	81	27	collect	0	207	341	5122	25052	4878	3748
0	0	91	27	65	collect	0	65	75	5047	24987	4953	3813
0	0	42	65	65	collect	0	185	502	4545	24802	5455	3998
0	0	90	65	27	traversal	0	40	0	4545	24762	5455	4038
0	0	86	27	25	collect	40	111	171	4374	24651	5626	4149
0	0	64	25	80	collect	0	229	110	4264	24422	5736	4378
0	0	63	80	24	collect	0	204	35	4229	24218	5771	4582
0	0	183	24	11	traversal	0	45	0	4229	24173	5771	4627
0	0	185	11	3	traversal	0	74	0	4229	24099	5771	4701
0	0	8	3	68	traversal	0	9	0	4229	24090	5771	4710
0	0	189	68	2	collect	128	26	20	4209	24064	5791	4736
0	0	172	2	67	collect	0	57	954	3255	24007	6745	4793
0	0	173	67	2	traversal	0	35	0	3255	23972	6745	4828
0	0	191	2	5	collect	35	53	863	2392	23919	7608	4881
0	0	190	5	2	traversal	0	33	0	2392	23886	7608	4914
0	0	188	2	68	traversal	0	16	0	2392	23870	7608	4930

standard approach.

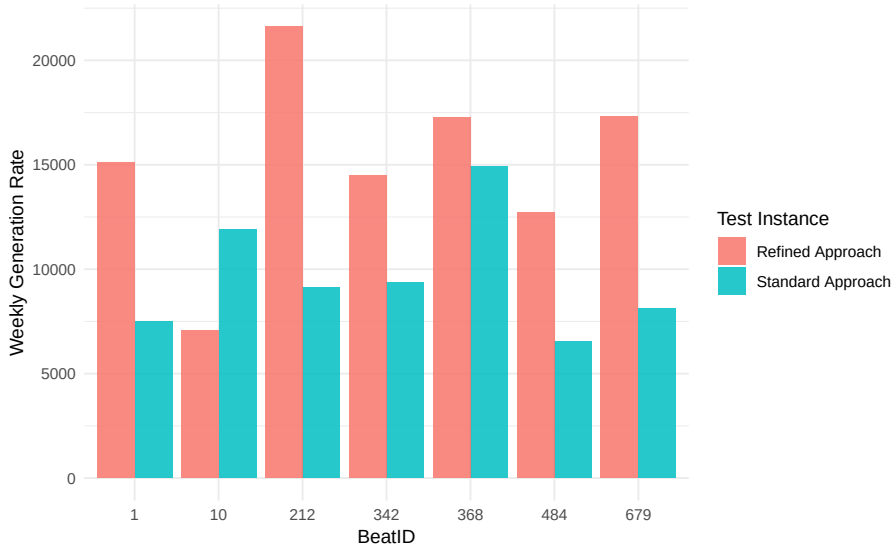

Figure 6.3: Estimated weekly generation rates for standard approach versus refined approach

Figure 6.4: Synthetic population and street network for beat 10

Table 6.3 shows the resulting generation rates for the two approaches to input data. The Table shows that generation rates differ considerably. This is because of the difference in the underlying assumptions with regards to how generation rate is estimated for each approach. The standard approach assumes that segment length is directly proportional to population and in turn to generation rate, since an estimate of $5kg/m/week$ is used. In contrast, using the more refined approach of a synthetic population, no underlying assumption as to population density is made resulting in unique generation rates per street segment.

This is significant as the generation rate will directly affect how accurate the route is. With an incorrect route based on poor generation rate estimates routes could exceed the shift length, or be too short and result in larger vehicle fleets being used to service the metropolitan area. Even without the optimised vehicle route, the generation rate estimate on its own is a tangible contribution, as it can be used by waste managers to better split suburbs into service areas and to plan waste collection activities.

Table 6.3: Generation rates for test instances using the refined approach and standard approach

Beat	Refined approach [kg/week]	Standard approach[kg/week]	% Change
1	15118	7515	101%
10	7052	11911	-41%
212	21603	9116	137%
342	14501	9357	55%
368	17256	14899	16%
484	12729	6544	95%
679	23149	8121	185%

6.2.2 Activity time

The next point of comparison is vehicle activity times. Since the input data parameters differ between the test instances, comparing the resulting activity times is of interest. Table 6.4 shows the vehicle activity times when comparing the standard approach and the refined approach. Figure 6.5 shows vehicle activity for the seven test beats in hours. The figure shows activity times for all routes, for this reason any beats with a total activity time over 8 hours requires an additional vehicle. As expected the vehicles spend the majority of their time collecting waste, followed by traversing or deadheading segments and lastly spend the least amount of time on offloading waste at intermediate facilities. In two of the seven cases more than one vehicle is required to service the area.

Table 6.4: Activity times per beat

	Refined Approach				Standard Approach			
Beat	Collect	Offload	Traversal	Total	Collect	Offload	Traversal	Total
1	4.00	0.80	2.79	7.59	4.93	0.17	1.34	6.44
10	1.92	0.27	1.19	3.37	7.81	0.25	1.64	9.70
212	5.78	1.07	4.52	11.37	5.98	0.17	1.52	7.67
342	2.38	0.80	2.59	5.78	6.13	0.17	2.25	8.56
368	5.65	0.80	2.97	9.42	9.77	0.25	2.21	12.23
484	4.08	0.53	1.35	5.96	4.29	0.27	0.82	5.38
679	3.74	0.80	3.24	7.77	5.33	0.08	1.48	6.89
Mean	3.94	0.72	2.66	7.32	6.32	0.19	1.61	8.12

Figure 6.6 shows the same information but for the instances based on the standard approach. In the case of the instances based on the standard approach, the vehicles again spend the majority of their time on collection, with offloading and traversing a smaller proportion of total time spent in this case. Table 6.5 further emphasises this by showing

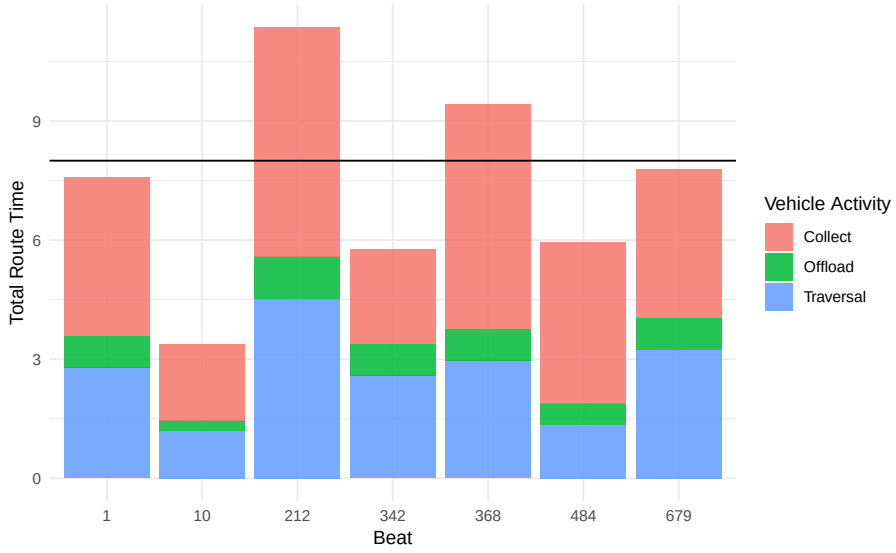

Figure 6.5: Activity times per beat for instances using the refined approach to input data.

activity times per beat, as a proportion of total servicing cost, for instances based on the two approaches. What is apparent from this table is that collection time comprises only 54 percent of activity time for the refined approach, but 78 percent of the time for the standard approach. This is again likely a function of the fact that the standard instance does not differentiate between required and non-required edges. Vehicles therefore have less opportunity to traverse segments as all segments must be serviced.

Offloading times for the standard instances averaged 0.72 hours (43 minutes) as opposed to the 0.19 hours (11 minutes) for the refined instances. It is clear from both these numbers that vehicles make multiple intermediate facility visits throughout their routes, since both numbers are higher than the 16 minute and 5 minute input variables. As a proportion of total time spent the intermediate facility visits take an average of 10 percent for the test set and 2 percent for the standard set. This difference is once again significant as it impacts the feasibility of routes. If routes were to be developed for a metropolitan area using an assumption of a 5 minute offloading cost, the routes could practically prove to be infeasible, as the actual time spent by vehicles offloading would be longer and would mean that the vehicle would likely be unlikely to complete its route within the shift.

Finally, Table 6.5 also shows that the average time spent traversing segments is 2.66 hours for the standard set and 1.61 hours for the refined set. As a proportion of total route time this represents 20 percent in the case of the standard set and 36 percent in the case of the refined set.

6.2.3 Route collection efficiency

A good summary of the information presented above is *collection efficiency*, which is a metric defined for the purposes of this dissertation and based on the principles of Lean, as set out by Poppendieck et al. (2011). Broadly speaking, Lean thinking classifies activities into those which add value, and those that do not add value, with the aim of reducing waste within any particular system. By applying the same principles to waste collection the *collection efficiency* can be defined as the value adding cost (or time) divided by the total route cost (or time).

The value adding portion of the route is the collection, while traversal and offloading

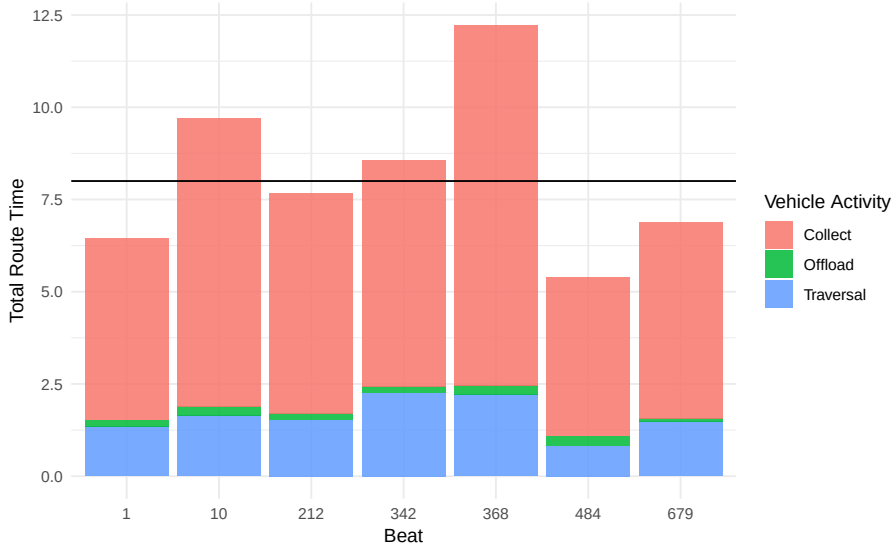

Figure 6.6: Standard set activity times per beat.

Table 6.5: Activity times per beat, as a proportion of total servicing cost, for the refined instances and standard instances

	Refined Instance			Standard Instance		
Beat	Collect	Offload	Traversal	Collect	Offload	Traversal
1	53%	11%	37%	77%	3%	21%
10	57%	8%	35%	81%	3%	17%
212	51%	9%	40%	78%	2%	20%
342	41%	14%	45%	72%	2%	26%
368	60%	8%	31%	80%	2%	18%
484	68%	9%	23%	80%	5%	15%
679	48%	10%	42%	77%	1%	21%
Mean	54%	10%	36%	78%	2%	20%

are considered non-value adding. Figure 6.7 shows the collection efficiency for the beats tested, while the information is also contained in Table 6.6. In all cases the collection efficiency for the refined instances is lower than for the standard instances. This is a direct result of the higher waste generation estimates, the lower vehicle velocities, the higher offloading cost and the lower number of required edges. The collection efficiency for the refined instances is estimated at 54 percent, while the standard instances average 78 percent.

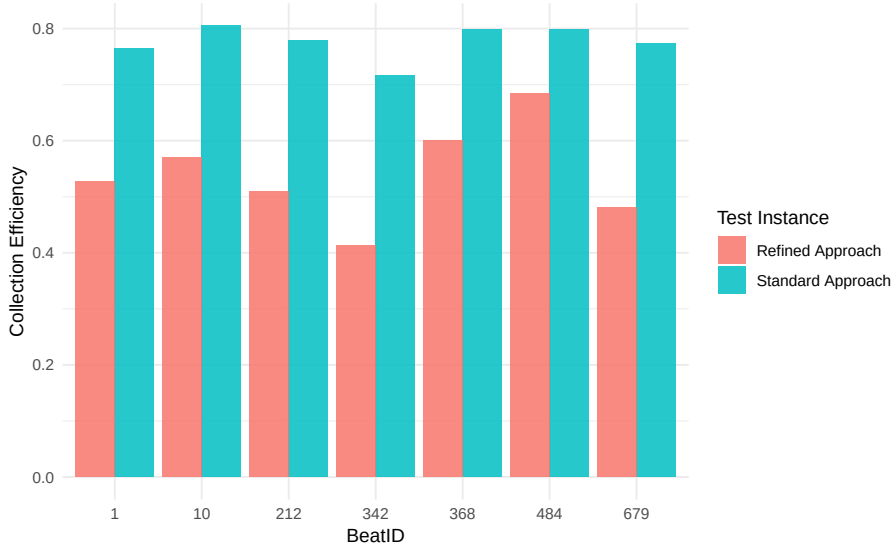

Figure 6.7: Collection efficiency per beat for standard instances versus refined instances.

Table 6.6: Refined and standard instance collection efficiency

Beat	Refined Instance Collection Efficiency	Standard Instance Collection Efficiency
1	53%	77%
10	57%	81%
212	51%	78%
342	41%	72%
368	60%	80%
484	68%	80%
679	48%	77%
Mean	**54%**	**78%**

It is important at this point to clarify the comparison that is presented here. By taking the same service areas (or beats) and solving the collection problems using the same MCARPTIF algorithms, but *differentiating* the input parameters the effect of only the input parameters on the solutions can be evaluated. A lower collection efficiency for the refined instances therefore does not represent a lower quality solution but a solution based on a different set of underlying assumptions and input data than in the standard case.

Consider beat 679 for example. Table 6.4 shows that for this particular beat the solution will allow for 3.74 h of collection and 3.24 h of traversal in the refined case, while in the standard case collection is estimated to take 5.33 hours and traversal only 1.48 h.

In addition the total route length in the refined case is 7.77 h and 6.89 h in the standard case. Consider now for example that both these solutions are provided to a waste manager to aid in allocating vehicles to collection beats. Should the manager be given the solution from the standard case, the manager is likely to allocate this vehicle to another small area to ensure that the vehicle utilises its full 8 h shift. This decision would be based on the assumptions that the vehicle services segments at 20 km/hour and spends 5 min at the transfer stations etc. However in practice, GPS data shows that the vehicle travels at average only 4.13 km/h when servicing this particular beat (see Table 4.10 in Chapter 4 for reference), and will on average spend 16 minutes at the transfer station. Given a solution based on this information, the waste manager will likely allocate this vehicle to only this beat, as the estimated route duration is just less than 8 hours.

What is therefore implied by the fact that collection efficiency is lower in the refined cases than in the standard cases is that the MCARPTIF parameters used in literature are likely too optimistic and produce solutions that differ greatly to those based on the data presented in this dissertation. An interesting future contribution in this regard would be to use the same GPS data set, extract the actual routes taken by the collection vehicle drivers and compare both the refined and standard cases to the actual routes for accuracy.

6.3 Conclusion

Throughout this dissertation methods were developed and tested for estimating MCARPTIF input parameters. These include generation rate, service cost, deadheading cost and offloading cost, amongst others. In this chapter these estimates were used to develop a set of seven *refined instances* which were compared to a set of seven *standard instances*. The refined instances were developed using parameters estimated in this dissertation, while the standard instances were developed using known parameter estimates in literature, as presented by Willemse and Joubert (2016a). Both refined and standard instances were then solved using the MCARPTIF algorithms and the results compared. The chapter aims to demonstrate that input data has a significant effect on collection routes and as such showed how generation rates and activity times differed for routes based on the same beats, but with differing parameters. While neither the standard or the refined approach were compared to actual collection routes and a claim can therefore not be made with regards to accuracy it does demonstrate the estimation of input parameters should be crucial part of the MCARPTIF solution process.

Chapter 7

Conclusion

Accurate, reliable and updated input data is a crucial aspect in the success of designing and improving any system, particularly complex real world systems with various confounding variables. Waste collection is a good example of this, as solution strategies rely both on robust mathematical models and on accurate parameter estimates to improve collection routes. The aim of this dissertation was to address the latter, by showing that publicly available data sets, as well as GPS data from waste collection vehicles, can be used to estimate Mixed Capacitated Arc Routing Problem with Time Restrictions and Intermediate Facilities (MCARPTIF) parameters, and to solve real world instances. The key input parameters addressed include landfill visit durations, segment service and deadheading costs and waste generation rates. In referring back to the research question posed in section 1.2, this chapter summarises the outcome of the research questions, highlights the contribution made by answering the research questions and explores future contributions to literature that this dissertation uncovers.

7.1 Landfill visit durations

Throughout the course of normal operation, waste collection vehicles visit landfills or transfer stations to dispose of the waste collected by the vehicle. The amount of time spent travelling to the particular landfill or transfer station, as well as the time spent at the landfill or transfer station impacts the collection potential of the vehicle in a shift. For this reason the accurate estimation of drop-off durations at waste disposal facilities is crucial to MCARPTIF solution success. Landfill visit durations were successfully estimated for a landfill in a metropolitan area in South Africa, using vehicle GPS data and geofences based on publicly available landfill locations. Landfill visit durations were found to average higher than those used in previous Capacitated Arc Routing Problem (CARP) variants, but was very similar to estimates reported by Wilson and Vincent (2008), which is promising. Attempt were made to fit the drop-off duration estimates to distributions, however these were unsuccessful and point towards confounding variables within the data. The average landfill drop-off duration was estimated at 16 minutes. In addition the effect of congestion was explored as a possible reason for longer drop-off durations within the sample. Contributions in this area include the methods used to derive the above drop-off durations as well as the estimates themselves, which can be used to solve other CARP waste collection variants that incorporate waste disposal.

7.2 Estimating street segment service and deadheading times

Collection and traversal costs proved to be the most challenging of the MCARPTIF parameters to estimate. To achieve this research outcome, GPS data as well as publicly available street network data was used. GPS points were snapped to a street network, and vehicle velocities over the network was calculated. A particularly challenging aspect of calculating velocities per segment was the fact that time intervals in the GPS data were often too long for vehicles traversing shorter street segments. This was however overcome by using velocity estimates from points as the vehicle enters and exits these segments. The next crucial aspect to the analysis of vehicle velocity through service areas was to separate instances of vehicles servicing and traversing segments. To this end a number of variables were tested, namely the service day of a particular beat, the service vehicle of a particular beat, and whether or not a segment was traversed multiple times within a single service day. All of these variables were successful, with the best performing variable being how many times a segment was traversed in a single day. By using these variables together, statistically distinct velocity distributions could be extracted for vehicle service and deadheading velocity in 6 of the 7 test beats. This represents the biggest contribution that this dissertation makes to the body of knowledge. This contribution is significant because vehicle velocity can not be assumed to be constant through different areas or street segments. To solve accurate routing instances therefore requires velocity estimates with sufficient granularity to capture the velocity differences within and between areas. The methods developed in this dissertation provide an opportunity at estimating vehicle velocity for future MCARPTIF problem sets.

In all of the test cases considered, the velocity of vehicles, both when servicing and when deadheading, was significantly lower than those used in CARP problems in literature. Also apparent throughout the analysis was that the velocity differs from area to area and that using the methods described could improve the accuracy of collection routes, and subsequently also collection costs.

7.3 Estimating street segment waste generation rates

The final MCARPTIF variable is that of waste generation rate. For waste collection applications, this variable is particularly important and tricky to compute. This is because, unlike other municipal services such as water or electricity, consumption (or in the case of waste, generation) can not be accurately measured at the household level. Since literature contains various per capita waste generation rates, the approach taken to overcome this problem was to use a synthetic population based on census data to approximate the population size of each street segment and use that to determine the weekly generation rate for the segment. The advantages of this are two-fold, the first is that the effects of population density are considered in the vehicle routing solution. Instead of assuming a uniform population density each segment has a unique density and subsequently a unique generation rate. The second advantage is that better knowledge of the residential population allows for a better understanding of which segments require waste collection. This too is an improvement on estimates made in existing waste collection literature.

7.4 Generating and solving MCARPTIF instances from real world estimates

All of the real world parameters discussed above were combined and tested on seven randomly selected benchmark instances which were solved using the constructive heuristics presented by Willemse and Joubert (2011). These were considered the refined set and were compared to a standard set made up of parameter estimates from Willemse and Joubert (2016a). The aim of the comparison was to use the same service areas, and solve the routing problems using the same solution algorithms, but to vary the input data to test the effect of the input data on the solutions. Overall it was found that solutions varied considerably between the refined set and standard sets and that solutions from the control set had less realistic collection efficiencies. This is due to the higher velocity estimates, lower waste drop-off durations and the fact that the control set did not account for street segments without populations presents. The contribution of this chapter is that it demonstrates the importance of input parameters on solution quality. In addition it demonstrates that with the right data sources it is possible to extract a street network, estimate MCARPTIF parameters and produce a feasible solution. While this process is not fully automated it does prove that a data-driven MCARPTIF model that estimates parameters and produces efficient routes routinely is feasible and practical.

7.5 Future research

While this dissertation did contribute to the body of knowledge, many opportunities remain to expand this field of research towards practical models that drive cost savings for municipal waste collection operations. In the below section some of these opportunities are explored and their viability discussed.

7.5.1 Landfill visit durations

Future work aimed at estimating landfill visit durations could include using the methods presented here to replicate the analysis across all landfills in a metropolitan area, in order to study how visit durations differ between facilities, and potentially to identify the variables that lead to longer visit durations. A good approach to separating drop-off durations between peak times and quieter periods of the day could be k-means clustering. The idea being that clusters of visit durations might be present in the data above and below 13 vehicles (as was visually demonstrated in Chapter 3. This might potentially yield better results.

In terms of CARP variants for waste collection, an interesting addition to the body of knowledge might be a problem variant with different visit durations for each landfill, which would allow vehicles to potentially travel further to a landfill with a faster drop-off duration. Given the findings on landfill congestion presented in Chapter 3, it would also be conceivable to have a CARP variant where a tally of the number of vehicles within each landfill is kept, and where a cost penalty is added for vehicles that visit already congested landfills.

7.5.2 Estimating street segment service and deadheading times

While velocity estimates were obtained for a number of residential service areas in a metropolitan area in South Africa, avenues to better estimates are by no means exhausted. Additional predictor variables could be investigated when it comes to separating servicing

and deadheading activities. An example of this would be speed limits. Speed limits are likely to have a directly proportional relationship to vehicle travel velocity, and the vehicle velocity as a proportion of the speed limit might be insightful as to what the vehicle is busy doing, particularly on busier roads.

Another approach worth pursuing would be one based on clustering, instead of the statistical approach followed in this dissertation. Vehicle velocities would then be separated using unsupervised machine learning using the categorical variables already identified in chapter 4.

Further research could also focus on the effect of traffic. Data on traffic is fairly readily available through service providers such as Google Maps. Research could therefore consider the impact of the time of day on collection routes. For example, it might be faster to service a busy street segment during off peak hours than during rush hour traffic, or to service areas closer to the vehicle depot towards the end of the shift when traffic might delay the vehicle's return to the depot.

Another potential area to explore is that of performance measurement for collection crews. With accurate data on demonstrated route completion times, collection crews could be given collection targets that are realistic and achievable. Collection crews could then be encouraged to meet these targets, which will drive operational excellence. Real time monitoring of these performance indicators could also flag inefficient collection beats where crews are consistently over or under performing.

7.5.3 Estimating street segment waste generation rates

Future work on this aspect of the dissertation could include using weighbridge data to calibrate the per capita waste generation estimate. If weighbridge data was available, each vehicle route could be linked to the actual mass of the waste dropped off. The waste can then be allocated to the synthetic population to produce a better per capita generation rate. The synthetic population can also be used to distinguish between generation rates in high or low income areas, since this data is contained in most census data. Another potential area of interest could be to better estimate collection cost, based on actual waste generation rates. A total route cost could be calculated using the travel distance of the vehicle, fuel cost, maintenance costs and collection crew salaries. This can then be apportioned to the synthetic population to determine which areas are more costly to service. Information like this would be useful in making strategic waste management decisions such as where to build intermediate facilities.

7.5.4 Fully integrated parameter estimation and routing system

While this dissertation tested seven relatively small collection areas, future work could test these data estimation techniques on larger collection areas, or even on an entire metropolitan area. Ultimately the goal would be to fully integrate the process of parameter estimation and live routing algorithms. A database of GPS points, weighbridge data and various other relevant variables would be updated on a daily or weekly basis. The MCARPTIF parameters would then be recalculated and efficient collection routes constructed using the updated data. This would also provide valuable insights into collection operations which could be used to update Key Performance Indicators (KPIs) and drive continuous improvement within waste management operations. Using the techniques discussed and developed in this dissertation, it is likely that cost savings can be realised in future for waste collection operations, particularly in developing nations.

Bibliography

Aliahmadi, S. Z., Barzinpour, F., and Pishvaee, M. S. (2021). A novel bi-objective credibility-based fuzzy model for municipal waste collection with hard time windows. *Journal of Cleaner Production*, page 126364.

Bautista, J., Fernández, E., and Pereira, J. (2008). Solving an urban waste collection problem using ants heuristics. *Computers & Operations Research*, 35(9):3020–3033.

Beigl, P., Lebersorger, S., and Salhofer, S. (2008). Modelling municipal solid waste generation: A review. *Waste Management*, 28(1):200–214.

Belenguer, J. M. and Benavent, E. (2003). A cutting plane algorithm for the capacitated arc routing problem. *Computers & Operations Research*, 30(5):705–728.

Belenguer, J.-M., Benavent, E., Lacomme, P., and Prins, C. (2006). Lower and upper bounds for the mixed capacitated arc routing problem. *Computers & Operations Research*, 33(12):3363–3383.

Benavent, E., Campos, V., Corberán, A., and Mota, E. (1992). The capacitated arc routing problem: lower bounds. *Networks*, 22(7):669–690.

Benjamin, A. M. and Beasley, J. (2010). Metaheuristics for the waste collection vehicle routing problem with time windows, driver rest period and multiple disposal facilities. *Computers & Operations Research*, 37(12):2270–2280.

Bivand, R., Lewin-Koh, N., Pebesma, E., Archer, E., Baddeley, A., Bearman, N., Bibiko, H.-J., Brey, S., Callahan, J., Carrillo, G., et al. (2019). Package 'maptools'.

Boeing, G. (2017). Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139.

Chen, L., Gendreau, M., Hà, M. H., and Langevin, A. (2016). A robust optimization approach for the road network daily maintenance routing problem with uncertain service time. *Transportation research part E: logistics and transportation review*, 85:40–51.

Chen, L., Hà, M. H., Langevin, A., and Gendreau, M. (2014). Optimizing road network daily maintenance operations with stochastic service and travel times. *Transportation Research Part E: Logistics and Transportation Review*, 64:88–102.

Christiansen, C. H., Lysgaard, J., and WøHlk, S. (2009). A branch-and-price algorithm for the capacitated arc routing problem with stochastic demands. *Operations Research Letters*, 37(6):392–398.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Corberán, Á. and Laporte, G. (2015). *Arc routing: problems, methods, and applications*. SIAM.

Dangi, M. B., Urynowicz, M. A., Gerow, K. G., and Thapa, R. B. (2008). Use of stratified cluster sampling for efficient estimation of solid waste generation at household level. *Waste Management & Research*, 26(6):493–499.

Delignette-Muller, M. L. and Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34.

Ehmke, J. F., Campbell, A. M., and Urban, T. L. (2015). Ensuring service levels in routing problems with time windows and stochastic travel times. *European Journal of Operational Research*, 240(2):539–550.

Gendreau, M., Ghiani, G., and Guerriero, E. (2015). Time-dependent routing problems: A review. *Computers & operations research*, 64:189–197.

Ghiani, G., Guerrieri, A., Manni, A., and Manni, E. (2015). Estimating travel and service times for automated route planning and service certification in municipal waste management. *Waste Management*, 46:40–46.

Ghiani, G., Guerriero, F., Laporte, G., and Musmanno, R. (2004). Tabu search heuristics for the arc routing problem with intermediate facilities under capacity and length restrictions. *Journal of Mathematical Modelling and Algorithms*, 3(3):209–223.

Ghiani, G., Improta, G., and Laporte, G. (2001). The capacitated arc routing problem with intermediate facilities. *Networks*, 37(3):134–143.

Ghiani, G., Laganà, D., Laporte, G., and Mari, F. (2010). Ant colony optimization for the arc routing problem with intermediate facilities under capacity and length restrictions. *Journal of heuristics*, 16(2):211–233.

Ghiani, G., Laganà, D., Manni, E., Musmanno, R., and Vigo, D. (2014). Operations research in solid waste management: A survey of strategic and tactical issues. *Computers & Operations Research*, 44:22–32.

Ghiani, G. and Laporte, G. (2000). A branch-and-cut algorithm for the undirected rural postman problem. *Mathematical Programming*, 87(3):467–481.

Golden, B. L. and Wong, R. T. (1981). Capacitated arc routing problems. *Networks*, 11(3):305–315.

Haklay, M. and Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18.

Harland, K., Heppenstall, A., Smith, D., and Birkin, M. (2012). Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. *Journal of Artifical Societies and Social Simulation*, 15(1):1–15.

Hoornweg, D. (2012). What a waste: A global review of solid waste management. Technical report, World Bank.

Huynh, N., Namazi-Rad, M.-R., Perez, P., Berryman, M., Chen, Q., and Barthelemy, J. (2013). Generating a synthetic population in support of agent-based modeling of transportation in sydney. In *20th International Congress on Modelling and Simulation (MODSIM 2013)*.

Jiménez-Meza, A., Arámburo-Lizárraga, J., and de la Fuente, E. (2013). Framework for estimating travel time, distance, speed, and street segment level of service (los), based on gps data. *Procedia Technology*, 7:61–70.

Joubert, J. (2014). Multi level iterative proportional fitting. Technical report, MATsim.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161.

Karak, T., Bhagat, R., and Bhattacharyya, P. (2012). Municipal solid waste generation, composition, and management: the world scenario. *Critical Reviews in Environmental Science and Technology*, 42(15):1509–1630.

Kenyon, A. S. and Morton, D. P. (2003). Stochastic vehicle routing with random travel times. *Transportation Science*, 37(1):69–82.

Kiilerich, L. and Wøhlk, S. (2018). New large-scale data instances for carp and new variations of carp. *INFOR: Information Systems and Operational Research*, 56(1):1–32.

Kinnaman, T. C. (2009). The economics of municipal solid waste management. *Waste Management*, page 2615.

Laporte, G., Musmanno, R., and Vocaturo, F. (2010). An adaptive large neighbourhood search heuristic for the capacitated arc-routing problem with stochastic demands. *Transportation Science*, 44(1):125–135.

Liu, C. and Wu, X.-w. (2010). Factors influencing municipal solid waste generation in china: a multiple statistical analysis study. *Waste Management & Research*.

Lum, O., Golden, B., and Wasil, E. (2018). An open-source desktop application for generating arc-routing benchmark instances. *INFORMS Journal on Computing*, 30(2):361–370.

Manson, N. (2006). Is operations research really research? *Orion*, 22(2):155–180.

Medina, M. (1997). The effect of income on municipal solid waste generation rates for countries of varying levels of economic development: A model. *Journal of Solid Waste Technology and Management*, 24(3).

Minghua, Z., Xiumin, F., Rovetta, A., Qichang, H., Vicentini, F., Bingkai, L., Giusti, A., and Yi, L. (2009). Municipal solid waste management in pudong new area, china. *Waste management*, 29(3):1227–1233.

Mourão, M. C. and Pinto, L. S. (2017). An updated annotated bibliography on arc routing problems. *Networks*, 70(3):144–194.

Müller, K. and Axhausen, K. W. (2012). Multi-level fitting algorithms for population synthesis. *Arbeitsberichte Verkehrs-und Raumplanung*, 821.

Pidd, M. (2010). Why modelling and model use matter. *Journal of the operational Research Society*, 61(1):14–24.

Polacek, M., Doerner, K. F., Hartl, R. F., and Maniezzo, V. (2008). A variable neighborhood search for the capacitated arc routing problem with intermediate facilities. *Journal of Heuristics*, 14(5):405–423.

Poppendieck, M. et al. (2011). Principles of lean thinking. *IT Management Select*, 18(2011):1–7.

Qdais, H. A., Hamoda, M., and Newham, J. (1997). Analysis of residential solid waste at generation sites. *Waste Management & Research*, 15(4):395–405.

Ross, S. M. (2017). *Introductory statistics*. Academic Press.

Saeed, M. O., Hassan, M. N., and Mujeebu, M. A. (2009). Assessment of municipal solid waste generation and recyclable materials potential in kuala lumpur, malaysia. *Waste management*, 29(7):2209–2213.

SANRAL (2009). Design guidelines for single carriageway national roads. Technical report, South African National Roads Agency.

Solid Waste Management (2016). 3rd generation integrated waste management plan. Technical report, City of Cape Town.

Steyn, L. (2016). *Developing a Residential Waste Generation and Service Analysis Model*. Undergraduate mini-dissertation, University of Pretoria.

Steyn, L. J. and Willemse, E. J. (2018). Using vehicle gps data to infer offloading times of waste collection vehicles at transfer stations. In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–6. IEEE.

Taş, D., Dellaert, N., van Woensel, T., and De Kok, T. (2014a). The time-dependent vehicle routing problem with soft time windows and stochastic travel times. *Transportation Research Part C: Emerging Technologies*, 48:66–83.

Taş, D., Gendreau, M., Dellaert, N., Van Woensel, T., and De Kok, A. (2014b). Vehicle routing with soft time windows and stochastic travel times: A column generation and branch-and-price solution approach. *European Journal of Operational Research*, 236(3):789–799.

Troschinetz, A. M. and Mihelcic, J. R. (2009). Sustainable recycling of municipal solid waste in developing countries. *Waste management*, 29(2):915–923.

Wertz, K. L. (1976). Economic factors influencing households' production of refuse. *Journal of Environmental Economics and Management*, 2(4):263–272.

Wilcoxon, F., Katti, S., and Wilcox, R. A. (1970). Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259.

Willemse, E. and Joubert, J. (2011). Constructive heuristics for the residential waste collection problem. In *ORSSA Annual Conference, Elephant Hills Hotel, Victoria Falls, Zimbabwe, 18-21 September 2011, pp 19-28*.

Willemse, E. J. (2016). *Heuristics for large-scale Capacitated Arc Routing Problems on mixed networks*. PhD thesis, University of Pretoria, Pretoria. Available online from http://hdl.handle.net/2263/57510 (Last viewed on 2017-01-16).

Willemse, E. J. and Joubert, J. W. (2016a). Benchmark dataset for undirected and mixed capacitated arc routing problems under time restrictions with intermediate facilities. *Data in brief*, 8:972–977.

Willemse, E. J. and Joubert, J. W. (2016b). Constructive heuristics for the mixed capacity arc routing problem under time restrictions with intermediate facilities. *Computers & Operations Research*, 68:30–62.

Willemse, E. J. and Joubert, J. W. (2016c). Splitting procedures for the mixed capacitated arc routing problem under time restrictions with intermediate facilities. *Operations Research Letters*, 44(5):569–574.

Wilson, B. G., Agar, B. J., Baetz, B. W., and Winning, A. (2007). Practical applications for global positioning system data from solid waste collection vehicles. *Canadian Journal of Civil Engineering*, 34(5):678–681.

Wilson, B. G. and Vincent, J. K. (2008). Estimating waste transfer station delays using gps. *Waste management*, 28(10):1742–1750.