THE FACULTY OF ENGINEERING AND THE BUILT ENVIRONMENT

Department of Civil Engineering – Centre for Transport Studies

The integration of informal minibus-taxi transport services into formal public transport planning and operations – A data driven approach.

Prepared by: Jacobus Frederick du Preez (DPRJAC007)

Prepared for: A/Prof. Mark Zuidgeest & A/Prof. Roger Behrens

October 2018

Plagiarism Declaration

I know the meaning of plagiarism and declare that all the work in the document, save for that which is properly acknowledged, is my own. This thesis/dissertation has been submitted to the Turnitin module (or equivalent similarity and originality checking software) and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.

Signed by candidate

JF du Preez

Executive Summary

The MiniBus Taxi (MBT) mode is poorly understood by planning and operational authorities, yet plays a big role in the economies of developing countries transporting the workforce to and from their places of employment and offering employment to thousands in the operations of these services, as well as the numerous rank-side services and amenities offered to patrons.

In recent years, research focussed on mapping paratransit services, including MBTs, in cities of the developing world has contributed significantly to the understanding of the mode in terms of its spatial extent in its respective service areas.

In South Africa, experience has shown that the wholesale replacement of MBTs with scheduled services is an unattainable goal. Instead, planning authorities and researchers have, more recently, shown interest in investigating feasible methods of integrating the scheduled and unscheduled services as hybrid planned-trunk and paratransit-feeder networks.

The objective of this research is to present the case for simple methods of planning and carrying out onboard surveys of paratransit services to classify and to better understand the operations of individual routes, identified route classes, the network as a whole, as well as revealed passenger demand for the services and, ultimately, how this information can be wielded in the planning and implementation of hybrid routes or networks.

The data central to this study consist of onboard captured MBT data, which was collected with a public transport data capturing application using GPS enabled smartphones in the City of Cape Town from April to August 2017 as part of a City of Cape Town's Transport and Urban Development Authority (TDA) data collection project. The purpose of the project was to clarify the actual extent of MBT services within the City and to improve the representation of the MBT mode in the City of Cape Town's travel demand model.

An Android smartphone application, purpose-built for collecting operational information onboard public transport vehicles, was used to collect spatial and temporal data on the operations of a sample of active MBT routes in Cape Town. The application, which saw some functionality updates specifically for the project, was used to collect the following information per MBT trip:

- Location of stops;
- Time of arrival and departure at stops;
- Number of passengers boarding and alighting at each stop;
- The relative boarding and alighting stop of each specific passenger;
- The amount paid in fare money per passenger at each stop;
- The actual path travelled by the vehicle as a GPS route trace; and
- The origin and destination route description of each route captured.

It is estimated that there are more than 800 active and operational routes in the Cape Town. The objective of the data collection project was to survey each one of these routes for a pre-specified number of trips. As the project was still underway when this research was carried out, the information listed above collected for a sample of trips for 278 routes (556 if the reverse direction is considered as a unique route designation) formed the basis of this study.

During the course of this study, the analyses of these data have shown that while the operational characteristics of individual routes are relatively consistent and stable, it is possible to distinguish between different service typologies within the larger route network.

From the raw data structure listed above, the operational characteristics that were calculated for each trip and aggregated at the route level included:

- Trip and route distances;
- Average operating speeds;
- Travel times;
- Number of stops per trip;
- Load factors between stops along the route; and
- Fare rates and trip revenues.

In addition to the identification of the operational characteristics of the MBT network, service classes and routes, the outcomes of the study include providing a framework of methods for the collection, extraction, cleansing, analysis and visualisation of the data. It also includes the identification of metrics which are key in describing the difference in service types.

The descriptive operational characteristics that were calculated for each trip record, inbound and outbound per route, were evaluated to establish whether they can be used to determine if different service typologies can be observed in the data. It was found that simple *k-means* clustering procedures may be used to classify the routes into separate, distinguishable service classes. For the purpose of this study, it was decided, nominally, that the classification should be executed for three classes. Three was subjectively considered a good value to be inclusive of traditional Trunk and Feeder or Distribution, route types as well as the possibility of the existence of a yet to be defined third type.

The clustering procedures were carried out for different combinations of the operational variables for which the most consistent results were obtained for the combination *distance – stop density*[1] *– passenger turnover*[2]. Analysis of the within-class operational characteristics indicates that these three service classes clearly differ in terms of their stop frequencies, distances, speeds and their spatial network coverage.

[1] Trip distance divided the number of stops
[2] Number of passengers divided by the number of stops

The study furthermore provides evidence that the understanding of the MBT network and sub-networks of service classes within this network, including its interaction with other public transport modes and infrastructure, provides planning and operating authorities with key information for effectively planning and implementing hybrid networks.

Finally, the study demonstrates many additional insights can be garnered from these data by implementing improved statistical sampling and survey methods at the route level and by analysing aspects of the data that were not considered central to the research. These aspects include route adherence studies, origin – destination studies and methods of expanding the onboard data samples accurately by marrying it with data collected during static rank departure and arrival counts.

Ultimately, the study shows that an unprecedented knowledge of the operations of MBT routes and networks may be obtained through detailed yet simple analysis of onboard data and that this knowledge may be very useful in the planning and operations of integrated public transport networks.

Table of Contents

List of Figures

List of Tables

Appendices

List of Abbreviations

AFC	Automatic Fare Collection
APC	Automatic Passenger Counter
AVL	Automatic Vehicle Location
BRT	Bus Rapid Transit
CITP	Comprehensive Integrated Transport Plan
COCT	City of Cape Town
COV	Coefficient of Variation
CPTR	Current Public Transport Record
CSC	Contactless Card System
DoT	Department of Transport
CSV	Comma Separated Values
EFA	Exploratory Factor Analysis
GPRS	General Packet Radio Services
GIS	Geographical Information Systems
GPS	Global Positioning System
GTFS	General Transit Feed Specification
GTFS-RT	General Transit Feed Specification Real Time
IoT	Internet of Things
IRPTN	Integrated Rapid Public Transit Network
ITDP	Institute of Transport Development Policy
ITP	Integrated Transport Plan
KML	Keyhole Markup Language
LCD	Liquid Crystal Display
MBT	Minibus Taxi
NHTS	National Household Transport Survey
NLTSF	National Land Transport Strategic Framework
NLTTA	National Land Transport Transition Act
OD	Origin-Destination
OL	Operating Licence
OLB	Operating License Board
PCA	Principle Component Analysis
PRE	Provincial Regulatory Entity
PTI	Public Transport Interchange
SHP	Shapefile
SQL	Structured Query Language
TDA	Transport and Urban Development Authority
TCQSM	Transit Capacity and Quality of Service Manual
TCRP	Transit Cooperative Research Programme
TRS	Transport Reporting System
TTC	Toronto Transit Commission
WCSS	Within Sum of Squares

1 Introduction

1.1 Background and context

In the African context, and elsewhere in the developing world, paratransit refers to public transport services that are available to the general public and are operated by individuals or small to medium or even large sized private businesses. Paratransit in the developing world context is often referred to as 'informal transport' (Cervero & Golub, 2007).

In South Africa, paratransit refers to the MiniBus Taxi (MBT) industry. MBT transport services, mostly operating as unscheduled public transport services, are, in the most cases, regulated and operated by privately owned businesses. Most commuter taxi services are affiliated with an operator association and these taxi operator associations manage ranking facilities and represent the operators that are licensed to operate a specific route or corridor between ranking facilities.

The MBT industry in South Africa has its origins in the 1950's and is closely linked with the Group Areas Act of 1950. Due to the physical separation imposed between the different races and groups of people, the need for municipal bus services was exacerbated. The 1955 municipal bus boycott and driver strikes led to a large number of drivers losing their jobs. These drivers, who knew the transport network well, began transporting passengers in their private capacity as the demand for transportation began to increase. Over time, the capacity of these vehicles increased from the large saloon cars, Valiants and Chevrolets mostly, to the MBTs seen today (Mxolisis, 2006). Although the industry is often described as informal, and it did indeed have an informal genesis, the MBT industry has managed to penetrate the formal economy (Behrens *et al.*, 2016) and has an estimated annual national revenue value of R40-Billion (Dolan, 2014).

In Cape Town, from which most of the data that will form the basis of this research originates, ±7,500 MBTs (Behrens *et al.*, 2016) are operated. The Cape Town MBT industry is operated entirely by private owners with the majority of vehicles being 14-16 seaters. The City of Cape Town's 2017 – 2022 draft Comprehensive Integrated Transport Plan (CITP) indicates that there are between 9,500 and 10,100 registered operating licences for 16-seater buses (including driver). All other capacity vehicles offering this type of services together make up approximately 600 vehicles.

Services are, in theory, operated on routes for which operating permits are granted by the Provincial Regulatory Entity (PRE). Each licensed route is associated with an Operating Licence Board (OLB) number. While these services operate on a specific route, or a permutation thereof, there are no fixed stops other than the ranking facilities and perhaps a few layby areas along the way. Operators alter their routes seeking to maximise profits by adopting time-saving strategies and following travel demand dynamically (Saddier *et al.*, 2016). Passengers are collected and dropped off along the route at any point, often illegally and obstructing the general flow of traffic.

1.2 Problem Statement

Despite being a R40-billion industry and transporting more than 10 million passengers nationally every day, no scientific network planning and/or operational design efforts go into optimising paratransit services. Due to its informal beginnings, the system and routes have

organically developed into what it is currently seen. Instead of transport planning principles being applied, trial and error, first-hand experience, institutional knowledge and extremely effective “ears-on-the-ground” techniques have been employed in order to ensure these services are as lucrative as possible to the private operator companies. Municipalities and regulating authorities do make an effort to understand these services, issue and regulate licensing permits and provide basic infrastructure, such as ranking facilities. There is, however, no evidence that route designs or operational planning is carried out at any level other than defining the legal route designation on which operator agencies are permitted to ply their trade.

The basis upon which all transport planning is carried out, is the understanding of how, why and when travel occurs. In order to understand the above, and to plan for and design or improve public transport services, it is necessary to have reliable data of both the demand and supply of transport. Methods of collecting these data vary significantly. In the absence of modern ticketing systems and/or passenger counting and/or vehicle tracking technologies, the traditional method of collecting data on existing public transport services consist of manual, paper-based, onboard surveys. For this reason, amongst others, very little large-scale datasets of MBT operations in South Africa exist and prior to this data collection effort, and excluding numerous route mapping projects, no significant onboard data collection of MBT services have been carried out in South Africa. Gaibe (2009) provides early evidence of the insight that can be gleaned from onboard data while Saddier & Johnson (2017) shows how operational data can be collected using smartphone applications. Despite this, no significant efforts have been made to collect and analyse extensive MBT onboard datasets.

Since paratransit services rarely adhere to the fixed routes as described by their operating licenses, it is all but impossible to carry out onboard surveys of paratransit operations using traditional methods of data collection. These traditional methods, while easily implementable for scheduled services which run on fixed routes and stop at fixed points, consist of enumerators undertaking the journey and documenting the time the vehicle arrived and departed at every stop, taking count of the number of passengers boarding and alighting at each stop and recording the fare paid. More modern methods of public transport data collection are possible with modern automated fare collection (AFC) systems and integrated ticketing. These methods enable the automation of data collection on public transport services making the analysis of routes, route segments and stops quite simple.

For paratransit services, however, neither of the above is possible unless the operator has an AFC system that facilitates data collection and analytics. While studies proposing these methods have been undertaken (Van Zyl & Labuschagne, 2008), this is something which is yet to be seen adopted in the 'informal' paratransit services in the developing world.

In South Africa, the adoption of contactless card system (CSC) AFC in the MBT industry has also been investigated. While these systems have obvious benefits to both the owners and the passengers, owners have expressed concerns that these systems showed no benefit towards the driver and drivers are likely to sabotage the AFC systems in order to necessitate cash payments, leaving owners in a worse off financial position (Muwanaula, 2013).

1.3 Research Basis

This research proposes that large scale collection of detailed data of MBT operations using GPS enabled smartphones will provide previously unattainable insights into the industry, how it operates, and how these data can be used to efficiently plan and improve services to the benefit of all stakeholders.

Since the dawn of the age of personal smartphones, the majority of which are equipped with high quality Global Positioning System (GPS) chipsets, and the en masse development of smartphone applications, mapping and data collection of these seemingly chaotic transport services have become much simpler to carry out.

It is possible, using a GPS enabled smartphone equipped with a purpose-built software application, to digitally map the route trace of a paratransit vehicle while collecting all operational characteristics of the route/trip undertaken. Depending on the sophistication of the application used, the information that can be collected in this manner can include, inter alia, the following:

- Vehicle specific details: type, make and model, capacity of the vehicle;
- Location of stops;
- Route trace of the trip;
- Number of passengers boarding and alighting at each stop;
- Gender, age group and ethnicity of each passenger;
- Relative boarding and alighting position of each passenger; and
- Dwell time at stops.

1.4 Research Objectives

The main objective of this research is to investigate how these data can be used to contribute to both the planning and design processes of paratransit services in cities in developing countries. The specific aims of this study are:

- To obtain a clear understanding of the actual operating characteristics of MBTs services;
- To assess whether these data can be used to develop methods of improving the planning of routing and coverage of services;
- To determine how detailed onboard survey data can be used to effectively plan and design MBT services;
- To develop operational metrics of MBT operations that can be used to measure performance of services; and
- To develop a classification system for MBT service types and to determine if this classification system may be useful for integration with other public transport modes.

The desired outcome of the study is to obtain a detailed overview of the physical operational characteristics of the MBT industry and to provide a framework by which the effective strategic, tactical and operational planning of these services, and possibly the integration with other modes, can be carried out.

1.5 Research Questions

Based on the objectives and desired outcomes of the research, the following questions are posed:

- What are the operational characteristics of paratransit services, i.e., average spacing of stops, average passengers per trip or trip segment, operational speed and segment speed between stops?
- What strategic and tactical and operational metrics can be extracted from these data to serve and inform the planning and design processes?
- Can these services be classified in terms of operation type (i.e., trunk or feeder, express or all-stop)?
- Can this information be useful in planning and policy making pertaining to the implementation of hybrid public transport and paratransit services?

1.6 Scope and Limitations

1.6.1 Data Collection Project Scope

As the data upon which this study is based was sourced as part of a City of Cape Town planning project, the scope of the data collection was dictated by the specifications and requirements of this project. As the objective of the project was to determine the extent of services within the City of Cape Town Metropolitan and to determine which routes are active and operational, more emphasis was placed on cost-effectiveness and comprehensiveness in terms of coverage than on collecting data that are statistically representative of the actual operations on the route or network level.

The scope of this project required for each route, that the route is surveyed on two normal weekdays (Tuesday, Wednesday and Thursday), a Friday and on a Saturday. Each route was to be surveyed three times during the morning (5:30 – 9:30) and afternoon (15:00 – 19:00) each and twice during the middle of the day (11:00 – 14:00) and that these trips surveyed consist of an outbound and an inbound component, i.e. a round-trip. The aim was, therefore, to survey each route back and forth eight times on four different days, i.e. for a total of 32 trips per direction.

It is estimated that by surveying a representative sample of vehicles for full day operations and sampling on the route level would cost approximately five (5) to eight (8) times as much as the amount for which this data collection project was costed.

More detail surrounding the data collection is provided in section 3.2.2.

1.6.2 Incomplete Dataset

As the data collection project had not yet been completed at the time of conducting this research, the “complete” dataset was not available. Some routes had been surveyed more frequently than other routes while certain routes had not been surveyed at all. The analyses and procedures undertaken and developed in this study do, however, take this into account, dealing with this bias where appropriate.

1.6.3 Research Scope

Inbound and Outbound Trips

Each route surveyed was assigned a unique route identifier (route ID). For the return (inbound) trip direction, the letter R was used as a suffix. The return direction for the route ID MBT001 would, for example, be MBT001R.

While the direction of the trip surveyed was recorded, determining what the peak commuter direction is for each route, was not considered part of the scope of this research.

Passenger Origin-Destination Information.

The passenger Origin-Destination (OD) information, which is discussed briefly in Section 3.2.2, was collected as part of the data collection project but this information was considered secondary to the purpose of this research and was, therefore, not assessed in any detail.

Data Cleansing

Data cleansing procedures, as described in section 3.3, were applied to the data. The scope and purpose of the cleansing exercise was not to arrive at a dataset that was entirely free of errors but rather to retain data with values that fit within the bounds of sensible operational characteristic values. As the objectives of this study place more emphasis on developing a framework of MBT onboard data analysis than specifically evaluating the City of Cape Town's MBT network, the level of error retained from the raw data after applying some rule-based cleaning procedures, was considered acceptable.

Sole Data Source

The data collected, while not statistically representative of all MBT services' operations, was not expanded or compared to other available data sources, such as the static taxi rank departure surveys that took place from 2013 onwards.

Even though the data collected are not statistically representative, a result of the inadequate sampling methods implemented, the aim of this research is to demonstrate how these types of data collection methods and efforts can be used to create an unprecedented level of MBT service operational understanding through onboard surveys and the analyses of the data. Secondary data sources, such as the static rank departure surveys carried out from 2013 onwards, while offering the ability of expanding the data collected to a daily operations level, were not applied.

Collection Method

The data collection process was dictated by the data collection project's scope. As a cost reduction consideration, enumerators were instructed to leave the vehicle after surveying the trip at the last stop (destination rank) and board another vehicle to survey the return trip to the origin rank. Information related to the ranking time of vehicles and the time required to complete a roundtrip was therefore not captured in this survey process.

Arrangements were also made with some 'queue marshalls', where possible, that enumerators be allowed to skip the passenger queue and get access to the first vehicle in the vehicle queue. This required all passengers boarding at the origin rank to be allocated the same boarding time – passengers were therefore logged once the trip started and no their actual boarding

times. For this reason, no information related to passenger boarding times was captured. Upon arrival at the destination rank (end of a trip) the enumerator would capture the disembarkation of the remaining passengers, allocating the same alighting time for all passengers getting off at the last stop. The duration of the trip is calculated from the difference between the trip start and the alighting time of the first passenger getting off at the last stop of the trip.

1.7 Report Structure

The structure of the report is sequenced such as to guide the reader through the methodology implemented and thought process of the author throughout the study.

This section of the report provides the reader with background to the 'informal' paratransit industry and, more importantly, the study, highlighting the problem statement and listing the objectives and desired outcomes of the research.

The reader is introduced to the data which is central to the study in order to obtain an understanding of the source, structure and limitations of the data.

The aim of the literature review is to provide the reader with a comprehensive account of previous studies that have been carried out that are relevant to this research and to demonstrate what the status quo is in this field of research.

The review of these studies aims to illustrate the potential of how these existing data collection and analysis methods can be expanded to yield previously unattainable levels of understanding of the industry.

This section provides a more detailed overview of the onboard operational data, its collection and it's structure.

This section also gives an explanation of the sequence of processes that comprised the collection, collation, analysis and interpretation of the data in order to satisfy the objectives discussed in the introduction.

The MBT operations section of the report delves into the analysis of the data in order to gain an understanding of both the operations of the MBT services as well as the structure of the MBT network.

The purpose of the analysis is to not only understand the nature of these operations but to identify the variation in service types provided in order to determine which operational characteristics may be instrumental in classifying the MBT routes into differentiable classes.

This section of the report explores methods that may be used to classify MBT routes in terms of their service types.

The understanding gained about the underlying structure and interrelationship of operational characteristics provides guidance as to which of these characteristics describe the most of the variance in the observed data and, ultimately, is used to classify the routes by means of clustering analysis.

Finally, this section of the report provides a visual representation of the spatial distribution of the resultant route classes and an overview of the intra-class distribution of operational characteristics.

The results of the data analysis, visualisation and the outputs of the route classification are discussed in more detail in this section of the report.

The resultant route classes and the distribution and relationships within the full data set, and within each route class, are evaluated in order to obtain a deeper understanding of the services and discuss how these understandings can be wielded to improve planning and operations.

This concluding section of the report recaps the objectives and desired outcomes of the research and summarises the processes, results and key findings against these objectives.

This section of the report identifies further aspects of paratransit operations that could be considered in future research based on similar data collection efforts.

While the research contributes greatly to the understanding and body of knowledge surrounding informal paratransit data collection and planning, it is acknowledged that due to limitations surrounding the sampling and scope of the data collection study as well the specific objectives of the study that significant further insights can be garnered from the data.

2 Literature Review

This chapter provides an overview of the status quo of transport planning related to urban paratransit in developing countries. The purpose of the chapter is to illustrate the importance of the mode and to describe its role in the wider transport network.

A detailed review of literature pertaining to the collection and analysis of formal public and unscheduled paratransit data was undertaken to give an account of the most recent and relevant research that have been carried out.

2.1 Transport Planning

Transport planning refers to the making of policies and decisions and the planning of infrastructure and service investments that are needed for the effective movement of people and commodities from one place to another.

Considering the vast sums of public funding that is needed for the investment in transport infrastructure, it is of paramount importance that all viable alternatives are considered when investment or policy decisions are made regarding transport. These decisions are made on the national, provincial and local levels and should ideally incorporate the input of all stakeholders – from government agencies to the consumers of transport.

In order to make informed decisions on the most appropriate and viable alternatives, as much information pertaining to the specific transport or area related problem is required.

This information is compiled from the various levels of data that are available explaining the transport system and the transport users – the supply and the demand. These data are either sourced directly by survey processes that have been specifically designed for the purpose or are extracted from other internal or external sources.

Data are collected (directly or indirectly), analysed and evaluated in order to make a judgement or decision, i.e. investment, policy change or do nothing. The data collection and analysis step is carried out to create an understanding of the problems faced by the transportation systems and affected communities (Meyer, 2017).

Without good data, of the population and of the existing transport systems, it is not possible to understand the dynamics of how, why, where, when and at what cost people travel, be it on a daily or periodic basis. It is also, on this basis, impossible to understand what the obstacles are that prevent people from traveling. Without fully understanding the transport environment, it is not possible to improve public transport offerings.

Ultimately, transport planning is concerned with the provision of the effective transport systems – infrastructure, services and policies, and at the heart of the planning process is data.

2.1.1 Public transport planning and operations design

Since public transport involves more variables such as routes, stops, schedules, mode choice, vehicle types, subsidies, etc., the planning thereof is inherently more complex than private systems and highway planning (Black, 1995; Vuchic, 2005).

The basis of the planning methodology is the understanding of the demand for travel. While the planning process differs based on the maturity of the network, the fundamental concepts

of demand, capacity, routes, stops, spacing of stops, spacing of routes, frequency of services, fleet size, vehicle size remain the same (Black, 1995).

All of the above is influenced by the unique typographies, land-use patterns as well as political and economic constraints, amongst other factors that are unique to a particular urban area (Black, 1995).

2.2 Paratransit in developing countries

2.2.1 History and nature of paratransit services in developing countries

Paratransit services in the developing world generally originated as a result of poor formal services provision by both the public and private sector. These demand responsive services started emerging where private operators began to provide transport services to cater for the gap in the public transport markets. In Tanzania, the term Daladala was coined to refer to the then illegal private paratransit services that charged one *Dala* (One US Dollar) per trip (Behrens *et al*., 2016). In Nairobi, in similar circumstances, Matatu operators started operating in response to gap in the public transport market in the densely populated area surrounding Nairobi. Both these origins have similarities with the origins of the South African MBT industry, which saw its genesis in response to poor service provision and political boycotts against the then Apartheid government's municipal provide bus services.

In other developing countries in Africa, where regulated bus services were provided by private companies, as a result of government tenders, these services often failed due to mismanagement or political influence and were replaced by a multitude of smaller privately-owned minibus transport operators to serve the community needs (Ousmane, 2008). One of the major drawbacks of these services is that they are mostly operated for profit with no subsidy from governments and, as such, do not serve as a social service (Ousmane, 2008).

Despite their unlikely origins, these services are all major players in the daily transport of passengers in their respective areas of operation.

These poorly regulated, or often unregulated, services are not limited to the African continent. They also exist in many of the cities in developing countries in Latin America and Asia. In most of these cases, paratransit services compete with the publicly owned transport systems and are used extensively.

According to Vuchic (2005), these services thrive in areas where the following criteria exist:

- Low vehicle ownership;
- Poor public transport systems;
- Low cost more important than safety, comfort or reliability;
- Low labour cost;
- Lack of rational transportation policies to coordinate modes and increase system efficiency;
- Lack of organisational ability to introduce priorities for public transport vehicles on roads and streets; and
- Lack of funds for investment in higher order public transport modes.

2.2.2 Share of daily passenger trips

In the developing world, paratransit is often the dominant mode of transit available to the general public, especially captive public transport users. The City of Cape Town's (CoCT) Integrated Transport Plan (ITP) for 2013 to 2018 states that MBTs enjoy approximately 13% of the modal share of all daily passenger trips made (CoCT, 2013). This translates to about 27% of all public transport trips.

According to the 2013 South African National Household Travel Survey (NHTS, 2013), the modal split in Gauteng is 30% in favour of the MBT mode, which is about 72% of the mode share of public transport modes (excluding non-motorised modes).

Although MBTs are a major presence in Cape Town, Metrorail in Cape Town has a much larger share of daily passenger trips when compared to other South African metropolitan areas. In Gauteng, only about 5% of all daily trips are passenger rail trips (NHTS, 2013) whereas, in Cape Town, this is 25% (CoCT, 2013). On average, as estimated in the 2017 National Land Transport Strategic Framework (Department of Transport (DoT), 2017), the MBT mode accounts for about two thirds of public transport trips made in South Africa.

In other African cities, where public transport services and infrastructure are in an even more dire state, this modal share is generally much higher. In Dar es Salaam, Tanzania for example, approximately 60% of all daily trips are made by *Daladalas*. In Nairobi, Kenya, this proportion is 34% (Behrens *et al*., 2016).

These estimated figures are based on combinations of surveys, such as classified counts, rank arrivals and departure counts and operating licensing data. The need for onboard surveys, however, came about because these figures are assumed to be grossly underrepresented due to inadequate survey methods.

2.2.3 Operational aspects

Paratransit services in the developing world generally have the following characteristics (Vuchic, 2005):

- They provide frequent services during peak period and wherever high demand exists
- Service frequencies and provision are low and unreliable during the off-peak periods when demand is low;
- Most vehicles have seating for all passengers but overcrowding during high demand periods results in very low levels of safety and comfort; and
- Paratransit services are self-sustaining and require or receive no subsidies from government.

The size of paratransit service fleets varies from city to city in the developing world (Behrens *et al.*, 2016). The type and size of vehicles operating also varies significantly from city to city from the 14-16 seaters that comprise the bulk of the South African city fleets to the midibuses (30 seaters) and buses (40+ seaters) that operate the paratransit networks in cities such as Dar es Salaam, Nairobi, Bangkok, Port Moresby and Mexico City.

Many of the large cities in these developing countries have very poor public transport systems and suffer from heavily congested transport networks (Vuchic, 2005). The fact that these services operate in general traffic and are poorly regulated and policed exacerbates the

congestion problems. As operators compete for passengers and often stop illegally and obstruct general traffic, safety standards can be considered low (Vuchic, 2005).

Apart from the ranking areas, at route termini, that are, in most cases, provided by the local or provincial government (depending on the city and country), there are often no fixed stops along routes – passengers are picked up or dropped off at any point along any road. These routes also vary in nature from area to area and even within areas – some routes operate services similar to traditional trunk networks, while other operate similar to feeder and distributor routes and provide connection to high demand land uses, other MBT routes and higher order mode networks.

The City of Cape Town's Current Public Transport Record (CPTR) of 2004/5 estimated that approximately 55% of services operate as line-haul or trunk services while 30% and 15% operate as feeder and distribution services respectively.

Fare costs for paratransit are generally higher than both rail and bus modes as passengers value the demand responsive element and short waiting times during peak periods.

2.3 Public transport data collection and analysis

Since the objectives of this research are related to extracting as much information as possible from onboard paratransit data collected, this section forms the focal point of the literature review. Literature that documents the status quo of public transport and paratransit data collection were studied in order to obtain an understanding of the latest collection and analysis methods that have been applied in recent studies and to determine, if possible, how these methods can be applied to this research.

Only literature that focuses on road-based public transport modes have been considered with traditional bus services, bus-rapid transit and paratransit (developing world only) being the service types evaluated.

For the purpose of this research, only data collection related to existing transport systems are considered. Studies related to the planning of transport systems such as household questionnaires and other studies or surveys related to the estimation of public transport demand (i.e. macroscopic demand modelling) have not been considered relevant to this research.

2.3.1 Automated Fare Collection Systems

Many public transport services across the world have implemented automated fare collection (AFC) systems. A fare collection system is a key interface between transit operators and their passengers (Toronto Transit Commission (TTC) Fare Collection Study, 2000:3). These systems consist of fare collection boxes installed in the vehicles and at fare collection points at stations (Sanchez-Martinez & Munizaga, 2016). Depending on the system, passengers can pay for their journeys with cash, paper-based tickets, smartcard tickets or contactless bank cards at these points.

Most modern AFC systems make use of a contactless smart card (CSC) payment system using prepaid transport fare cards. CSCs allow for quick entry and exit to the system where the user merely taps their CSC against the fare box to complete the transaction. Examples of

such systems include Cape Town's MyCiti *myconnect* card, Gautrain's Gautrain Card, London's Oyster Card, Hong Kong's Octopus card and many others.

In addition to collecting the journey's fare from passengers' cards, the AFC system is also able to collect data specific to each fare transaction. This includes the timestamp, and identifiers related to both the fare card and the fare box or gate (Sanchez-Martinez & Munizaga, 2016). Put differently, the system captures the time and place of every transaction.

Across an entire transport system with many vehicles operating on multiple routes, this level of data collection produces high-resolution information related to the performance of each individual element of the system.

Depending on how fare rates are calculated (zone system, flat rate, distance based) and whether the passengers are required to "tap" in and out of the system, the system is also able to collect data that can be used to accurately estimate OD information by connecting the boarding and alighting locations and times of a passenger or smartcard.

Many studies investigating the use of AFC data to improve public transportation planning and operations have been carried out.

Sanchez-Martinez and Munizaga (2016) look at harnessing big data in public transportation. This workshop report discusses not only the uses of AFC data, but looks also at various other sources of public transport related data that can be harnessed and utilised to improve the public transport industry. This report highlights some of the key studies that have been carried in the realm of AFC and GPS data.

Gschwender, Munizaga, and Simonetti (2016) investigate the use of smartcard and GPS data for public transport policy and planning. This study uses Transantiago (Santiago, Chile) as an example of a successful use case of the Santiago public transport system. This study describes the methods developed to extract, inter alia, information such as public transport OD matrices, and public transport vehicle speed profiles.

Munizaga and Palma (2012) present a methodology for estimating public transport OD matrices from smartcard and GPS data. This study uses detailed public transport boarding information to an obtain estimate of alighting time and position. The result of this is a disaggregated public transport OD matrix.

Sanchez-Martinez and Munizaga cite Tamblay, Galilea, and Muñoz (2015) who in their study propose a zonal inference model which enables the estimation of a zonal OD matrix if passively collected smartcard transaction data are available.

2.3.2 GPS Data, Automatic Vehicle Location Systems and Automatic Passenger Counters

Similar to the data collection ability of AFC systems, Automatic Vehicle Location (AVL) and Automatic Passenger Counter (APC) systems are capable of gathering an enormous quantity and variety of operational, spatial and temporal data (Furth, Hemily, Muller & Strathman, 2006).

Furth and Muller (2014) derive methods of determining the distribution of passenger waiting time from AVL data. The study also shows examples of how service reliability can be

measured as the generalised cost of waiting and how improvements in reliability can reduce the generalised cost of waiting as much reductions in headway.

Hemily (2015) discuss the overall uses of Intelligent Transportation Systems (ITS) data in public transportation. This study provides high-level overviews of the potential uses of data generated by AVL, AFC, APC and other systems that produce data such as fault logging, events etc. It also identifies the various challenges and difficulties presented by the production of so much data and makes recommendations on research required to position the industry in order to make optimum use of the ever-growing masses of data that the future will provide.

Van Zyl and Labuschagne (2008) investigate if MBT owners are willing to install onboard passenger counting systems combined with GPS route tracing. The proposed system, which is fitted with an LCD screen which displays geocentric adverts, providing the owner with an income stream that would, theoretically, offset the inconvenience of being monitored. This research also summarises the objectives and requirements of the 2007-2020 South African Public Transport Strategy (DoT, 2007a) and 2007 Public Transport Action Plan (DoT, 2007b). These documents emphasise the importance of Integrated Rapid Public Transport Networks (IRPTN) and how MBTs must play an important role as feeders to the core rail and road trunk corridors.

Van Zyl and Labuschagne (2008) confirm that the status quo in collecting CPTR data on MBT operations is by rank surveys where the arrival and departure of all vehicles and the boarding and alighting numbers are captured. Off-rank roadside monitoring surveys are also carried out to capture the operational activities of vehicles that do not load or unload passengers at the ranks. Van Zyl and Labuschagne (2008) carried out experimental data collection efforts by installing in an operating MBT the following equipment:

- Onboard computer;
- GPRS modem and antenna;
- GPS antenna and receiving module;
- Door sensor;
- Security cameras; and
- An LCD monitor.

The above was installed in a taxi operating a route in Cape Town. The data collected was, in essence, similar to the structure of the MBT data collected and analysed for this research but on a much smaller scale and with a higher degree of automation in collection methodology. The outcome of the study is demonstrating the usefulness of statistical data related to MBT operations with information on route adherence, location of stops, route segment occupancy, etc.

A similar study (2013 – ongoing) is being carried out by others in Tshwane, South Africa to determine the operational business value of the minibus taxi industry as part of the Tshwane Rapid Transit (TRT) *A Re Yeng* Bus Rapid Transit (BRT) Project. This study, commissioned by the City of Tshwane, monitored the full-day operations of samples of MBT vehicles for a pre-defined period of time and the sampling was carried out on the operator association level in order to get a representative measure of the specific operation. This was repeated for a number of routes and operator associations in order to determine indicative business values of

individual operators and to calculate the level of compensation required to negotiate an agreement with operators to give up their operating licenses and discontinue operating certain routes and/or corridors.

2.3.3 Paratransit Data Collection

The traditional methods of MBT data collection in South Africa have their limitations. One of the largest shortfalls, as pointed out by Gaibe (2009), is the fact that most MBT surveys are conducted at ranks where either the arriving or the departing vehicles and their occupancies are recorded. This obviously does not account for the turnover of passengers (boarding and alighting of passenger) along the route between the two end-points. It also does not account for the probable deviation in route that MBTs exercise to increase their passenger capture if needed. These surveys are generally paper based and require a lot labour hours to capture, process, clean and convert into useable formats to transfer the data into useful information.

Moodley, Aucamp and Wood (2005) acknowledge that traditional methods are useful in giving an indication of the current services but that some route operations cannot be surveyed using traditional methods.

The methods employed by Van Zyl and Labuschagne (2008), Gaibe (2009) and the *A Re Yeng* study address some of these shortfalls.

2.3.4 Crowdsourced Data

Jeff Howe (2006) coined the term ‘crowdsourcing’ in a 2006 edition of Wired Magazine when describing the rise of online stock free image-sharing exchanges as opposed to outsourced professional stock photography companies. A further example he wrote about is open source software coding and how groups of volunteer coders could write code just as well as professional software developers employed by companies, such as Microsoft or Sun Microsystems but at a fraction of the price. Crowdsourcing refers to a “participative online activity where a heterogeneous group of people voluntarily undertake a task of mutual benefit (Estellés and González, 2012). A good example is Wikipedia, which exists as a result of the same model where the collective knowledge of all contributors grew into a comprehensive encyclopaedia (Howe, 2006).

As technology has advanced and internet access became more widely available, so has the number of industries served by this collective contribution of users.

In the public transport crowdsourcing realm, much emphasis has been placed on providing real time information of public transport services and measuring the reliability performance of these services against their scheduled arrival and departure time.

Lau and Ismail (2015) discuss the benefits of passengers using smartphones to provide real time information of bus locations and congestion estimation as opposed to the transport service providers supplying the information to the benefit of the commuters.

An example of the above is the smartphone application Moovit, which allows transit passengers to passively or anonymously collect real-time public transport information. Users of the application transmit their speed and location data to Moovit who then in turn supplies the aggregated data, together with the public transport schedules back to other users of the application in order to enhance the trip planning ability of the community. The application

also allows passengers to report on delays and the reasons for delays, how full the bus/train is and other information that might be useful to commuters.

The Moovit End User License Agreement describes the application as *"a platform for planning your trips through use of public transit in certain countries. The platform enables you to a. view the location of transit line stops of certain transit agencies supported by the application and any related static information (name of Transit Agency, Transit Line Stops locations, transit lines timetables, frequencies and travel routes); b. plan and optimise your trip based on comprehensive proprietary trip planning algorithms which combine the public transit information with certain "wisdom of the crowds", dynamic information which derives from other users of the platform, whether in real time and whether based on an estimated bases based on other users' previous experience; c. post certain content, which would later appear on the platform, next to each of the relevant transit line stops; and d. communicate with other users of the application"*.

This sort of application does not only provide a new level of convenience to the users of public transport by providing them with real time trip planning abilities, the amount of data collected over time provides a massive amount of repository of information that provides insights into how, where, when and why people move around a city.

These data can assist authorities in developing high-resolution origin-destination matrices as well as aggregated and anonymised data on the attributes of the users or passengers.

This application, however, and others like it[3], is concerned with scheduled public transport services as they rely on comparing the user supplied data with schedules and analysing these against each other to report back to the user. Converting schedules to General Transit Feed Specification (GTFS) and General Transit Feed Specification Real Time (GTFS-RT) and making it freely available allows any person capable of creating their own application or website to public schedules and/or real time information of whichever agency's data they wish to use. GTFS and GTFS-RT both also rely on schedules, fixed stops to be in a usable format.

Cairó, Salcedo, Gutierrez-Garcia (2015) investigates methods of sourcing user knowledge to profile unscheduled transport services by analysing daily travel itineraries of public transport commuters in Mexico City, Mexico. This paper proposed a consolidated approach to crowdsource knowledge related to unscheduled public transport services provided by a number of unrelated service providers (city government or private operators), information not supplied by officials to commuters. This research acknowledges that solutions such as open data and GTFS cannot be applied to unscheduled and unregulated public transport systems such as Mexico City's. The services are not only unscheduled, there are no real time sources of data on any of the operations. From this research came the refinement and application of the *expert user heuristic* (Sendra and Cairó, 2014), a heuristic that uses the collective knowledge of the expert public transport users that computes the shortest path (analogous with lowest generalised cost) from a given to a given destination based on distance, transfer times, waiting times and speed of public transport services. This information is used on a smartphone application platform, SmartPaths, the objective of which is to design efficient routes for unscheduled public transport trips for users in Mexico City.

[3] Citymapper, Here, TriMet NextBus, GoMetro

Ching *et al.*, (2013) introduced two approaches that are different from crowdsourcing in that a specific group of people carry out the 'sourcing' task. These approaches, "fleetsourcing" and "flocksourcing" (forms of guided crowdsourcing) were used to collect data on the bus systems in Dhaka, the Capital of Bangladesh.

Fleetsourcing refers to the use of smartphones as AVL systems within a fleet of public transport vehicles using either the driver or the onboard ticket agent as data collector (Ching *et al*., 2013). The smartphones may also be programmed to collect other information such as ridership and overcrowding, etc., acting as a passenger counting device.

Flocksourcing refers to the outsourcing of focussed data collection to a specific team of people for a given purpose (Ching *et al*., 2013). This method of data collection was used to collect the Cape Town MBT data on which this research is based. Data collection teams are trained in the use of smartphone application which they will use to collect the data with.

The objective of the Dhaka study was to determine the viability of flocksourcing and technical feasibility of fleetsourcing (Ching *et al*., 2013). The data collection effort was carried out by eight local people over a week, collecting data on two bus lines. The data collection can best be described as an onboard passenger survey with GPS vehicle route tracking.

The data that were collected included, inter alia, the following: Demographic information, trip purpose, satisfaction levels and qualitative levels of overcrowding.

2.3.5 Informal Transit Route Mapping

Recent years have seen a rise in the use of flocksourcing principles to use smartphone devices equipped with GPS chips to map the alignments of scheduled and unscheduled transit services in the developing world, providing passengers with information on available routes that have not otherwise been provided by operators or authorities.

The Digital Matatus Project (Klopp *et al*., 2014; Williams et al., 2015; Klopp *et al.*, 2015; Klopp & Cavoli, 2017, Klopp *et al.,* 2017) was one of the first initiatives that used smartphones with GPS devices to collect data and map unscheduled and semi-formal (or informal) public transport services in Nairobi, Kenya.

Other studies were being carried out in Dhaka (Ching *et al.*, 2013) and Mexico City (Cairó *et al.* 2015) more or less during the same period (data collection efforts on these projects were all around 2012).

The Digital Matatus project *flocksourced* Geographic Information System (GIS) data by having a team of students ride the buses[4] (matatus) and map the routes using smartphone applications that were already available to download. A number of applications were tested but the application that was found to be most suitable and adaptable to their data collection needs was the *MyTracks* App developed by Google for Android devices.

The data that were collected during this study included the GPS route trace and information related to stops along the way such as if there was a bench or shelter, the stop's name or visible signage (Klopp *et al*., 2015). Where GPS service quality was lost and incomplete

[4] In some cases the students would ride the routes in private vehicles adding the stop location details separately.

routes were found, the routes were fixed manually using GIS software. GPS quality and accuracy were, however, in most cases found to be comparable with hand-held GPS devices.

One of the main objectives of this study was to collect GIS data and to make the data directly available to the public, in the form simplified maps, to provide them with comprehensive information about routes, stops and schedules (where they exist). Another objective of the was to data on these services and format the data in an open format such as GTFS and to make this data available to external parties to develop tools such as trip-planning applications (Klopp *et. al,* 2015) for the benefit of commuters.

Since GTFS requires schedule information and the matatus do not follow fixed schedules as in formal public transport, hypothetical schedules had to be created in order to feed into the route planning software Klopp *et al.,* 2015).

In essence this project mapped the previously unmapped public transport routes in a rapidly developing country. While this information is very useful for commuters who previously relied on word of mouth and personal experience to get around, the information generated from these data was just as useful for the operators to optimise their services, and planners of a BRT route in Nairobi who needed to get an understanding of the service coverage to plan their ridership and frequency surveys (Klopp *et al.*, 2015). The data have also been used by external development or finance entities such as the World Bank, the European Union and the African Development Bank to carry out spatial studies such as community access to health and other facilities (Klopp & Cavoli, 2017). This study also led to the United Nations Habitat (UN-Habitat) approaching the Institute of Transport Development Policy (ITDP) to carry out a similar study in Kampala (Uganda) and another Kenyan City, Kisumu (Klopp & Cavoli, 2017)

This project serves as a great example of how data can be collected on informal and semi-formal public transport to generate knowledge of the operations. Klopp *et al.* (2015) acknowledges that more research needs to be carried out to improve these data collection tools and techniques, not only to provide commuters with travel information but to also to improve analysis and modelling of these services in order improve infrastructure and operations.

Klopp and Cavoli also present the 2014 project “Mapa Dos Chapas” that was carried out in Maputo, Mozambique. The chapas are the Mozambican equivalent of Kenyas matatus, Accra’s trotros and South Africas MBTs. This study was very similar in nature to the digital matatus project where GPS coordinates of minibus routes and frequently used informal stops were collected but dissimilar in that the information was verified and approved by chapas industry representatives and Maputo City Council (Klopp & Cavoli, 2017). The resultant map from this project, similar to the matatus map, has been adopted by chapas associations and planning authorities, giving the chapas network additional importance in the planning processes.

Ndibatya *et al.,* (2016) present a similar mapping exercise that was carried out in Kampala, Uganda. The mapping was carried out using a custom-developed Android App called GoMetro Pro (an updated version of this App was used for the data collection in this study). As in many other developing countries, formal scheduled public transport services do not exist and instead privately owned minibus taxi type services are operated to fill the void.

The objective of this study was to gain an understanding of the network in geospatial terms, the premise being that this understanding will enable better decision making in order to improve the current service offerings. The study consisted of the following phases:

1. Establishing the location of taxi ranks and routes originating from the ranks;
2. Tracing the routes by having volunteers board the taxis from the ranks and riding selected routes, mapping the route trace using the GPS mapping App GoMetro Pro, tagging the location of informal stops along the way; and
3. Post data collection processing where data points were processed to ensure consistency and accuracy and converting data into a format to be used with GIS software.

One of the outcomes of this research was open data on informal public transport that could be used by authorities, software developers, researchers and entrepreneurs to improve travel for the customers of these services. The results also included a sample schematic map of all roads, taxi routes, taxi ranks and stops. In essence, this research comprised a data mapping exercise in order for stakeholders to be able to visualise and understand this component of the public transport network. The scope of the research did not include the collection and analysis of operational information related to these services.

2.3.6 Beyond Mapping

Gaibe (2009) and Gaibe and Vanderschuren (2010) present a 2008 onboard minibus taxi survey that was carried out using GPS devices – the first study of its kind that the author is aware of. The study was carried out as part of a larger planning project that was being carried out for public transport tourism services in Stellenbosch in the Cape Winelands District in South Africa. Gaibe (2009) uses this data collection as a case study to review the current (2008) CPTR data collection of MBT services.

The need to carry out the onboard surveys of MBTs arose from gaps that existed in the ability of the CPTR to provide detailed MBT passenger information. These gaps exist as result of the static data collection methods that are usually employed to capture the passenger and vehicle arrivals and departure at ranks.

GPS technology was used to ‘geo-code’ boarding and alighting stops and to create traces of the routes the taxis travelled. A paper-based survey form was used to capture the number of passengers boarding and alighting at each stop en route. Each passenger was handed a numbered ticket upon boarding and was asked to hand back the ticket when they alight, enabling the observation of passenger origin-destination information along each route. Each surveyor was assigned to a specific vehicle which they rode out all the trips the vehicle made during the day, recording the passenger volumes and stop locations mentioned above.

In summary, the study derived the following information from the onboard surveys (Gaibe & Vanderschuren, 2010):

- Total passengers boarding per taxi per route;
- Passenger travel times;
- Vehicle travel times;
- High boarding and alighting locations, en route or at ranks;
- Passenger origin and destination pairs;

- Geo-coded information on all alighting and boarding points;
- Exact routes travelled per taxi and how and when taxis deviate from Operating Licensing Board (OLB) route;
- Peak and off-peak vehicle utilisation;
- Total time taxis are not traveling i.e. ranking;
- Total number of taxis on a particular route and taxi operating hours;
- Total number of trips per vehicle; and
- Vehicle and rank capacities.

The outcome of this study was being able to determine exactly how MBTs operate on these routes and comparing this information to the previous round of CPTR passenger demand data collected. The study showed that a high number of passengers were collected outside of the ranks and that indicated that most taxis were not always adhering to their designated routes – i.e. not conforming to their permit requirements (Gaibe & Vanderschuren, 2010). Gaibe concludes that CPTR (now 'Transport Register') is an important tool for transport planning but that the information is not complete for flexible services, such as MBTs, and therefore require onboard type surveys to fill in the complete picture of their operations and passenger demands.

Saddier *et al.*, (2016) and Saddier *et al.*, (2017) discuss the *Accra Mobile Experiment*: the use of GPS smartphones to capture data and map onboard paratransit services in Accra. What set this project apart from the mapping efforts carried out elsewhere was the objective of providing the local transport authorities with operational data for transport planning (Saddier *et al.*, 2017). Saddier *et al.* (2017) show how these data could be used to quantify the reliability of these services in terms of travel time and waiting time variability. The data collection and mapping was carried out in Accra, Ghana in 2015 on the Ghanaian paratransit services called *trotros.* The mapping covered 315 routes, each surveyed for one round-trip between their respective origin and destinations. Similar to the Cape Town services, these services operate between fixed origin and destination points (stations or ranks) but vary in terms of departure frequencies, number of stops and travel times along the route and route adherence is generally a function of the profitability of travel demand (Saddier *et al.*, 2017).

The data collection effort carried out onboard these trotro vehicles over a week, recording over 1,200 trips on 315 routes. Each route was surveyed only once per direction, but this information was supplemented by station departure surveys that were carried out for one full week at each station. Theses station departure surveys were carried out in a similar manner in which the CPTR (now Transport Register) data have traditionally been collected.

Saddier *et al.* (2016) analysed aspects of the routes such as travel time, route distances and route/trip fare. Although the sampling was fairly limited, the analysis showed that a fair amount of variability of route characteristic can be observed. The data collection was carried out for the forward and reverse direction of each route (i.e. round-trip) so the variation in operational characteristics between the two directions could be measured. The study found that reverse direction trips were on average 43% shorter in terms of travel time while the median value of this measure was approximately 30% (Saddier *et al.*, 2016). The same comparison was drawn between the trip distance on a route between the forward and reverse trips, finding that the forward trips varied in length by approximately 17%.

Since the GPS coordinates location and timestamps of passenger boardings and alightings were collected en route, Saddier *et al.* (2016) were able to develop a spatial distribution of transportation demand. This level of information could provide key insights into where infrastructure investment could be prioritised to improve the operations of these services. The visual mapping of the routes, on the other hands, identifies which areas are underserved by transport services and can be used to design incentives to more equally distribute transport supply to underserved areas (Saddier *et al.*, 2016).

Saddier *et al*. (2017) looked at three aspects of service reliability that were analysed in these data; headway variability, travel time variability and route itinerary variability. The first two of these aspects are self-explanatory while the third is analogous with the concept of route adherence in capturing the flexibility of paratransit routes.

The findings of this research suggest that these services are relatively stable in terms of these aforementioned aspects, which in turn suggests that these services are not too dissimilar to formal bus services and that reform in the paratransit sector is and should be an achievable goal (Saddier *et al.*, 2017).

One important finding of this study was the demonstration of the relative simplicity, timeliness and frugality with which these types of data collection endeavours may be executed and with methodologies that are transferable to many other cities in the developing world.

Saddier and Johnson (2018) discuss a second wave of onboard surveys (full day operations of a sample of vehicles) and static ranks surveys that were carried out in Accra to establish an understanding of the operational characteristics of the *trotros* paratransit services. Detailed surveys were carried out onboard vehicles on six routes for a week in May 2016 while station departure counts were carried out for a full week in June 2016. This method of marrying the operational data from the onboard data with the departure frequencies and load factors leaving the ranks enable the study to calculate the total number of passengers transported on each route per day, the amount of revenue made per route per day of the week, the average headway between vehicles, the load factors on separate route segments and vehicle waiting times in queues.

Saddier and Johnson (2018) purposefully separated inbound and outbound trips for the difference peak periods to enable the determination of the differentiated route characteristics on the forward and reverse directions of each route. The study found that since vehicles typically only departed from the origin ranks once they were filled, vehicles spent more time waiting in queues than driving (only about 3.5 hours a day on average was spent transporting passengers). This also meant that vehicles were often to full to collect passengers at some of the common stops along the way. Vehicles therefore made optimal use of their seating capacities but the system itself was inefficient.

The operational characteristics that were quantified for each route included route lengths, operating speeds, travel times and passengers transported per trip, information that was expanded using the static station counts to develop a profile for each route for a day and different days of the week.

Coetzee *et al.*, (2018a) present a paper that discusses an "innovative way of conducting onboard vehicle surveys for minibus taxis" (Coetzee *et al.*, 2018, para. 1) using the GoMetro

Pro mobile app. This paper is based on the same data collection project and data used in this research and compares it to studies using the same technology, but different sampling methods, elsewhere – Rustenburg and Nelspruit.

Coetzee et al., (2018b) present a method of using this technology, with adequate route-level or operator level sampling, which enables the determination of business values of MBT operations. It focusses on a study carried out in Rustenburg for the development and implementation of the Rustenburg Integrated Rapid Public Transport Network (RIRPTN) – a project on which the author of this research was involved, carrying out the initial data analyses and assisting in the development of the preliminary revenue calculation methods. Sampling was carried out at the association level, making sure that a representative sample of the entire association's fleet was surveyed for full day operations (04:00 to 21:00) over at least eight days. Unique registration numbers were randomly selected from the list of vehicles registered with each association which enabled the capturing of all trips made during a range of days, capturing all passengers boarding and alighting and the fare revenues paid by passengers. Approximate daily profits or losses per vehicle could be calculated using the revenue information collected from the onboard surveys, operating cost estimations based on the distances and speeds of the vehicles and capital costs (i.e. vehicle purchase repayments).

2.4 Résumé

The research covered in this literature review that relates to scheduled public transport data only just begins to touch the subject. This is an area that has enjoyed much attention, especially so in the last five years with rapid advancements in technology, the emergence of "systems" such as the Internet of Things (IoT) and the ubiquitous use of smartphones in all aspects of day-to-day living which have produced large amounts of data related to the users and operation of public transport.

Studies related to data collection and analysis of paratransit services in the developing world are, however, limited in comparison. These studies have also either been fairly limited in their scope or in their coverage. Efforts have mostly gone into determining the extent of these services and to provide operators and/or authorities with geospatial information that can be used to improve the services or to provide the users with more information they can use to plan their journeys.

Gaibe and Vanderschuren (2010) allude to the potential of connecting geo-coded route data with operational survey data in describing the characteristics of services and the demand for them as opposed to the traditional rank surveys that inform CPTRs.

Ndibatya *et al.*, (2016) present a tool, GoMetro Pro mapping App, which is capable of carrying both the geo-coding of the trip route and stops collecting the passenger boarding and alighting points and volumes all within one smartphone application platform.

Deeper investigation of data collected using onboard smartphone methodologies (Saddier *et al*., 2016; Saddier *et al.*, 2017) have demonstrated the value that these techniques may have to planning and operating authorities and that these surveys may be relatively economical in comparison with traditional methods. Even more in-depth investigation into the operational characteristics has been demonstrated by Saddier & Johnson (2017) using the same methods of data collection.

Coetzee *et al*. (2018a) and Coetzee *et al.*, (2018b) present two studies which focussed on the data collection using the GoMetro Pro app. The first of these studies discusses the same data, and some uses thereof, that this research is based on. These studies, however, only begin to touch on the potential uses of such data.

The culmination of these various studies has demonstrated how simple and relatively inexpensive methods, assisted by technology, can be used to garner unprecedented levels understanding of paratransit service operations.

The research reviewed has revealed that these studies have either been very limited in their scope, i.e. the Digital Matatus project which put a lot of emphasis on mapping a network, or very limited in terms of their coverage, i.e. the 2017 Accra study which was very detailed in terms of the sampling and analysis of a select number of routes (Saddier *et al*., 2017; Saddier and Johnson, 2018).

What is lacking from this body of research, is the development of methods that enable the understanding of the operational and service coverage aspects of the paratransit network in relation to itself as well as of the operations of specific and individual routes within this network and how they fit into the greater public transport network.

3 Methods

3.1 Research Design

Having reviewed the literature related to the collection and analyses informal paratransit data, this chapter deals with the collection and analysis of the data on which this study was based.

Since the primary objectives of this research were to, inter alia, develop methods of analysing onboard data of unscheduled services; to apply these methods to data collected in the City of Cape Town in order to develop a better understanding of the local MBT operations; to use the findings to improve service offerings and to develop simple methods that can be used to enrich data collection, analysis and visualisation of future informal transit mapping initiatives, the structure of the data driven component of the research is as follows:

- Data collection;
- Data cleansing;
- Data analysis and visualisation to obtain an understanding of the underlying structure of the data and a high-level understanding of the nature of local MBT operations;
- Classification of service typologies according to observed variables in the cleansed data;
- Spatial and temporal visualisation of observed service typologies and their interaction with the operating environment;
- Further analysis of the intra-class variability of operations; and
- Discussion on how these methods can be applied to studies of a similar nature or enrich studies that aim to map paratransit networks.

3.2 Minibus Taxi Onboard Data Collection Study

3.2.1 Background to study

Onboard surveys of the City of Cape Town's MBT routes were carried out between April 2017 and March 2018 as part of a City of Cape Town contract[5].

The purpose of the data collection was to clarify the actual extent of the MBT network in Cape Town and to update and improve the minibus taxi mode in the City of Cape Town's travel demand model which is poorly understood and represented in the existing model.

The latest review of the of the City of Cape Town's comprehensive integrated transport plan (CITP) estimated that approximately 800 MBT routes currently operate in the metropolitan area. According to the City of Cape Town's Transport Reporting System (TRS), there are 102 operator associations operating between 185 ranks with between 7,000 and 10,000 unique vehicle registrations in operation, carrying about 550,000 passengers a day.

A total of 6,035 unique registrations were observed at the various ranks in Cape Town between January and March 2013 and it is common knowledge that there are many illegal operators that do not terminate at the ranks (City of Cape Town CITP, 2017). It is estimated that 46% of MBTs operating were doing so illegally.

[5] Project Number: 3652/185C CoCT Survey and Data Collection - Onboard Minibus Taxi Survey

The City of Cape Town acknowledges that the MBT industry in Cape Town is not understood well enough and, as such, advertised a tender for the collection of onboard route and passenger data in October 2016. This tender was awarded to a public transport mobile application development and data collection start-up company called GoMetro (referred to in literature review). The author joined this company in a project management and specialist consulting capacity in December 2016 with work starting in the planning of the data collection immediately. A pilot study (proof of concept) was carried out between January and April 2017 with the full study commencing 12 April 2017.

The scope of services required as specified in the tender document (City of Cape Town, 2016) can be summarised as follows:

- Onboard surveys of MBT routes which operate within the City of Cape Town boundaries;
- Collection of the location of all stops along these MBT routes together with the number of passengers alighting per stop or ranks;
- The method of survey must be able to track the routes and count passengers to determine the passenger demand on all MBT routes in Cape Town;
- All information must be electronic and geo-referenced;
- Taxi route identification, passenger link volumes and stop boarding/alighting information must be provided in a universal GIS format;
- Manual or electronic survey methods will be considered but the tenderer must be able to show the estimated level of accuracy as well as the type of cross checks or quality control mechanisms that will be built into the process;
- Each round trip MBT route must be surveyed the following number of times:
    - Three taxi trips in the peak periods (AM and PM); and
    - Two taxi trips in the off-peak period (better defined as inter-peak);
- These trips will be surveyed on two typical weekdays (Tuesday – Thursday) as well as a Friday and Saturday and no surveys will take place during school holidays or on public holidays;
- Routes travelled will be measures with a GPS, including:
    - GPS track in shape file or Keyhole Markup Language (kml) format
    - Timestamp of track
    - Vehicle information i.e. registration number, type of vehicle
    - Route information i.e. origin-destination of route and route name
    - Trip time per route
- All stop locations where passengers board or alight en route, including:
    - Coordinate of stop;
    - Timestamp of stop;
    - Number of passengers boarding at stop;
    - Number of passengers alighting at stop;
    - Total number of passengers onboard between stops;
    - Fare information per route;
- Stakeholder consultation with MBT industry to facilitate surveys.

The peak periods mentioned above were defined as follows:

- AM peak period: 5:30 – 9:00;

- Inter-peak period: 11:00 – 14:00; and
- PM peak period: 15:00 – 19:00.

Trips surveyed needed to start within these time frames in order to be counted as a legitimate trip belonging to the peak period in the collected dataset. In other words, a trip that started at 09:31am did not fall into any of the peak periods.

As detailed in the limitations in section 1.6, the total number of trips to that were required to be surveyed per route was 32; 8 x 2 on typical weekdays, eight (8) on a Friday and eight (8) on a Saturday. This was, however, not achieved in all cases with some routes being surveyed more frequently than others during the period of data collection.

3.2.2 Data Collection

The GoMetro Pro Android smartphone App underwent two large and fundamental functionality updates specifically for the collection of the data on this project. The App was used during the Kampala study (Ndibatya *et al.*, 2016) but only the mapping function was used. The application also had, before the Cape Town study, the ability to count passengers boarding and alighting en route and at ranks and geo-coding the passenger stops along the route traced. This allowed each passenger boarding and alighting location to be referenced in a spatial, as well as a temporal manner.

A variation to the original project scope required origin-destination information for passengers to be collected as well as the operational information. A system of handing out numbered tickets to passengers (Gaibe, 2009; Gaibe & Vanderschuren, 2010) was trialled during the pilot study but was not found to work effectively for a number of reasons; (1) passengers were averse to interaction with surveyors, (2) this required a considerable amount of post-processing and data capturing, which was not budgeted for and (3) unacceptable amounts of errors were found using this method.

An update to the App was developed during the pilot study where the boarding and alighting of each passenger was captured. In order to protect the proprietor's (GoMetro) Intellectual Property (IP), the exact mechanics of this method will not be described but essentially this allowed the connection of the relative boarding and alighting time and location of each passenger along the route.

The onboard data was collected using the GoMetro Pro App, the user interface of which is shown in Figure 1, using smartphones and uploaded to a cloud-based central databased at the end of each shift. From this online database, the data could be analysed online using Structured Query Language (SQL) queries or exported in comma separated value format (CSV) format to be analysed externally using excel or any other statistical or programming packages.

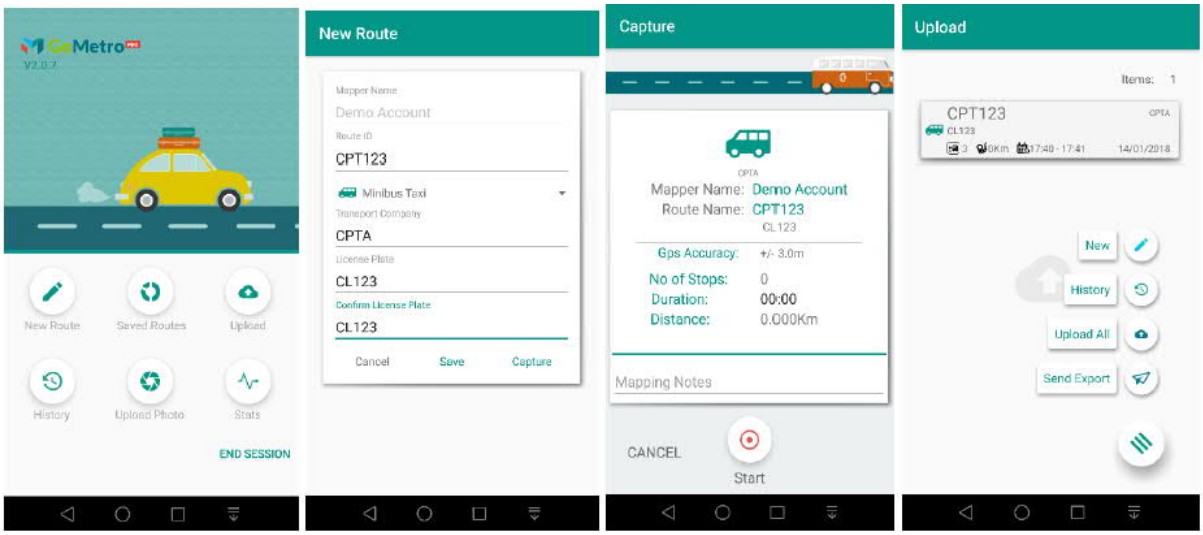


Figure 1: GoMetro Pro app route capture interface (Source: Coetzee, Krogscheepers & Spotten (2018))

3.2.3 Pilot Study / Proof of Concept

During the pilot study, the objective was to survey 10 different routes originating at three (3) specific ranks. The number of roundtrips and days on which these routes were to be surveyed was the same as the scope limitations mentioned above – three trips per route per commuter peak period, two per off-peak period, two weekdays, Friday and Saturday.

The purpose of the pilot study was to refine the data collection method, identify weaknesses in the tool, identify what additional training would be required to streamline the full study data collection effort and, most importantly, to engage with the MBT industry stakeholders (drivers, passengers, owners and operator associations) to facilitate the surveys.

The study was not without its 'hiccups' as some of the operator associations and owners were not happy for the survey to take place on their routes or vehicles. The reasons they gave included concerns that this study's data will inform the City of Cape Town in their decision to revoke operating licenses with the roll-out of the future phases of the MyCiti bus services. Drivers were often also wary of the survey, stating that they did not like to be 'spied on' by the City or the owners, for that matter.

The pilot study was scheduled to be completed in three weeks with the full study to commence once the results of the pilot have been approved by the client, the City of Cape Town.

With the variation in collection methodology to include the passenger OD information, the routes surveyed during the pilot were resurveyed in order to demonstrate the validity of the study approach and data collection tool.

Extensive planning and training of surveyors took place during the period in which the client reviewed and approved the pilot study data submission.

3.2.4 Stakeholder Engagement

In order to facilitate the surveys and get all players on board with the methods and purpose of the study, many different stakeholders needed to be engaged and negotiated with. Although

describing this part of the survey project is not within the scope of this research, it is crucially important to acknowledge the importance thereof.

Coetzee *et al.* (2018) provide more details on the consultations, engagements and negotiations specific to this project. Both Coetzee *et al.* (2018) and Gaibe & Vanderschuren (2010) acknowledge that a good communication plan and transparent stakeholder engagement is necessary to ensure the study can be carried out successfully – something the author of this research can confidently attest to, being personally involved with and present at many of the initial engagements with taxi associations, umbrella bodies as well as with representatives of the national regulation bodies.

3.2.5 Full Study

The full study commenced on 12 April 2017 and continued until March 2018. The data on which this research is based was collected between 12 April and 19 August 2017 the extent of which is summarised as follows:

- 12,781 taxi trips (includes the reverse/return trip);
- 167,798 passenger trips;
- 76,732 stops;
- 282 taxi routes surveyed (564 if including reverse route);
- 5,624 unique vehicles surveyed (high probability of duplicates with incorrect registration numbers);
- 59 days surveyed between 12 April and 19 August; and
- All days except Sundays and Mondays included in the survey.

For every trip surveyed, the following information was captured:

- Route designation ID
    - Associated Origin & Destination;
- Vehicle registration number;
- Taxi Association (information discarded as this was very poorly captured during data collection);
- GPS trace of the route the vehicle took between each stop.;
- Coordinates of each stop location;
- Timestamp at each stop location;
- Number of passengers boarding and alighting at each stop location;
- Fare paid per passenger at each stop location;
- Ethnicity of each passenger boarding (information unreliable as too much subjectivity used and correct use of this function was not widely practiced by surveyors);
- Approximate age group of each passenger boarding (three groups used: young, adult, old. This information is not reliable as the boundaries of age groups are subjective); and
- Gender of each passenger boarding (this information should be reliable as it is a binary choice for the surveyor with much lower level of subjectivity).

3.2.6 Data Structure

Table 1 provides a sample overview of the structure of the trip data.

Table 1 Raw trip data structure

Vehicle Reg No	Trip ID	Origin	Destination	Route ID	Surveyor	Company	Date Mapped	Distance	Start Time	Travel Time	Revenue
888SMO3WP	4033856	BELLVILLE	LANGA	COCTG62R	Tshediso Motjopi	CATA	26/04/2017	13.41	16:11:28	00:41	240
BWTA02WP	3728785	BRACKENFELL	BELLVILLE	COCT538	Anelisa Ntwanambi	CATA Bellvile	20/04/2017	6.86	11:15:56	00:18	205
BWTA05WP	3729255	BRACKENFELL	BELLVILLE	COCT538	Anelisa Ntwanambi	CATA Bellvile	20/04/2017	6.41	12:32:59	00:15	154
CA0000	5115716	MITCHELLS PLAIN	MORGENSTER	COCTMBT099	Lwando Ngonongono	BVTA	17/05/2017	4.84	15:36:37	00:10	136
CA102968	4321269	BELLVILLE	EERSTERIVIER	COCTF89	Nasiphi Magenuka	Eersteriver	4/05/2017	17.23	07:48:16	00:45	163
CA10583	5085113	MITCHELLS PLAIN	WESTGATE MALL	COCTMBT112	Wendy Mantshi	Johannes Meintjies	17/05/2017	9.44	11:22:01	00:28	135
CA106948	5106594	DELFT SOUTH	MITCHELLS PLAIN	COCTMBT107R	Anelisa Ntwanambi	Delft Taxi Association	17/05/2017	10.89	16:09:00	00:24	110
CA106948	5193740	DELFT SOUTH	MITCHELLS PLAIN	COCTMBT107R	Atabile Mzokoshe	DTA	17/05/2017	9.59	12:40:38	00:20	140
CA11049	5113210	MITCHELLS PLAIN	NYANGA	COCTMBT101	Ovayo Nyembezi	cata m/plain	17/05/2017	12.43	14:56:33	00:53	170

Table 2 provides an overview of the format of the passenger boarding and alighting information.

Table 2 Raw passenger boarding and fare data per trip

Vehicle Reg No	Trip ID	stop 1 on	stop 1 off	stop 1 fare	stop 2 on	stop 2 off	stop 2 fare	stop 3 on	stop 3 off	stop 3 fare	stop 4 on
CA815689	10000521	11	0	10	0	1	0	0	1	0	2
CA249116	10000732	8	0	10	0	1	0	0	7	0	
CT4891	10001022	15	0	10	0	6	10	0	1	10	0
CA292851	10001282	15	0	8	0	2	0	0	3	0	0
CA292851	10001833	4	0	9	2	0	9	1	0	9	0
CT4891	10002311	15	0	10	0	1	10	0	1	10	0
CA154809	10002683	15	0	9	0	1	0	0	1	0	0
CA298042	10002935	4	0	9	0	4	9				
CY239140	10003620	15	0	10	0	1	10	1	0	10	0
CA545069	10004067	11	0	10	1	0	10	1	0	10	0
CA154809	10004361	5	0	9	0	1	0	1	0	8	0
CA901069	10004629	15	0	9	0	3	9	0	4	9	0
CA58198	10005157	15	0	10	0	2	10	0	2	10	0
CF121635	10005653	10	0	10	2	0	10	1	0	10	0
CA753263	10006326	6	0	10	3	0	10	1	0	10	0
CA1877351	10006546	15	0	10	0	4	10	0	1	10	0

Table 3 provides an overview of the relative boarding and alighting locations captured for each passenger.

Table 3 Passenger boarding and alighting location details

Vehicle Reg No	Trip ID	Boarding Location	Boarding Time	Boarding Stop ID	Alighting Location	Alighting Time	Alighting Stop ID	Travel Time	Ethnicity	Age Group	Gender
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.663988, -33.890129	08:01:05	4018710	00:08:41	WHITE	MIDDLE	F
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.681101, -33.882927	08:06:44	4018716	00:14:20	BLACK	MIDDLE	M
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.687634, -33.878834	08:08:41	4018717	00:16:17	BLACK	MIDDLE	F
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.678301, -33.884567	08:05:38	4018715	00:13:14	INDIAN	MIDDLE	F
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.687634, -33.878834	08:08:41	4018717	00:16:17	INDIAN	MIDDLE	F
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.654633, -33.893124	07:59:52	4018709	00:07:28	BLACK	MIDDLE	M
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.687634, -33.878834	08:08:41	4018717	00:16:17	INDIAN	MIDDLE	F
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.687634, -33.878834	08:08:41	4018717	00:16:17	INDIAN	MIDDLE	F
CF232813	4018707	18.630964, -33.905220	07:52:24	4018708	18.671148, -33.887875	08:03:03	4018713	00:10:39	BLACK	YOUNG	M
CF232813	4018707	18.667952, -33.888832	08:01:49	4018712	18.678301, -33.884567	08:05:38	4018715	00:03:49	WHITE	MIDDLE	F

Table 1 provides the lowest level of detail per trip, showing information on the route ID, the unique trip ID, the date and time the route was surveyed, the name of the surveyor, the vehicle travel time, trip distance, total trip revenue and the start and end coordinates.

Table 2 shows the number passengers getting on and off at each stop along the route with the fare paid by passengers getting on at each stop.

Table 3 shows higher level details about the stops and passengers. Every record in this table relates to a specific passenger trip, showing the boarding and alighting times and locations.

Every stop in Table 3 gets provided with a unique stop ID, which is related to the unique trip ID. The unique trip ID is critical to the data in that it connects the data in the three tables with each other.

The relationship between the stop ID and trip ID is that the first stop's ID will always be equal to the trip ID + 1. In other words, if the trip ID is 4018707, the value of the first stop, the trip origin rank, would be 4018708 and every subsequent stop value would add 1 to the previous value. In Table 3, which shows some of the stops for trip 4018707 shows how the stop numbering system works. The trip and stop IDs are assigned to a data record once it is uploaded to the central database to ensure the values do not already exist in the database.

In addition to the data tables, each trip record is saved as two Keyhole Markup Language (KML) file extension file – one file for the route trace and one file for the stops.

The two files that would be produced by the data export for trip ID 4018708 would have the following file names:

COCT538_4018707_trip.kml; and

COCT538_4018707_stops.kml.

The filenames of these two files contain both the route ID as well as the unique trip ID. This is essential in connecting the information in the data tables with the geospatial files, the importance of which will become more apparent later in this document.

Figure 2 shows a KML file for a single trip between Mitchells Plain and Bellville with five stops, three stops en route and the two rank stops at the origin and destination. It also shows the timestamp of arrival at each stop and the number of passengers boarding and alighting.

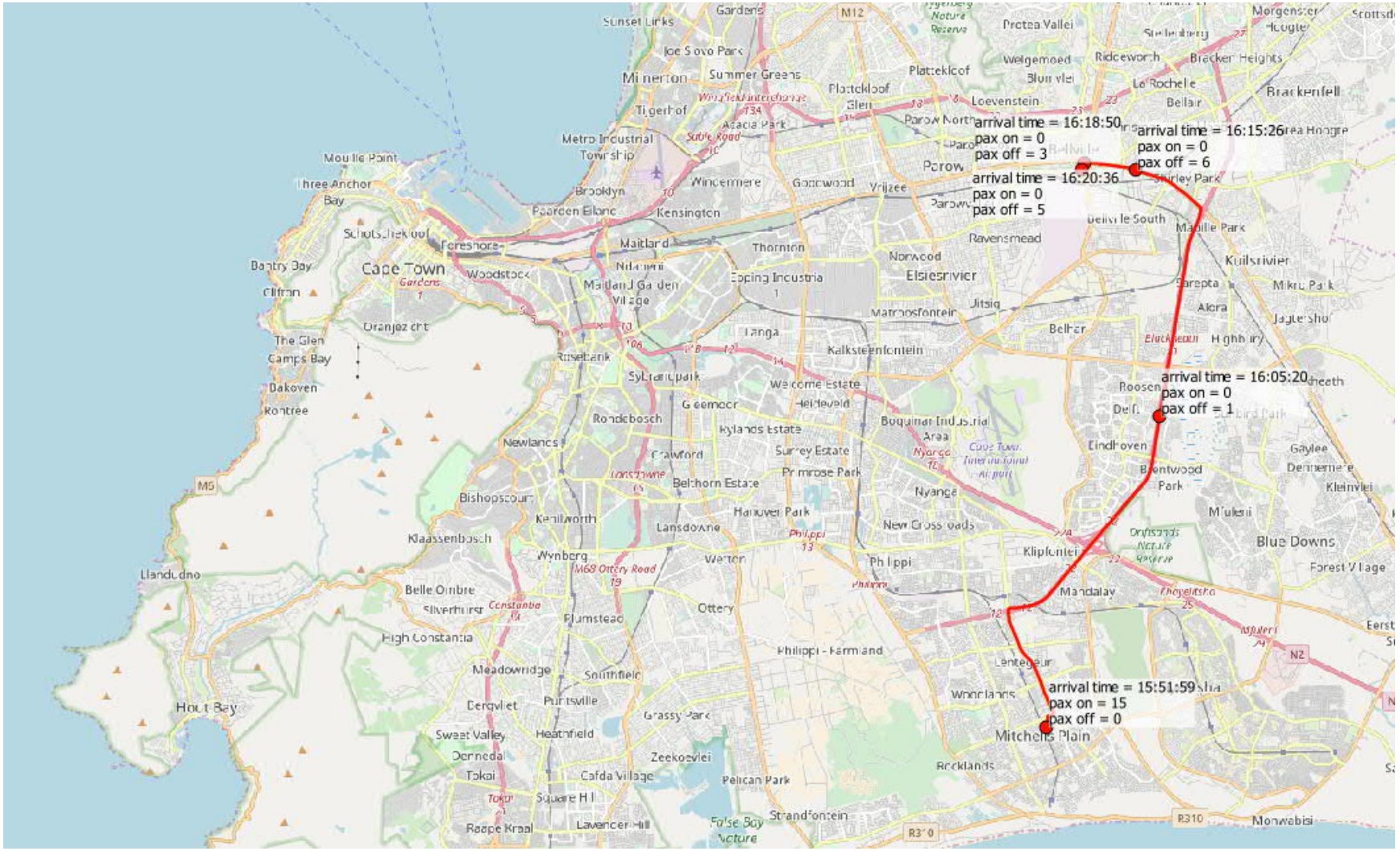


Figure 2: Example of trip geospatial file (Mitchells Plain - Bellville)

3.2.7 Extent of Data Coverage

Figure 3 shows the coverage of the routes surveyed in relation to the geographical area of the Cape Town region. Despite the fact that this dataset does not comprise the entire collection of the City of Cape Town's MBT routes, Figure 3 shows that the coverage of the MBT network is extensive, covering the majority of the most densely populated residential areas in the south east and connecting with the all commercial and industrial land uses in the central and western areas.

Figure 3: Coverage of routes surveyed

3.2.8 Operational Information for Analysis

Given the objectives of the research listed in section 1.4, especially the objectives concerned with developing a deeper understanding and classification system of MBT services, the information considered most important, in addition to the geospatial and temporal information collected, is the following:

- Trip route and distance;
- Operating speed;
- Vehicle travel times;
- Fare cost per trip;
- Trip revenue;
- Passenger turnover per trip; and
- Vehicle occupancies between stops.

Information determined from the data that is considered secondary to the research aims include:

- Passenger travel times;
- Passenger travel distances; and
- Passenger origin-destination information.

3.3 Data Cleansing

3.3.1 Types of errors

In order to make any meaningful inferences from data, it needs to be devoid of obvious errors and thoroughly 'cleaned' or 'cleansed'. This process is often the most time consuming and tedious part of data analysis (Kandel *et al.,* 2011).

Incorrect or inconsistent data can distort the results of models and analyses, nullifying the benefits of using a data-driven approach (Hellerstein, 2008).

Hellerstein (2008) presents different types of errors that can creep into data. The data error types that are applicable to this research include 'data entry errors', which is common when data collection is reliant on human data entry such as is the case in this research for some aspects of the data collection.

Another type of error that has been observed in this data for which no literature has been discovered is the incorrect formatting of timestamps with data uploads and downloads. For the purpose of this discussion, this will be called 'timestamp error'.

With the correct calculation of information such as travel time and speed being reliant on two correct timestamps, the error in one variable propagates to two others.

Incomplete records are also considered errors in this study. These errors occur when there is a break in GPS signal or some other equipment malfunction, which results in the trip data collected for said trip to be saved as an incomplete trip, for example a trip with only one stop (the origin), 12 passengers boarding and no passengers alighting.

3.3.2 Methods of cleansing the data

A method of investigating the underlying structure of the data to determine what the expected range of variables should be is called 'Exploratory Data Analysis' (EDA), a term coined by John Tukey in his 1977 book by the same title.

Tukey (1977) presents different EDA techniques: tables, five-number summaries, stem-and-leaf displays, scatterplot matrices, boxplots, residuals and outliers, amongst others. Many of these processes require human interaction to identify errors and to rectify them (Hellerstein, 2008).

The appropriate method or technique to cleanse the data with depends on the type of data – i.e. quantitative data, categorical data, identifiers, etc.

Quantitative data are numerical values or measures of some quantity and unit. Statistical methods of identifying outliers in this type of data are the foundation of the available cleaning techniques.

Categorical data represents a finite set of names, numbers or codes that have been assigned to the data to categorise them. An example of this is the predefined route IDs have been given to

the MBT routes prior to surveying them. Cleaning techniques for this type of error are usually more labour intensive than for quantitative data.

Identifiers are unique or arbitrary key values that are assigned to individual entries or subsets in a dataset and often have no meaning other than the identification property imbued upon them. The unique trip and stop IDs in the MBT data in this research are examples of such data. Since these values have been automatically generated upon upload to the central server, the only check that was done to these variables was to determine if duplicates existed. N duplicates values were observed.

Sources of 'dirty' data in qualitative data can be quite easily discovered by evaluating one column at a time for outliers (Hellerstein, 2008). This process is called univariate outlier analysis. In order to define what an outlier is, one need to determine what is considered unacceptably 'far' from the mean value. The most common metric by which this distance from the mean is described is standard deviation or variance (standard deviation squared. One also, however, needs to know if the data is normally distributed about the mean or if it is skewed in a specific direction from the mean. This skewness can be expected in some of the variables in the data analysed in this research as only non-zero values exist for concepts such as speed and distance.

Hellerstein describes Robust Statistics (Huber, 1972) as a subfield of statistics that describes the effect of errors in data variables on the distributions and develops methods or estimators that can eliminate such errors. One of the robust metrics methods of determining the centre of a dataset is using the median instead of the mean. Vastly different values can be determined using these two metrics.

The mean and median of an array {1, 3, 4, 7, 12, 5, 500} are 76 and 5, respectively. The mean is clearly in this case not representative of the majority of the values and 500 is clearly an outlier and should be rectified somehow.

A popular robust centre method, especially appropriate for skewed datasets with long tails, is the trimmed mean. This method refers to the recalculation of the mean after *k%* is removed on either end (low and high) of the quantitative data. A variation of this method is called winsorized mean. This refers to the method of replacing the trimmed *k%* with lowest and highest values that were not excluded (Hellerstein, 2008).

Where pure logic could not be used to eradicate erroneous data, the trimmed mean method was used. The winsorized mean method was not used as the MBT data is by nature fairly unpredictable and discarding *k%* of the data on either end still left a large quantity of data to analyse.

3.3.3 Cleansing the data

In cleansing the MBT data, categorical errors were first identified. The only categorical data of interest that was left to human entry was the route ID code. As a master list of the routes surveyed was kept, the captured codes were cross referenced with the master list to identify any codes that were captured that do not appear in the master list. The master list also had the dates on which specific routes were surveyed to compare the captured date with. The majority of route IDs could be rectified in this manner. When they could not, the scrutinised

route's KML file was plotted against other routes captured that day to determine which route most closely followed the same route alignment.

The rest of the errors sought were all quantitative. The fields that were scrutinised included the following:

- Number of passengers;
- Number of stops;
- Revenue;
- Distance;
- Speed; and
- Travel time.

The data cleaning and analysis was carried out using R, the software programming environment for statistical computing and graphics[6],

The raw data was imported into R where it was cleaned using a set of heuristics compiled in a script written for the purpose before it was analysed, also in R.

The script (Appendix A) was developed and applied to the data in order to obtain a dataset that can be used to analyse, explore and visualise the different facets of the data.

The script first removes obvious errors in the data such as trips where there were only one or zero passengers, trips with only one stop which is impossible since the ranks at either end are both counted as stops, trips with distance less than 1km as the shortest route surveyed had a distance of 1100m (Mitchells Plain Promenade – Mitchells Plain Town Centre rank). Trips with zero fare revenue were also considered incorrect and removed from the dataset – investigating individual entries for this occurrence was not feasible.

Since travel time and speed were calculated using the difference in trip start time and end time timestamps and the trip distance, if one of these timestamps was incorrect, the trip time could be negative, very large or very small. By eliminating these occurrences, impossibly achieved speeds and travel times were removed. To do this, all speed was first set to the distance divided by the trip time. Thereafter, trimming procedures were applied to remove the incorrect values on either side of the data spectrum.

The trimmed mean approach, as described above, of removing 2.5% on the low and high end of the data was applied to speed and travel time to address the aforementioned issue.

Figure 4 and Figure 5 indicate the operating speeds determined from the raw and cleaned datasets respectively.

[6] https://www.r-project.org/

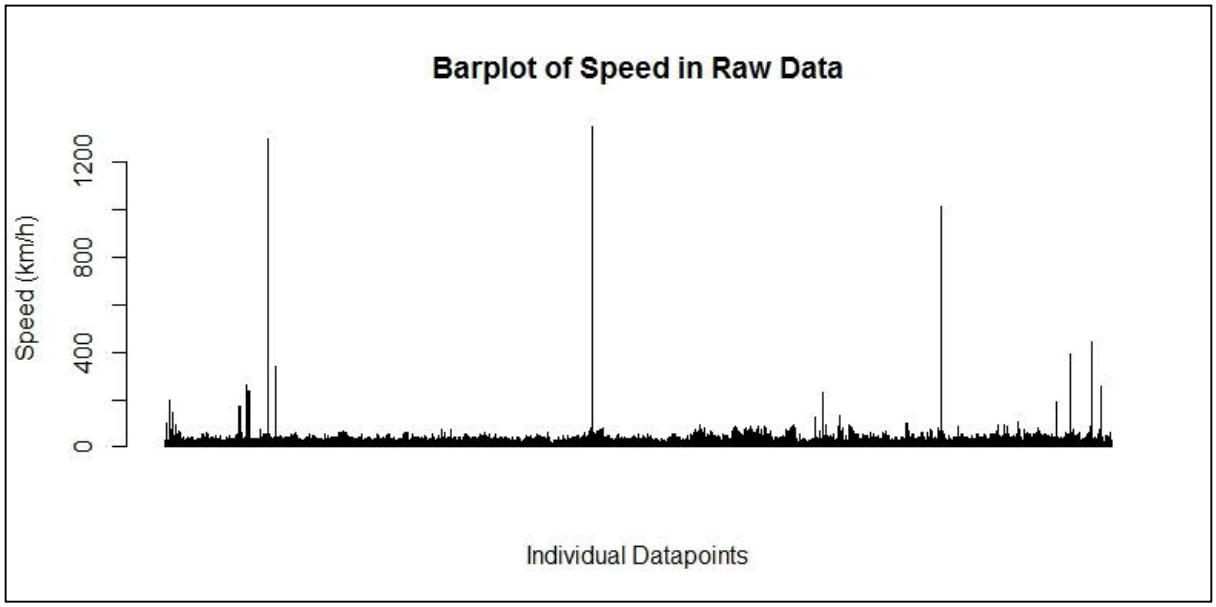


Figure 4: Operating speeds in raw data

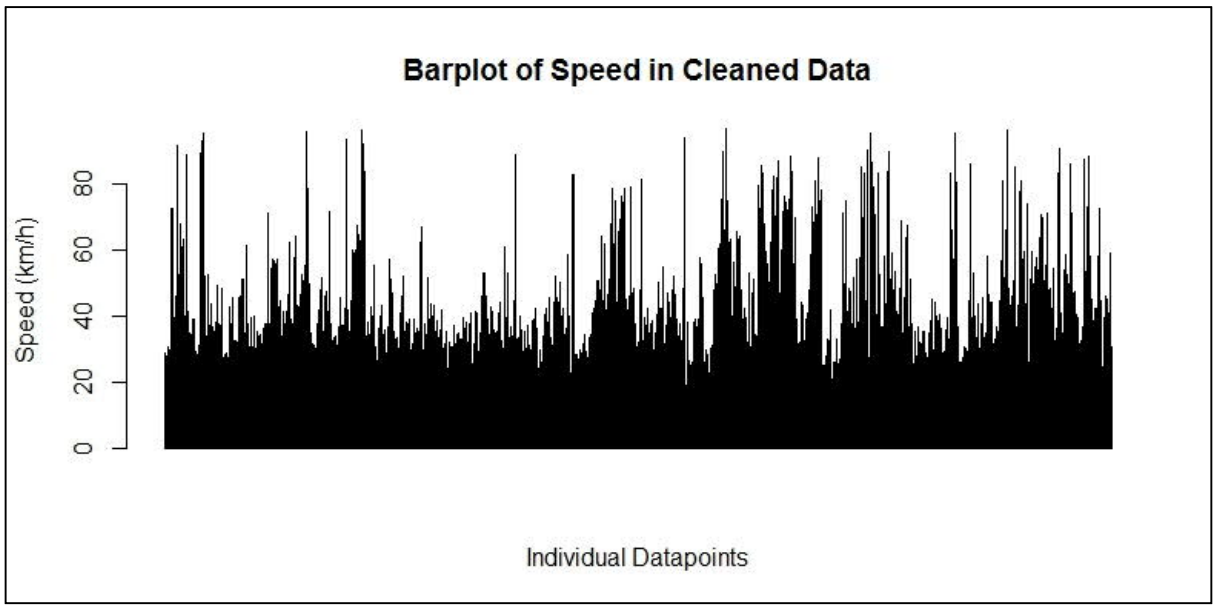


Figure 5: Operating speeds in cleaned data

While there are still a number of unlikely and relatively high speeds found in this dataset, the impossibly high values were removed from the dataset. The aim of this research is to propose a framework for using data of this type to understand the MBT operations and not the data analysis itself. The 2.5% trimmed off based on travel time and speed was to some extent a subjective exercise. This represented the smallest percentage that did not yield impossible speeds and travel times, i.e. 400km/h trips or 6-hour long trips. While the impossible records have been eradicated, a small proportion of improbable records have been retained.

3.4 Sampling Validation

As documented in the scope and limitations of the research, the sampling of the data collection project from which these data were obtained was not designed to be statistically representative; the margins of error found in these data are, therefore, unknown and cannot be

measured with conventional methods. Trip and route distances, which is discussed in the following sub-section, the only static characteristic of a trip that can be compared to something physical and tangible in the real world, i.e. the known distance of official routes, was used as a method of "validating" the data sampling.

The City of Cape Town's Transport Reporting System (TRS) was consulted to determine the length of all active designated MBT routes. Figure 6 shows the distribution of route distances as extracted from the City of Cape Town's TRS compared to the distances of the routes surveyed. It is important to note, however, that some routes are considered active as their operating licenses remain in effect but in reality, these routes are not operated. In other cases, licensed routes that are considered dormant are in reality operated. For these reasons, the survey schedule could not merely be a case of "ticking off" routes surveyed from the TRS list of active routes. On the ground reconnaissance needed to be carried for every rank from which surveys were conducted.

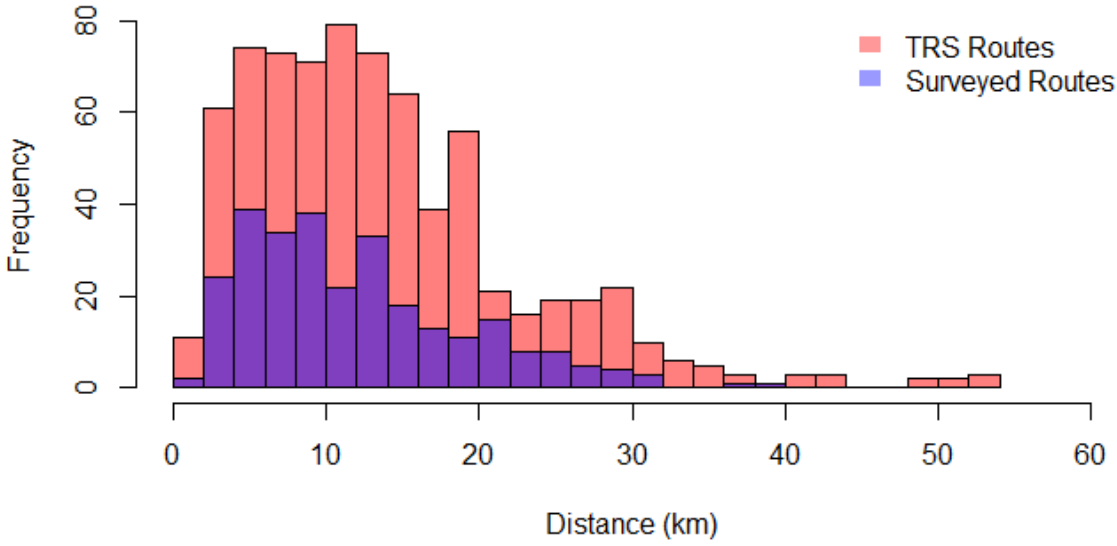


Figure 6: Histograms of distance of official TRS and surveyed routes (Source: TRS)

The margin of error (e) for the sampling rate shown in Figure 6, when considering each 2km distance band surveyed separately, is 19% with a 95% confidence level. The margin of error was calculated for each distance interval as follows:

$$e = \sqrt{\frac{Z^2 \cdot p \cdot (1-p)\left(1-\frac{SS}{N}\right)}{SS}}$$

Where:

e = margin of error;

Z = 1.96 - standard score for 95% confidence level;

p = 0.5 – worst case estimated proportion of the population that have the same attributes;

SS = sample size of route distance bands from survey; and

N = population size of route distance bands from TRS.

The comparison of the distances of the sample of routes surveyed with the entire route population suggests that routes surveyed are a fair representation of all routes within these distance intervals in the City of Cape Town.

Given that the sample of routes surveyed represent the City of Cape Town's active MBT routes well in terms of route distance, a key route service attribute, it can be concluded that despite the limitations surrounding the scope of the data collection study from which these data were sourced, and its lack of sampling design, that there are no major flags of concern. There is, therefore merit in the analysis and synthesis of results from these data.

Apart from the sample validation, the following can be inferred from comparing the surveyed route distance distributions with the route distances from the TRS:

- The distribution of route distances surveyed relatively closely resembles the TRS route distance distribution, providing a level of validation to the survey sample;
- The median distance of the surveyed routes 10.6 km;
- The median distance of the TRS routes is 12 km;
- Many longer distance routes may not have been surveyed; and
- The survey data contains data on a relatively higher proportion of shorter distance trips than found in the TRS route distances.

3.5 Analysing the Data

An exploratory approach was taken to determine the underlying structure of the data, in other words, to determine the following:

- The range of values observed for each operational attribute;
- The relationships that exist between these attributes;
- The operational attributes that contribute most to the variance in the data; and
- Methods that may be used to classify the data and the routes according to these attributes.

3.5.1 Visualising the underlying structure

The main goal of data visualisation is to effectively and efficiently communicate key information (Friedman, 2008). Different plot types were used in order to visualise the range, scale, and relationships in the different variables in the data.

The underlying data structure was first explored by means of visualisation to determine how classification could be achieved.

Bar plots of every observation of a given variable, as shown in Figure 4 and Figure 5, were used to visualise the spread of values for said variable. This was effective in the data cleansing exercise as well as in the general understanding of the range in which mean values could be expected to occur.

Boxplots were used to visualise the distribution of a given variable based on the following:

- Minimum;
- Maximum;

- Median;
- First quartile; and
- Third quartile

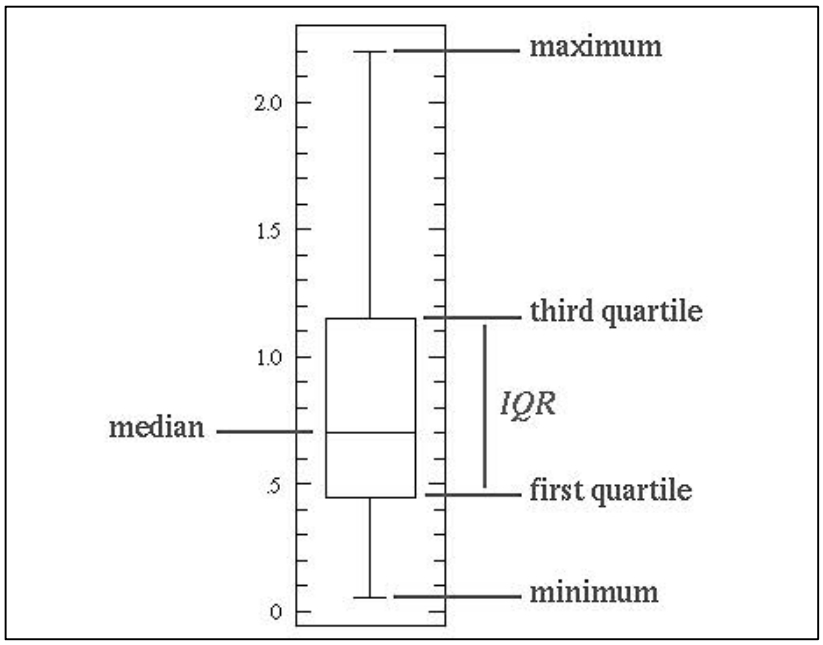


Figure 7: Boxplot definition (source: http://www.physics.csbsju.edu)

Boxplots provide a way of displaying the full range from minimum to maximum and the position of the first and third quartiles, indicating how far or close the majority of the data point values are from the median. The interquartile range (IQR) represents the range within which values are most likely to occur.

Simple bar style histograms were used to visualise the relative frequency of values for respective variables. A histogram is representation of the estimate of the probability density function of a quantitative continuous variable. It counts how many times a variable falls within a specific interval (or "bin") and displays a rectangle of each interval of which the height indicates the relative frequency of occurrence. Figure 8 is an example of a histogram – in this case showing the same data shown in Figure 5.

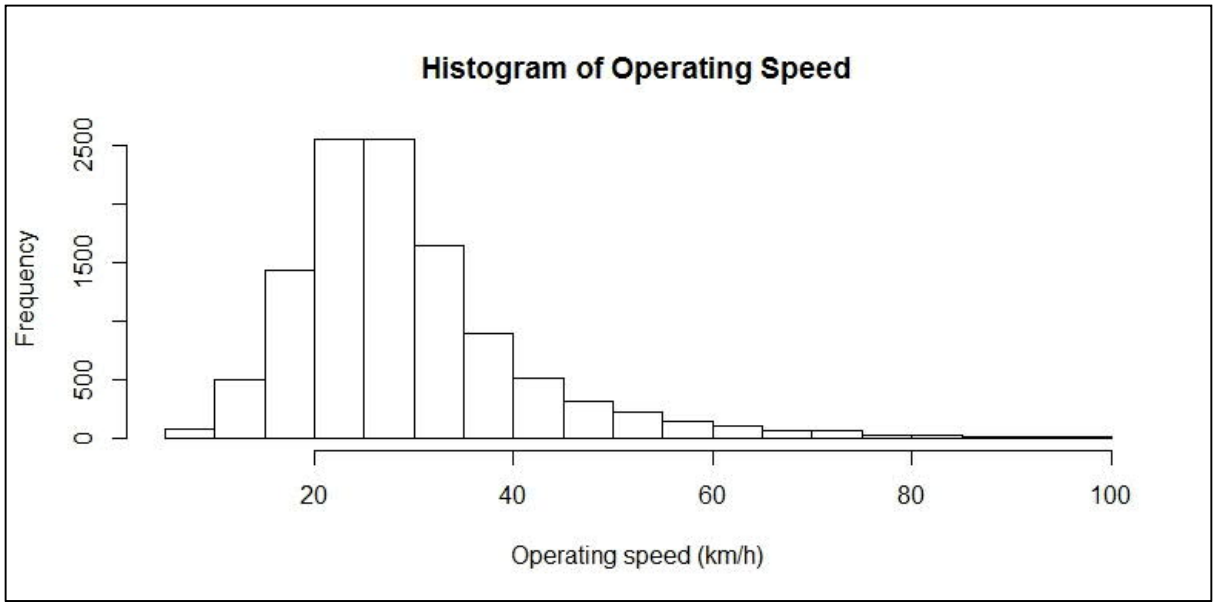


Figure 8: Example histogram

Scatter plots were used to visualise relationships, or the lack thereof, between certain variables. A two-dimensional (2D) scatterplot shows two variables plotted as Cartesian coordinates on a set of axes (X, Y). A 3D scatterplot adds another dimension (Z) to the plot on which a third variable's position can be plotted.

3.5.2 Variables Analysed

The following variables or metrics were analysed in order to get an understanding of the operations of MBT services:

- Distance;
- Speed;
- Travel time;
- Number of stops made;
- Number of passengers transported;
- Fare; and
- Revenue.

CHAPTER 4

4 MBT Operations

This chapter explores the operational characteristics data variables. It provides visual plots of the distributions of values for each variable investigated. It also provides discussions on the inferences made from the distribution and range of each one of these variables with respect to the expected and observed operations of MBTs in Cape Town. Lastly, this chapter identifies which variables could be instrumental in defining route type classifications. Understanding the service types operated by routes provides planning and operating authorities with information to use in improving MBT services and how best to integrate operations with formal or scheduled modes.

4.1 Trip and Route Distance

Trip distances are calculated from the route trace by adding the Euclidean distance between successive GPS coordinates. GPS coordinates were collected at a very high frequency (multiple points per second). The frequency of points does, to some extent, depend on the weather – which also influences the precision of the data. In poor weather conditions, these points may be up to 5m off. In general, however, the precision, accuracy and frequency of data points were of high quality.

Figure 9 shows the distribution in trip distances from the onboard survey data.

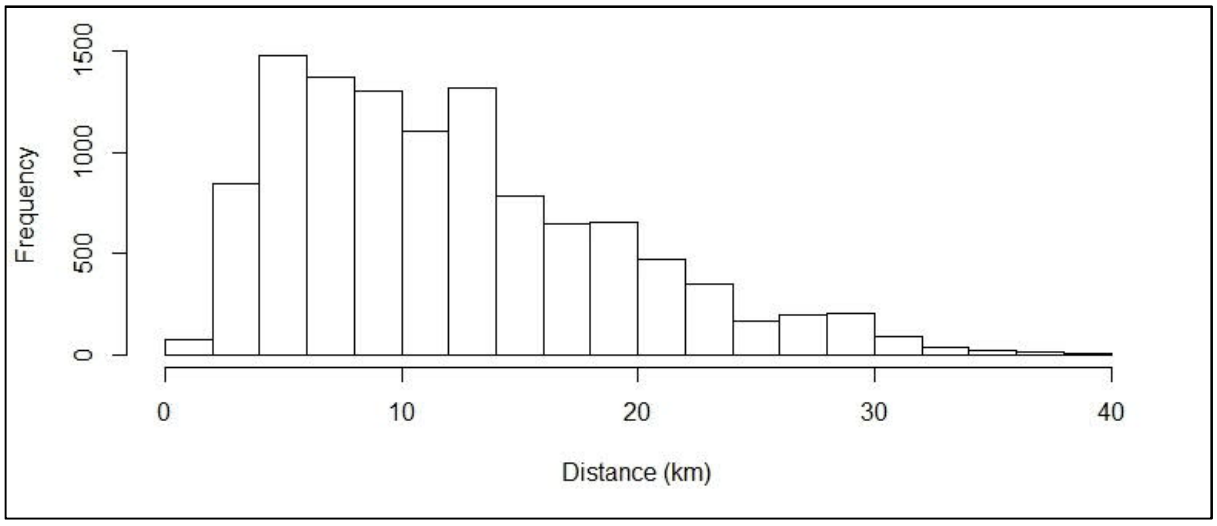


Figure 9: Histogram of distances of surveyed trips

Since the data analysed does not contain an equal number of trips for every route surveyed, the histogram above is skewed towards the distances of routes for which more entries in the dataset exist. Figure 10 shows a histogram of the route distances surveyed to account for the trip survey frequency bias.

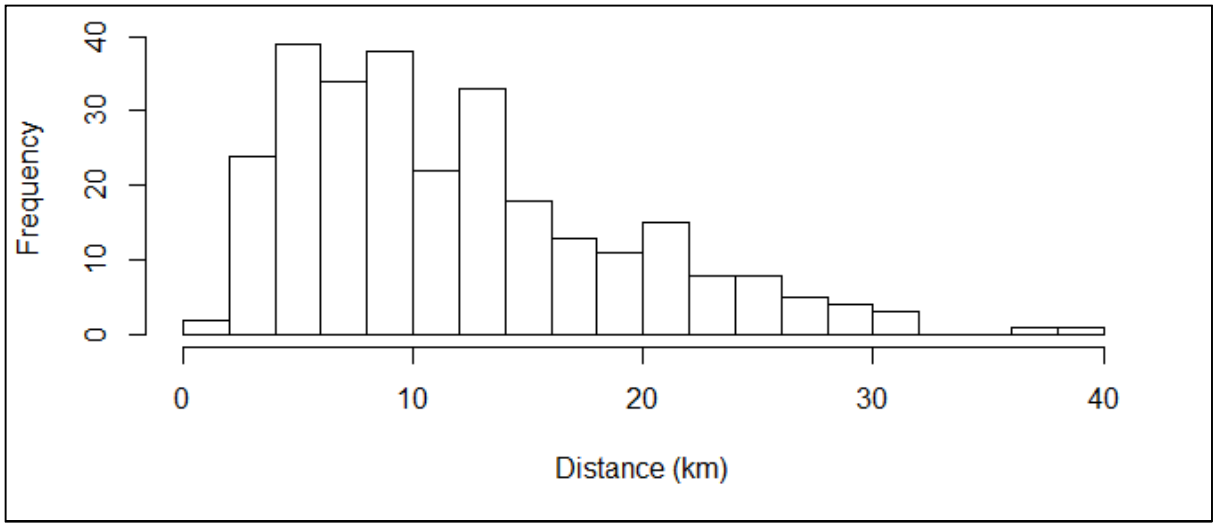


Figure 10: Histogram of distances of surveyed routes (including reverse routes)

It is important to note from these two histograms that this differentiation between trips and routes is only really meaningful for distance as this the only variable captured that has a (theoretically) static value. Since the vehicles, however, often do not adhere to their designated routes, there is a measure of variability in the observed trip distances.

Figure 11 provides an example of how the distance measured for a specific route may vary.

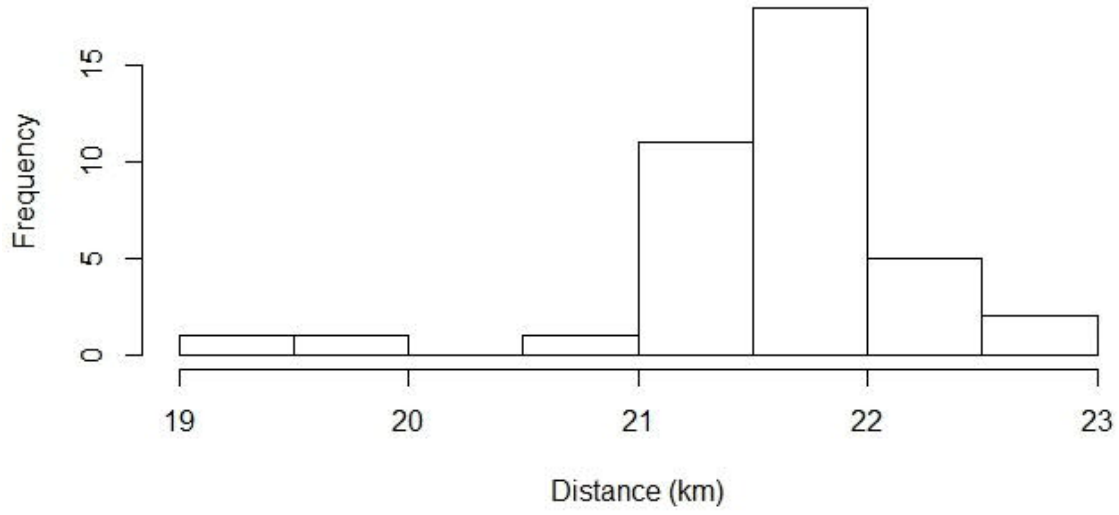


Figure 11: Example of variation of trip distance for a single route

This variation can be attributed mostly to the variation in the actual route the vehicle took between its origin and destination. Some measure of variability may also be attributed, in some cases, to GPS imprecision and inaccuracies due to weather and/or infrastructure interferences and equipment malfunction but this only constitutes a very small proportion of the variability.

4.2 Operating Speed

Figure 12 shows the distribution of operating speeds across all trips included in the cleaned dataset analysed.

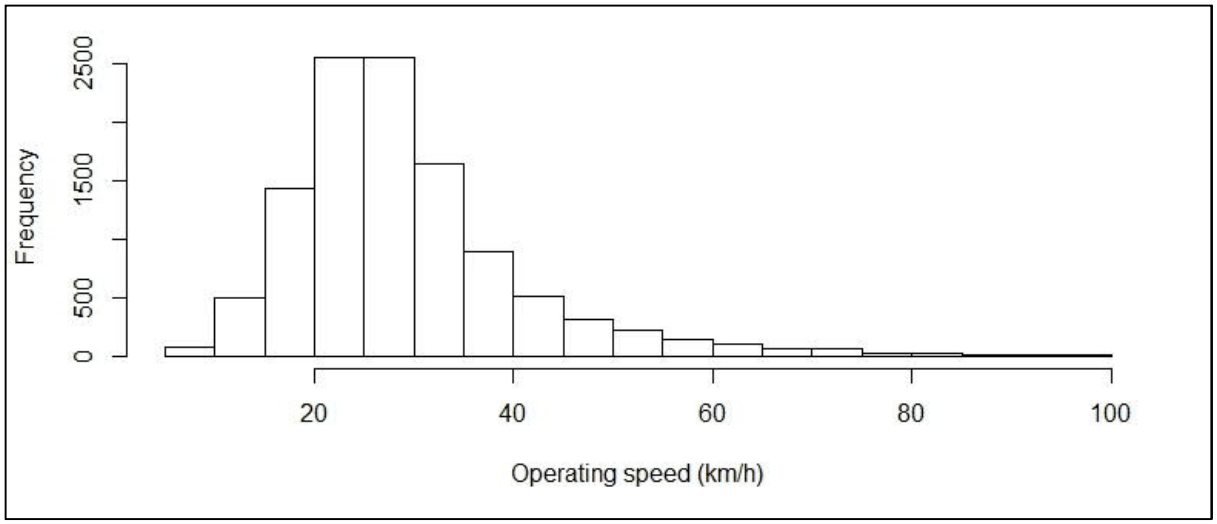


Figure 12: Histogram of trip operating speeds

The distribution of operating speeds seen in Figure 12 is also subject to the trip frequency bias discussed for the distances above. A higher number of trips exist for some routes and, therefore, distribution of speeds observed will be skewed by these results. Calculating the mean or even the median speeds of all routes and visualising that distribution will also not provide a good indication as to what the typical operating speeds are for a route.

Operating speeds typically vary according to the time of day as a result of passenger demand patterns, so the data would have to be split into time periods to observe the operating speeds for different times of the day.

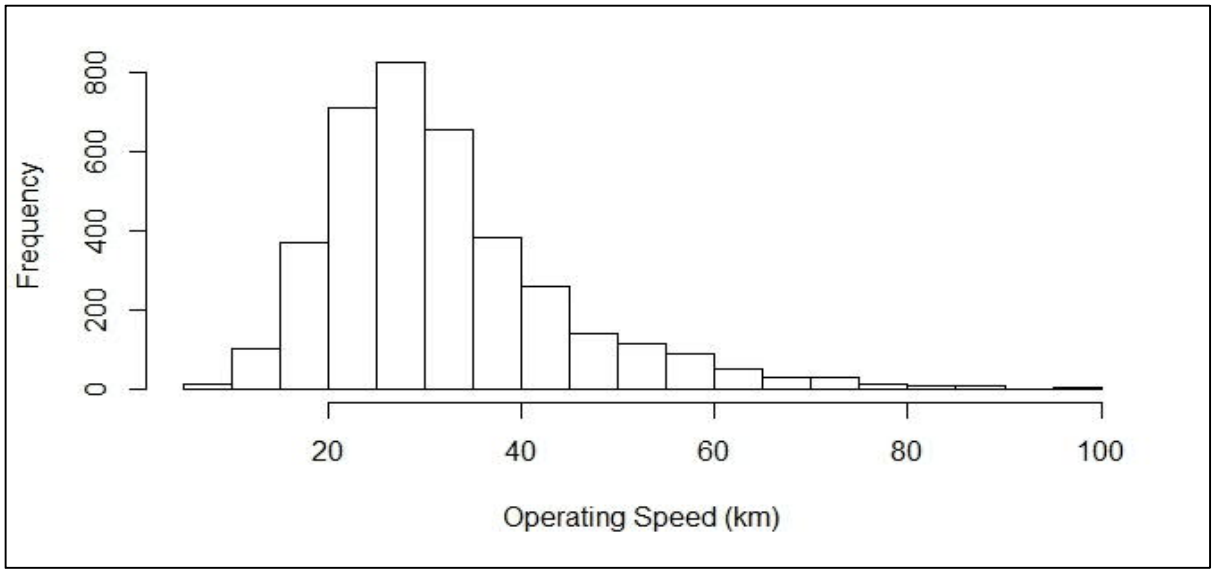


Figure 13: Operating speeds of trips surveyed during the morning (6:00 - 10:00)

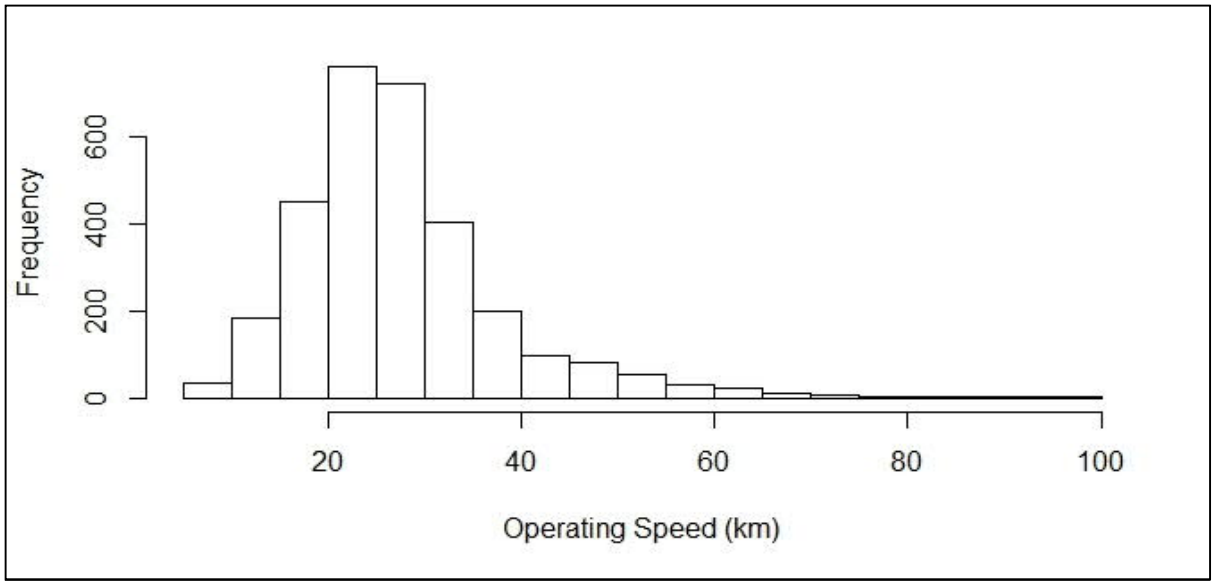


Figure 14: Operating speeds of trips surveyed during the inter-peak period (10:00 - 14:00)

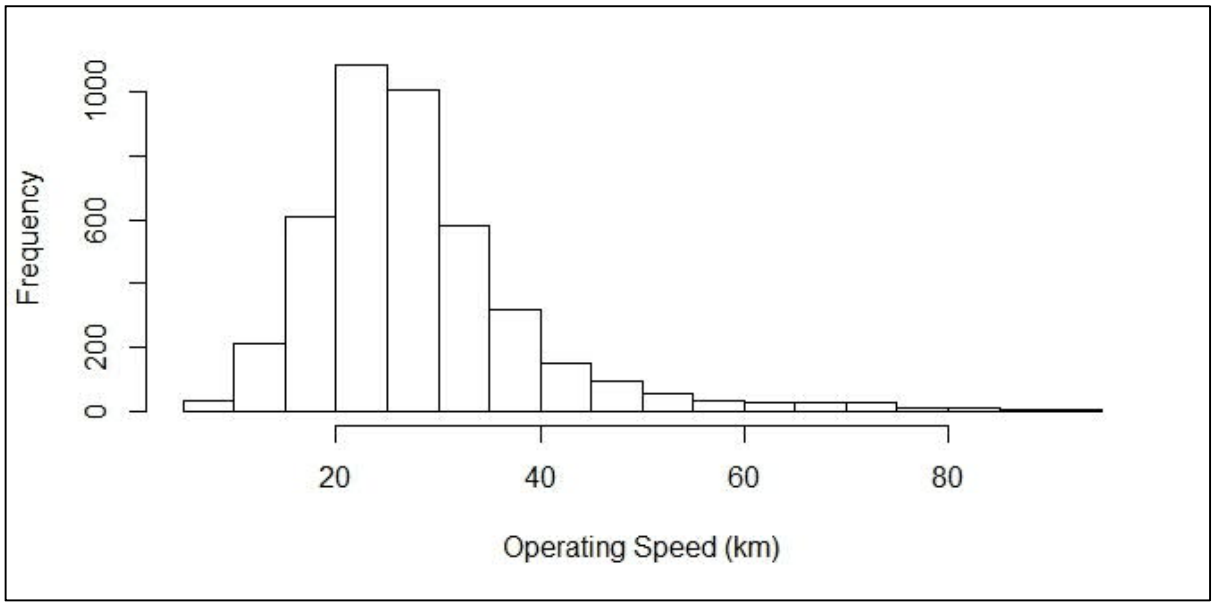


Figure 15: Operating speeds of trips surveyed during the afternoon (14:00 – 19:00)

Figure 13, Figure 14 and Figure 15 show, for the trips that were surveyed, that morning period trips' operating speeds (median speed = 29.1km/h) were slightly faster than the other two periods (median speeds 25.6km/h and 25.8km/h, respectively).

This is not, however, conclusive in stating that, on average, MBT operating speeds are faster during the morning than during the rest of the day. This statement may be true for some routes while the converse may apply to other routes. Each route operates differently and, by the very nature of these types of services, adapts to the circumstances, i.e. the passenger demands, congestion and special events. In the event of bus driver or train driver strikes, MBT is the only viable alternative for many captive public transport users. In these situations, a specific route may operate differently from its normal operations in order to capitalise on the profit to be made in transporting more passengers.

Under “normal” operations, however, some routes are expected to be faster in one direction than the other during peak commuter travel due to congestion and demand. In the counter peak direction, i.e. when a MBT returns to the rank (the reverse trip) to pick up another load of passengers, the vehicle would be traveling faster than for the forward trip it would most likely make fewer stops and congestion is often a unidirectional phenomenon. This applies mostly to trunk or line-haul type operations.

An example a route, with individual trip IDs shown in the vertical bars, exhibiting higher operating speeds in the counter peak direction is shown in Figure 16 and Figure 17.

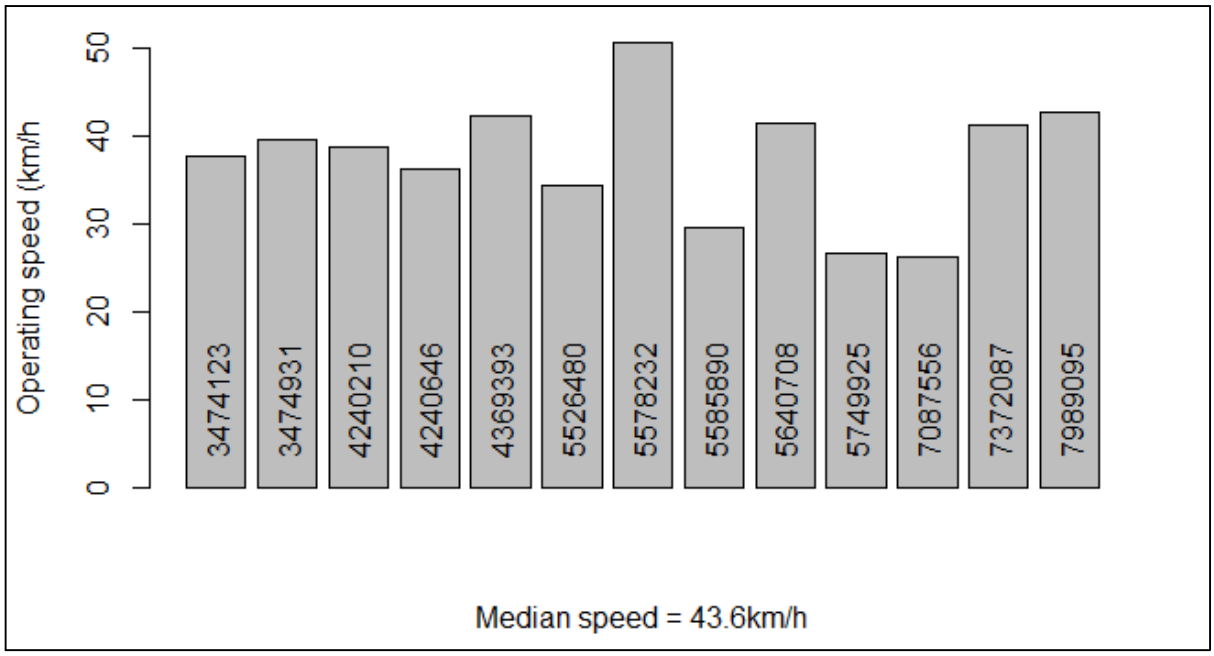


Figure 16: Morning peak operating speeds (Mitchells Plain to Bellville)

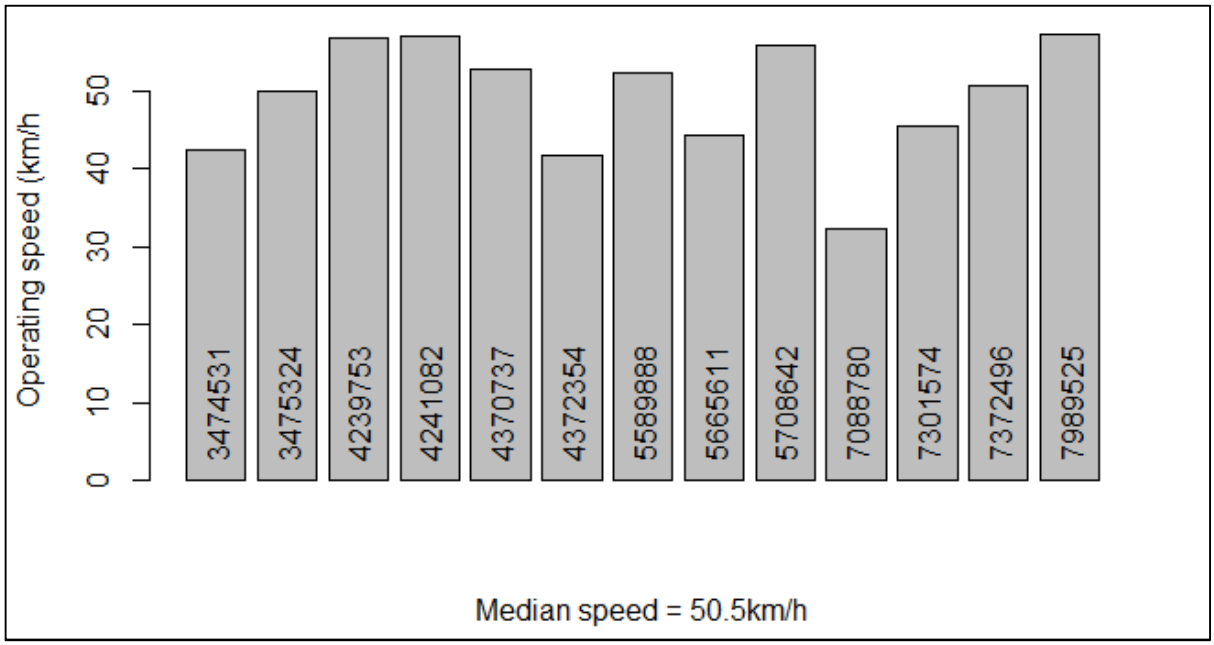


Figure 17: Morning peak operating speeds (Bellville - Mitchells Plain)

For the above route, the reverse direction speed is approximately 16% faster. Evaluating all routes surveyed, each route was on average 27% (median 20% and coefficient of variation 0.21) faster in a specific direction. This statistic was approximated by calculating the ratio of the maximum to the minimum of the average morning and afternoon speeds for each route.

4.3 Travel Time

Travel time was captured directly by the data collection tool as opposed to the speed, which was calculated from the route distance and travel time. The total trip travel time is the sum of travel times between stops where the timestamp and coordinates of each stop are recorded.

Figure 18 gives an overview of the distribution of vehicle trip travel times observed.

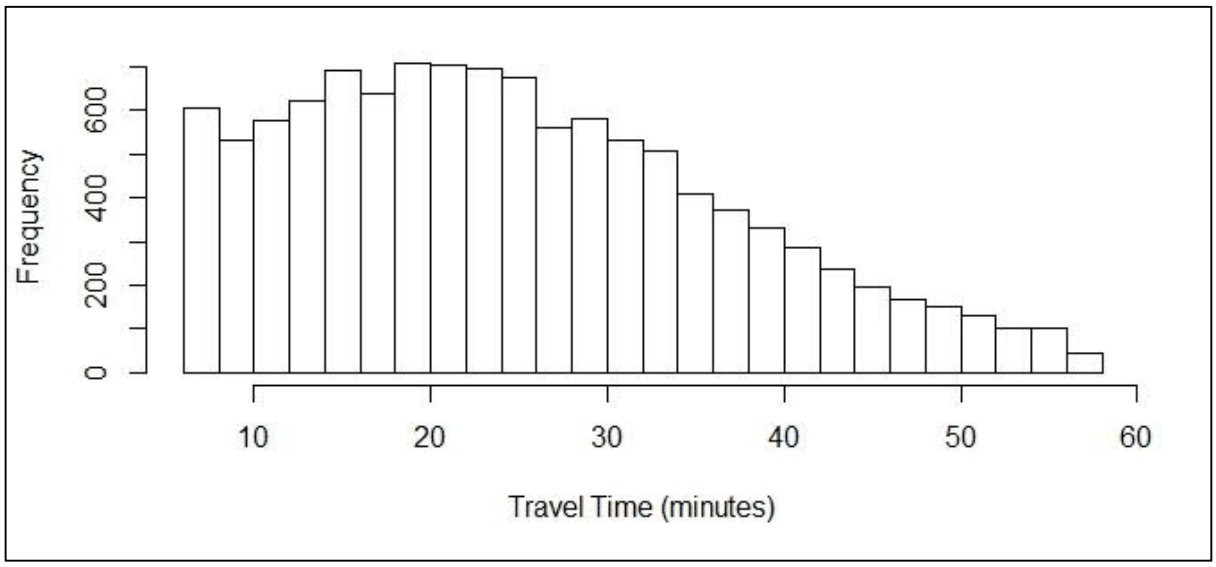


Figure 18: Histogram of vehicle travel times

The shortest trip travel time observed was approximately six minutes and was from Westridge to Hazeldene (areas in Mitchells Plain) with an approximate distance of 1.64km. Figure 18 shows that the bulk of the trips surveyed are between 15km and 25km distance.

4.4 Number of Stops

The hypothesis is that MBT routes exhibit some form of organic network structure where certain routes operate as traditional trunk type services while others operate as feeder or distribution type services. The nature of operation of a given route may, however, vary according to the time of day and direction of travel. As one of the outcomes of this research is to determine whether these services can be classified in terms or their service type, number or frequency of stops is an important descriptive indicator.

A histogram showing the distribution of the number of stops per trip (Figure 19) is shown purely to provide an indication of the range of values observed.

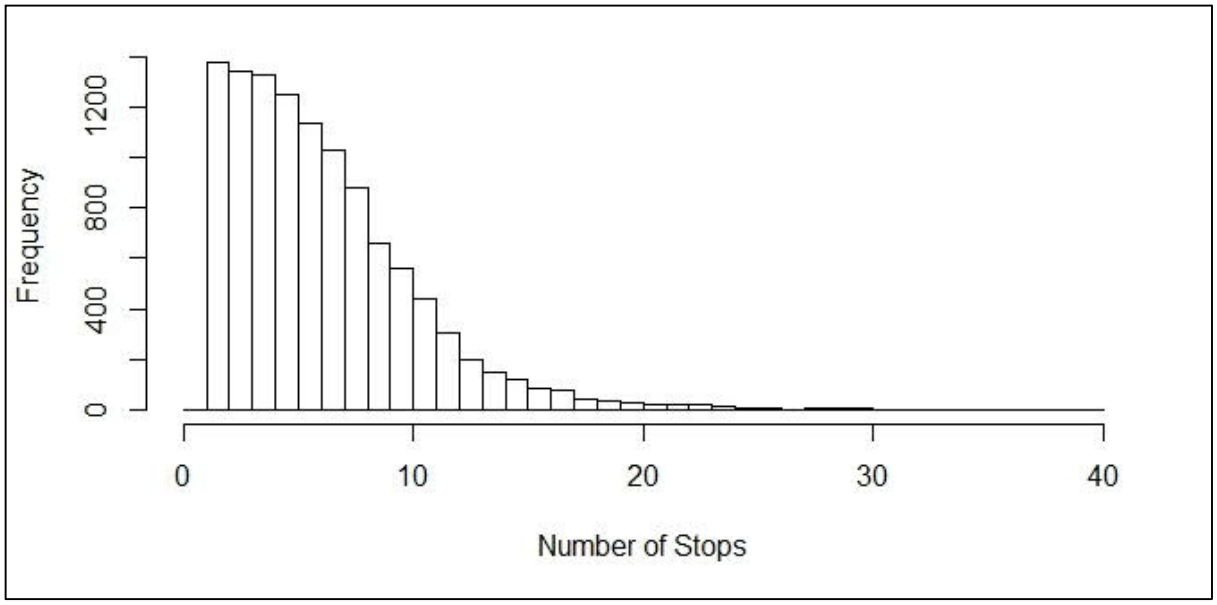


Figure 19: Histogram of stop frequency

Figure 19 shows that the majority of the trips surveyed only had two stops – i.e. the ranks on either end of the trip.

4.5 Number of Passengers

The number of passengers transported per trip by MBTs varies, as one would expect, according to travel demand which in turn varies, inter alia, according to factors such as the weather, special events, the land use of rank areas as well as the areas the routes traverse and the availability of other modes of public transport.

Figure 20 gives a high-level overview of the distribution of passengers transported per trip.

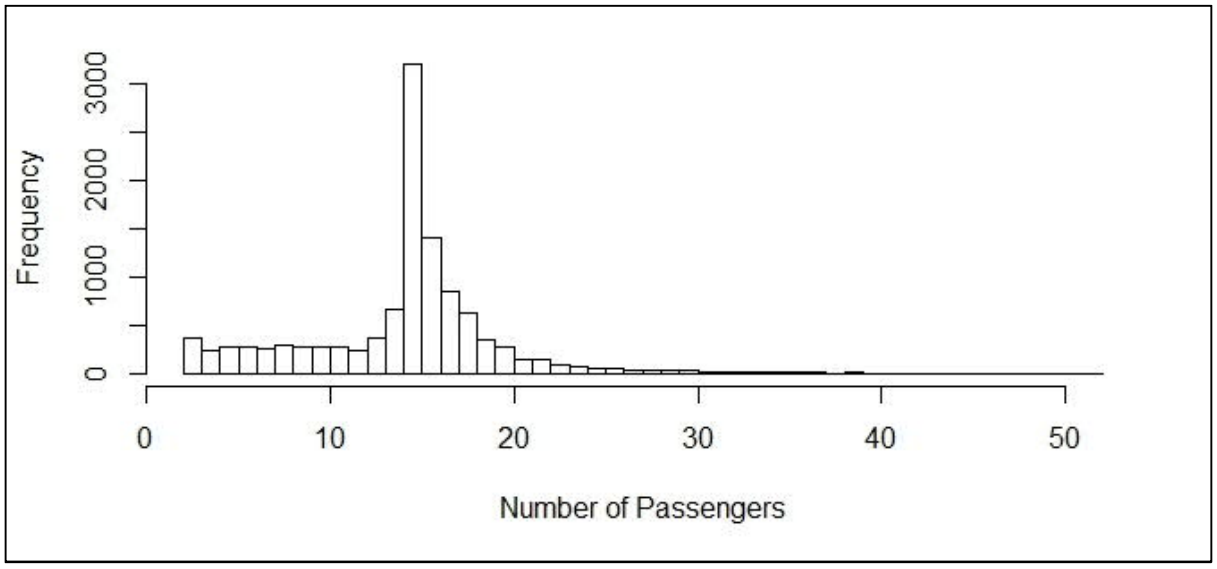


Figure 20: Histogram of number of passengers per trip

Figure 20 shows that the vast majority of trips surveyed had transported 15 passengers in total. It is also shown that for the routes and number of trips per route surveyed, a relatively

large proportion of trips only transport between 1 and 10 passengers. The maximum value found was 52 – a total of 11 trips surveyed transported more than 40 passengers.

4.6 Trip Segment Load

Segment load analyses has revealed the average and maximum segment loads per trip (the passenger volume on board between stops) are distributed as shown in Figure 21 and Figure 22 respectively.

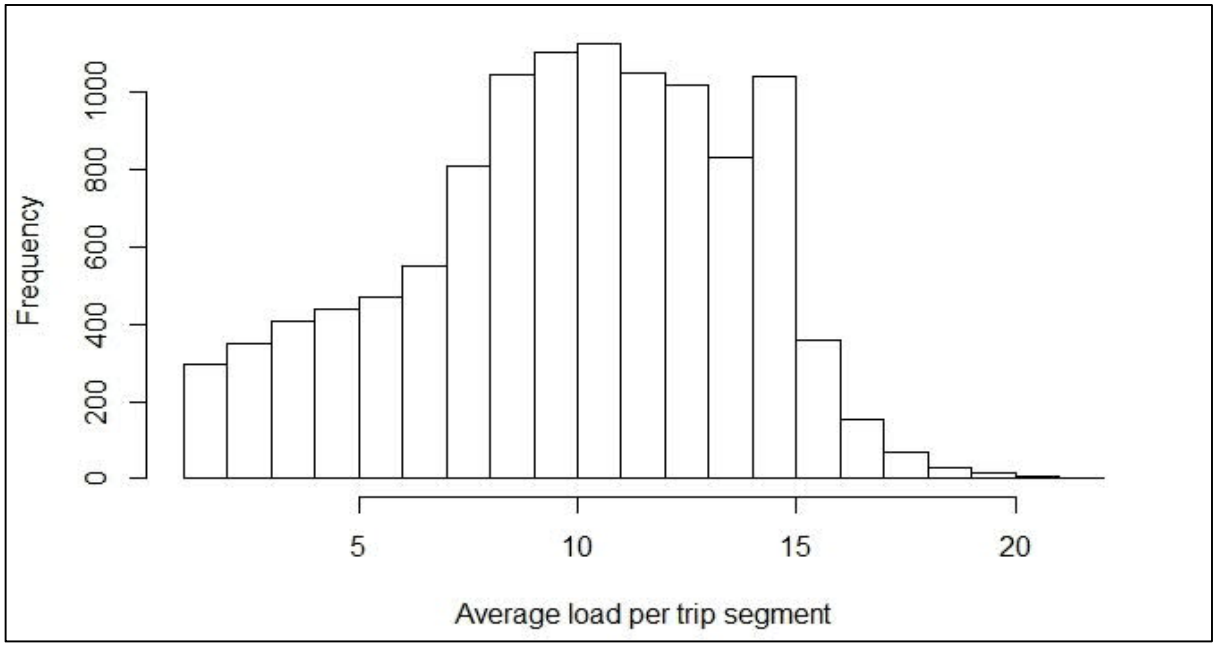


Figure 21: Average trip segment passenger load

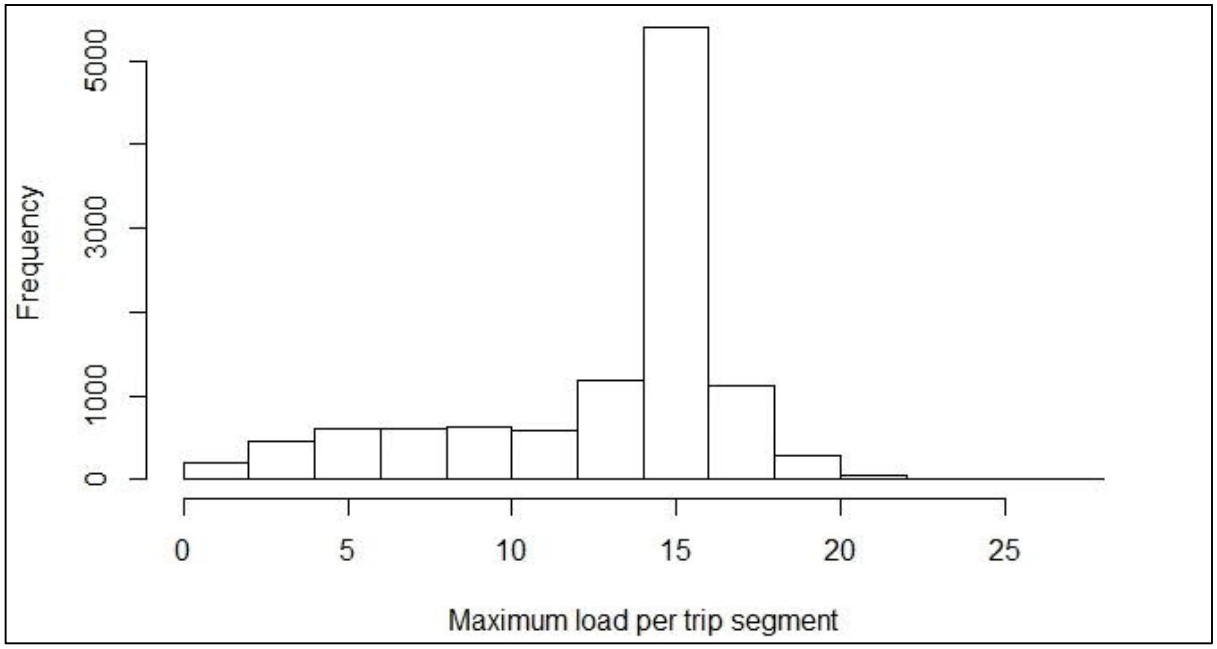


Figure 22: Maximum trip segment passenger load

Figure 23 shows the distribution of passenger numbers boarding at the 1st stop (the trips origin rank).

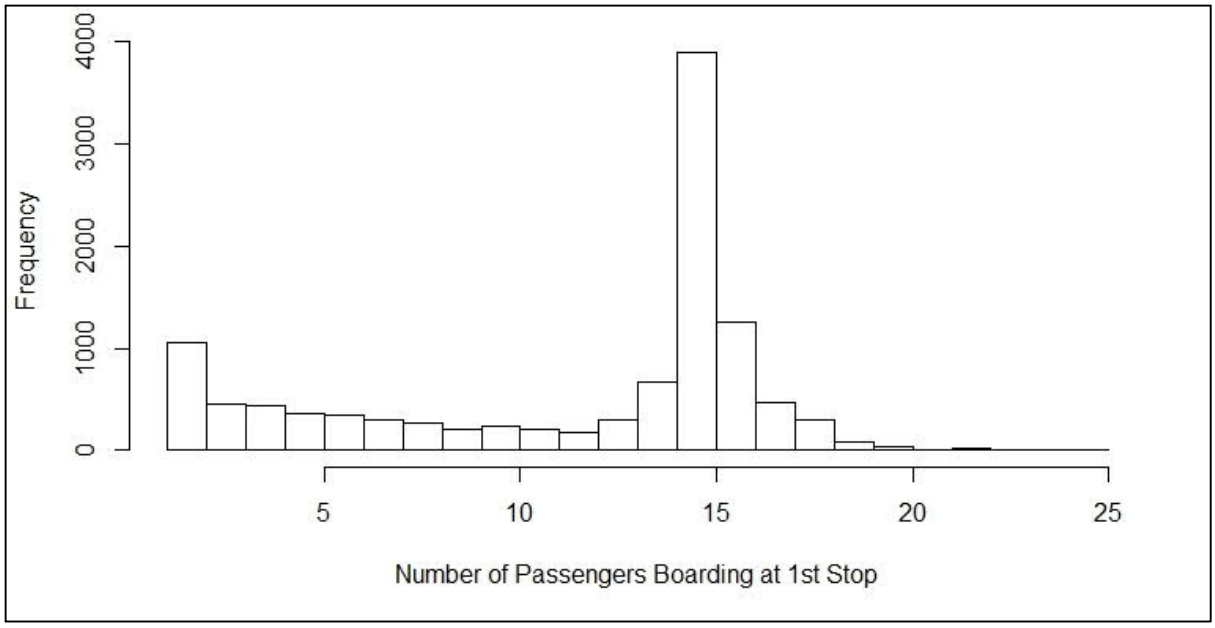


Figure 23: Passengers boarding at the 1st stop

As can be expected by these types of services that do not run according to schedule but are rather dictated by passenger demand, vehicles most often wait until they have a full load (15 or more passengers) before they depart from the rank. The high frequency of trips with only one passenger most likely refers to empty return trips where vehicles do not wait to fill up a load for the return trip but rather return as quickly as possible to the forward trip's origin rank to attempt to transport another full load, maximising revenue and profit made.

4.7 Fare

Each MBT operator is a cash business as all fares are paid in cash. Fares are generally regulated by regional MBT bodies that represent both the local and long-distance route operators (Mhlanga, 2017) and are usually prescribed as a maximum fare (Mxolisis, 2006).

Figure 24 shows the distribution of fare values across the trips surveyed.

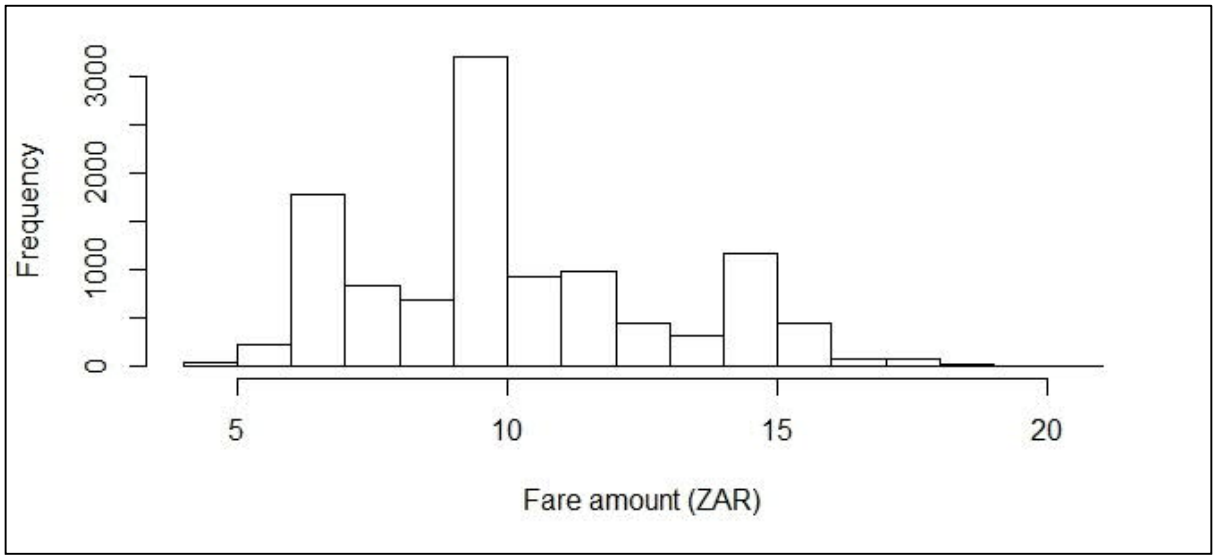


Figure 24: Histogram of fare per trip

The mode and median values observed in the range of fares was R10 while the mean value was found to be R10.55.

Taxi fares are not calculated on the basis of distance but what the fare for specific routes should be are agreed upon by the operator association bodies. Figure 25 provides an indication of the underlying relationship between route distance (trip distances used as a proxy in this case) and fare.

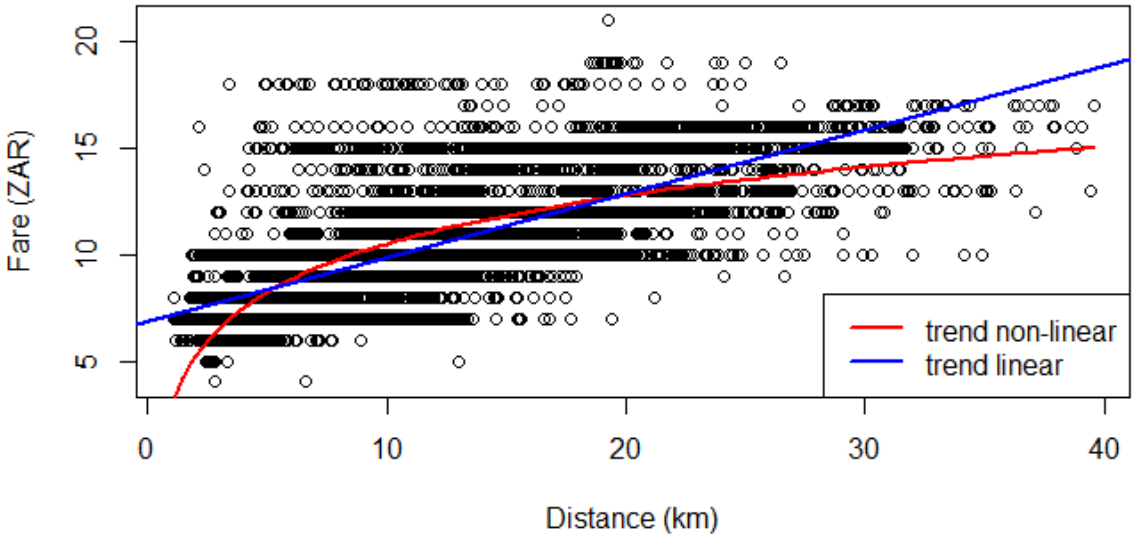


Figure 25: Relationship between trip route distance and fare

The curve ($y = 2.964 + 3.277ln(x) | \bar{R}^2 = 0.53$) fitted to these points by method of non-linear least squares parameter estimation is indicative of the increase of fare based on the distance of the routes. A linear function ($y = 6.890 + 0.303x \mid \bar{R}^2 = 0.56$) was fitted to these data purely for the purpose of comparing average per kilometre fare rates with other public transport modes – which are, however, not necessarily linear themselves.

With this linear relationship, the base fare is ZAR 6.89 with 30 cents per kilometre added thereafter. Both these relations are only valid for the observed distance range $1 \leq x \leq 40$.

4.8 Revenue

The revenue made on a trip depends on the total number of passengers transported and the value of the fare. Figure 26 shows the spread of revenue values observed.

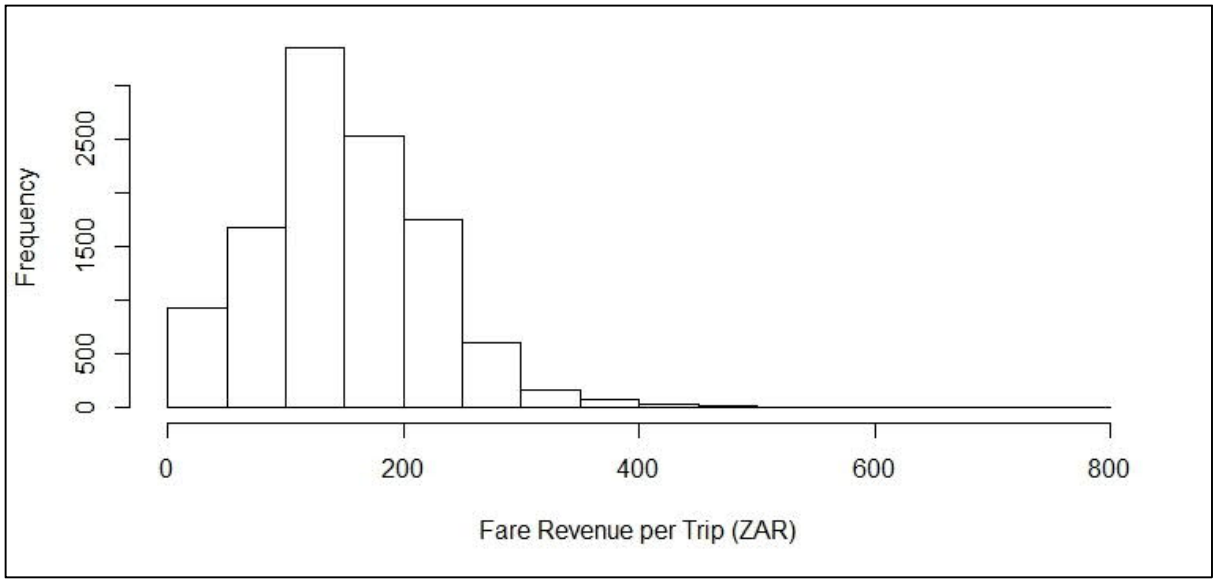


Figure 26: Histogram of trip fare revenues

Figure 26 shows that, regardless of trip distance and the fare, that MBTs tend to make on average around ZAR 150 in revenue. The 5th and 95th percentile revenue values determined were ZAR 40 and ZAR 270, respectively.

4.9 Relationships

As with the fare and distance, relationships exist between the other variables analysed. While the objective of this research does not include the development of robust models that explain the relationships between these variables, nor does the representativeness of the data allow for the development of statistically robust models, it is useful and interesting to visualise the data in terms of these relationships – or the lack thereof.

Developing an understanding the distribution of the data and the interrelationships between key variables within these data is serves both the objectives of obtaining a clear understanding of the operations of MBT services and classifying the route types by figuring out which attributes of the operations make the most sense to be used as variables on which to base a classification system.

Trivial relationships, such as distance/travel time, and number of passengers/revenue, inter alia that exhibit the expected relationships from these data, were mostly ignored in this discussion. Relationships that do not exhibit the expected function form have, however, been included.

While the relationships were evaluated in parallel with the analysis of the data distributions, and the knowledge about both these aspects are instrumental in the designation of the classification variables, the discussion on these relationships is included in Section 6.2. The discussion on the distributions and relationships are left until after the sections describing the classification in order for the reader to take cognisance of the observed route classes when considering this information.

4.10 Résumé

The visual representation of the distribution of the data provides an overview of the variation in service types offered by MBTs and despite the data limitations discussed in 1.6 and 3.3, it is clear that a variety of service types exist. It was, therefore, hypothesised that classifying these services according to service typologies will reveal different intra-class distributions of the operational characteristics examined in this chapter.

CHAPTER 5

5 Service Typology Classification

5.1 Background

To classify something is to categorise or to consider it as belonging to a specific group (Classify, 2018) based on ways that they are alike.

In order to classify the paratransit route types, it is necessary to understand which features are important in explaining how the observations vary from one another. One method that enables this is called Principle Component Analysis (PCA). PCA is a multivariate exploratory data analysis technique that makes it possible to visualise the relatedness (or variance) in data.

Another method that can be used to determine which features contribute to the data variance is Exploratory Factor Analysis (EFA). EFA is a method that attempts to identify patterns in data (Child, 2006) by reducing the observed variables into a smaller number of latent variables that share a common variance (Bartholomew *et al.*, 2011).

Ultimately, in order to be in a position to classify the routes it is required that the underlying data distribution structures are thoroughly understood, which was one of the objectives of the previous section.

5.1.1 Classification

One of the objectives of this research is to classify paratransit, in this case MBT, routes in terms of their service types or classes. The aim of this chapter is to establish a general service typology of MBT routes and to, if applicable, develop a method that determines to which class a route belongs.

Classification, which forms an integral part of data mining and machine learning, is concerned with assigning objects into defined categories based on their attributes. Examples of the application of classification include detecting spam email messages based on the subject and content, classifying of galaxies based on their shapes or categorising cancer cells as malignant or benign based on MRI scan results (Tan *et al.*, 2006).

Classification task input data is generally a collection of observations or records. Each record is characterised by a set of attributes (**X**) and a category, target or class attribute (***y***) (Tan *et al*., 2006). The attribute set may be comprised of discrete or continuous values or features while the class/category attribute must, by definition, be a discrete attribute (Tan *et al*., 2006).

Classification can also be described as the process of deriving a target function, or classification model, ***f*** that assigns each of the attribute sets **X** to one of the predefined categories (***y***). These models can either serve as explanatory tools or predictive tools, both of which would be useful in classifying the observed data and understanding the nature of paratransit operations.

Classification techniques are, however, most effective with describing or predicting datasets with discrete binary, dichotomous or nominal categories – in other words, qualitative data (Tan *et al*., 2006).

The quantitative data on which this research is based consists of records (observed paratransit trips) each of which inherently consist of continuous variables such as travel time, distance, number of passengers, and number of stops, etc.

Based on the fact that the definition of the target class (route type) of the paratransit data studied is unknown, an alternative "classification" method has to be used to divide the observed trip data into route type categories.

5.1.2 Discriminant Analysis

Discriminant analysis is a form of classification that finds a set of prediction functions based on independent variables that sort or classify the individual groups (Hintze, 2007). With discriminant analysis, known classifications of some observations are used to classify others and the number of classes is assumed to be known (Fraley & Raftery, 2002).

In the dataset on which this research is based upon, the classifications are not known but, for the purposes of this study, the number of classes or groups that describe the route operation type is three, namely, Trunk, Feeder and Hybrid. These names were chosen only for their parallelism with conventional public transport route type descriptions and will be reviewed on evaluation of the classification outputs.

5.1.3 Cluster Analysis

Cluster analysis is "the art of finding groups in data" (Rousseeuw & Kaufman, 1990). It divides data into groups that are meaningful, useful, or both, and is generally carried out for one or both of two purposes, namely, understanding or utility (Tan *et al*., 2006). It divides data objects into groups, or classes, based only on the attributes of the data objects.

Clustering for understanding is generally concerned with dividing objects, items, things into specific groups and assigning objects to these groups – in other words, classification. Cluster analysis is the study of techniques that help to understand data by automatically finding classes in the data (Tan *et al.,* 2006).

In the context of clustering for utility, cluster analysis provides an abstraction from individual data objects to the classes or clusters to which those data objects belong (Tan *et al*., 2006). In the context of clustering for utility, therefore, cluster analysis is the study of techniques that aim to find the most representative clusters of objects (Tan *et al.,* 2006).

Clustering or cluster analysis is concerned both with determining similarity and dissimilarity between objects. In general, cluster analysis attempts to minimise the intra-cluster distance (Euclidean or Manhattan) and maximise the inter-cluster distance for each variable on which the clustering was based.

One of the most efficient methods of clustering is called *'K-means Clustering'* (Ding & He, 2004). With K-means clustering, one selects the number of classes or cluster the data should be grouped into without knowing exactly what the definition of each group is. K-means clustering aims to partition data into K clusters by minimising the within-cluster sum of squares (WCSS), also known as 'sum of squared errors' (Ding & He, 2004).

K-means clustering, which was the choice of method used in this study, is carried out when the number of clusters is known. Choosing the optimum number of clusters is, however, not trivial and is often subjective and pre-determined by the research goal.

Describing MBT service routes, one would expect that some routes operate as traditional public transport trunk services, some routes operate as feeder or distribution services while others can be described as a hybrid between the former and the latter. This would lead to the obvious choice in the number of clusters as three clusters.

There are methods by which a more informed and subjective choice can be made by carrying out the clustering algorithms for a range of cluster number values and plotting the resultant WCSS against the number of clusters. The most efficient number of clusters can then be chosen from the graph from where the sharpest drop with respect to WCSS occurs.

5.1.4 Principle Component Analysis

Principle component analysis (PCA) is a multivariate dimensionality reduction technique that transforms high-dimensional data into lower dimensional data and identifies the dimensions with the largest variances (Ding & He, 2004). The objective of PCA is extract the most important information from a dataset by computing new variables called principle components as linear combinations of the original variables (Abdi & Williams, 2010).

Ding & He (2004) prove that principle components are effectively a continuous solution of k-means clustering membership indicators. PCA, by identifying the principle components and the underlying variables that contribute to their projection, identifies the variables that are most important in describing similarity or dissimilarity of individual data records.

PCA essentially computes eigenvectors and eigenvalues for linear combinations of the input variables. The eigenvector with the highest eigenvalue is defined as the principle component of the dataset, describing most of the variance in the data (Sayad, 2010).

While PCA is not an integral and mandatory part of cluster analysis, it can useful to determine which variables to use in the k-means clustering algorithm.

5.2 Clustering the Data

5.2.1 Route Types and choosing the number of clusters (k)

It was hypothesised that these data could be partitioned in a manner that identifies a hierarchy of services types. The City of Cape Town's CPTR (2004/5) differentiates between line-haul, feeder and distribution type services.

Another way in which these service types can be described is by their stop frequency. The Transit Capacity and Quality of Service Manual (TCQSM) (Transit Cooperative Research Programme (TCRP), 2013) provides definitions of service types by stopping patterns; Local, Limited-stop and Express. Although the TCQSM describes these services in the context of fixed-route operations, the definitions can be applied to paratransit for the sole purpose of classifying the service types.

Local services emphasise access over speed and are defined as a service that serves all stops along a route. These types of services can also operate as "flag-stop" services which allow passengers to be picked up and dropped off at any location on request (TCRP, 2013).

Limited Stop services balance access and speed and by definition serve only high-volume stops, providing passengers that use the service with fewer stops and therefore shorter travel times (TCRP, 2013).

Express services emphasise speed over access and are usually used for long distance trips. Express services often have only the two stops at either end of the assigned route, allowing the vehicles to operate at the maximum speed allowed by their operating environments.

Given the informal and often chaotic nature of these services, it is likely that some of these routes operate as express or limited stop services in the commuter peak periods and change their service to a feeder distribution function during the less busy during the inter-peak periods and vice versa.

The goal of the clustering, in this case, is to group the trip data collected into these service types based on the characteristics of the respective trips. As there is clear evidence of the temporal variation in the operations of routes, this clustering is also carried out for trips recorded during the different times of the day, i.e. the AM peak, inter-peak and PM peak.

5.2.2 Principle Component Analysis and Selecting the Clustering Variables

In order to carry out clustering, it is important to understand the underlying structure of the data, i.e. the linear relationship and correlation between certain variables and which variables are expected to, given the nature of the target clusters, contribute most to the grouping of said clusters.

For example, the data utilised for this study has as a variable the hour in which the trip was made (started), the route number, the date, the distance and duration of travel, the number of stops made, the number of passengers transported, maximum passenger load and speed. Common sense says that some of these variables will have little or no impact on the clustering, i.e. the hour in which the trip was made, while other variables are correlated, i.e. distance and the route number.

While PCA determines which variables describe the variance in the data the most, it does not make sense to include some of these variables in the PCA.

The variables that were carried into the PCA include the following:

- Distance (dist);
- Travel time (travelTime);
- Number of stops (numStops);
- Number of passengers (numPax);
- Speed (speed)
- Stop density (stopDens); and
- Passenger turnover (paxTurnover).

The full dataset was, therefore, trimmed down to contain only the above as shown in Table 4.

Table 4: Trimmed variables for PCA – showing the first six records in dataset.

Record	dist	travelTime	numStops	numPax	speed	stopDens	paxTurnover
1	9.7	20.0	7.0	16.0	29.0	0.72	2.3
2	9.8	26.0	8.0	18.0	22.7	0.81	2.3
3	9.7	20.0	6.0	15.0	29.0	0.62	2.5
4	9.9	25.0	8.0	18.0	23.7	0.81	2.3
5	9.7	26.0	8.0	20.0	22.5	0.82	2.5
6	9.8	25.0	7.0	15.0	23.5	0.71	2.1

Since these variables are in differing units and scales, the data were scaled in order to standardise the values, i.e. setting the means equal to zero. This was carried out by subtracting the variable's mean from each value and dividing the difference by the standard deviation:

$$X_{scaled} = (X - \mu)/\sigma$$

The result of the standardisation is a set of values for each variable with similar dimensions:

Table 5: First six rows of the scaled data

Record	dist	travelTime	numStops	numPax	speed	stopDens	paxTurnover
1	-0.3497	-0.4482	0.1090	0.3309	-0.0111	0.0219	-0.2908
2	-0.3296	0.0507	0.3578	0.7191	-0.5508	0.1971	-0.3108
3	-0.3526	-0.4482	-0.1397	0.1368	-0.0161	-0.1734	-0.1710
4	-0.3209	-0.0324	0.3578	0.7191	-0.4619	0.1877	-0.3108
5	-0.3425	0.0507	0.3578	1.1074	-0.5684	0.2116	-0.1710
6	-0.3310	-0.0324	0.1090	0.1368	-0.4761	0.0036	-0.3706

PCA carried out with these seven variables results in the following components showing the relative importance of each component (eigenvalues) in describing the variance in the dataset.

Table 6: Principle component eigenvalues

Importance of components	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	1.5603	1.5057	1.0642	0.9152	0.4353	0.3222	0.1852
Proportion of Variance	0.3478	0.3239	0.1618	0.1197	0.0271	0.0148	0.0049
Cumulative Proportion	0.3478	0.6717	0.8335	**0.9532**	0.9803	0.9951	1.0000

Table 6 shows that the first four components explain 95% of the variance in the data. The loadings (eigenvectors) for these components are shown in Table 7.

Table 7: Principle component loadings (eigenvectors)

Loadings:	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
dist	-0.494	0.379	-0.156		0.425	-0.177	0.611
travelTime	-0.241	0.523		0.508	0.203		-0.604
numStops	0.263	0.561		-0.265	-0.189	0.695	0.159
numPax		0.426	0.663	-0.237	-0.311	-0.474	
speed	-0.449		-0.185	-0.737			-0.463
stopDens	0.579			-0.256	0.738	-0.145	-0.144
paxTurnover	-0.302	-0.277	0.702		0.316	0.486	

The combined contribution, for the first four principle components, which explain 95% of the variance, shows that the variables most important in explaining the variance in the data are, in

order of importance, passenger turnover, number of passengers, speed, travel time, distance, number of stops and stop density. The relative difference in contribution of these variables is, however, relatively small at ±8%.

It was therefore decided that the traditional definition of services types, which are typically defined in terms of their distances and stop frequencies, shall be used to classify the routes. The variables distance, stop density (linear combination of number of stops and distance) and passenger turnover (linear combination of number of stops and number of passengers) were therefore chosen to carry out the first clustering experiment (km1).

5.2.3 K-means clustering

The first clustering experiment (km1) was carried out using as input variables distance (dist), number of stops (numStops) and passenger turnover (paxTurnover).

As one requires as an input to k-means clustering the number of clusters, an iteration of R's built in k-means clustering algorithm 'kmeans' to determine the number of clusters above which the returns begin to diminish. Figure 27 is a "scree plot", showing the descriptive effect of the number of clusters. It shows that after the first three clusters, the relative combined effect of increasing the number of clusters is not expected to significantly improve the within cluster sum of squares or "sum of squared errors".

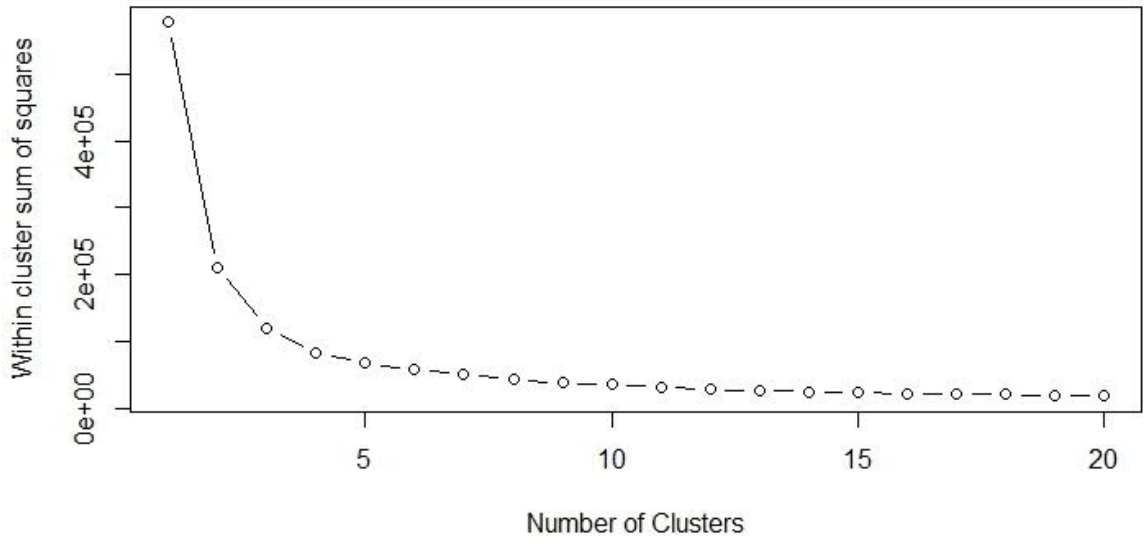


Figure 27: Scree plot for cluster experiment one (km1)

Figure 28 represents the three clusters identified projected in 3-Dimensional space. Each component does not represent the variables used to carry out the clustering (distance, number of stops and passenger turnover) but linear combinations of the components (as with principle components).

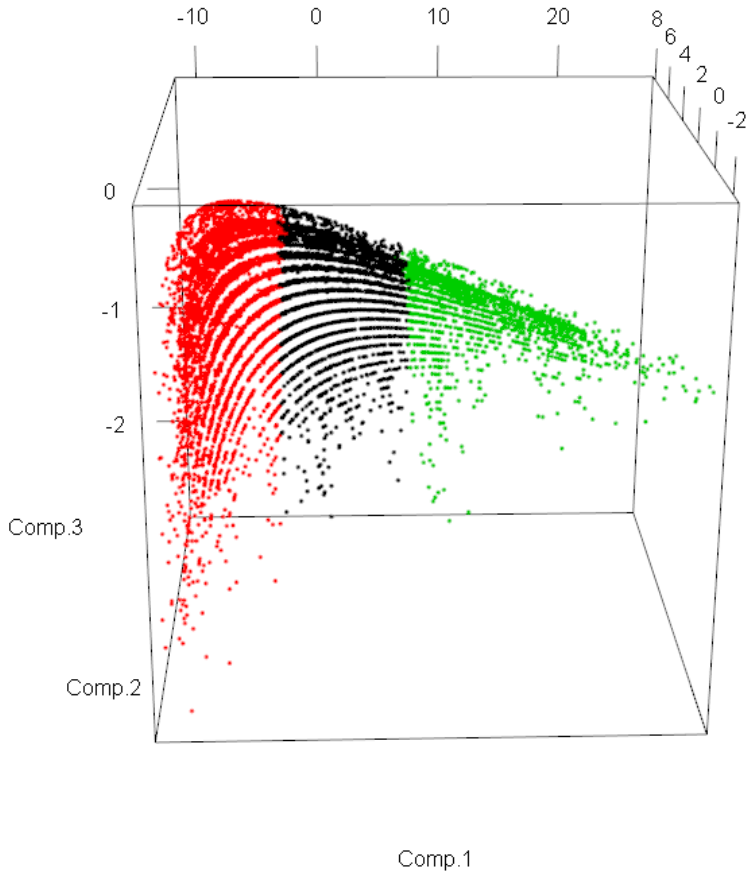


Figure 28: 3D plot of the cluster component projections (km1)

Two additional k-means experiments (km2 and km3) were carried out to benchmark the "goodness of fit" of km1.

K-means experiment km2 was carried out with clustering variables distance, speed and number of stops while km3 was carried out using distance, number of stops and number of passengers.

Since the clustering was carried out at the trip level and each trip belongs to a specific route, the "goodness of fit" approximation was to determine how well the clustering assigned or kept trips from the same route within the same cluster. Figure 29 shows that km1 performs better in the task of grouping the routes to specific clusters.

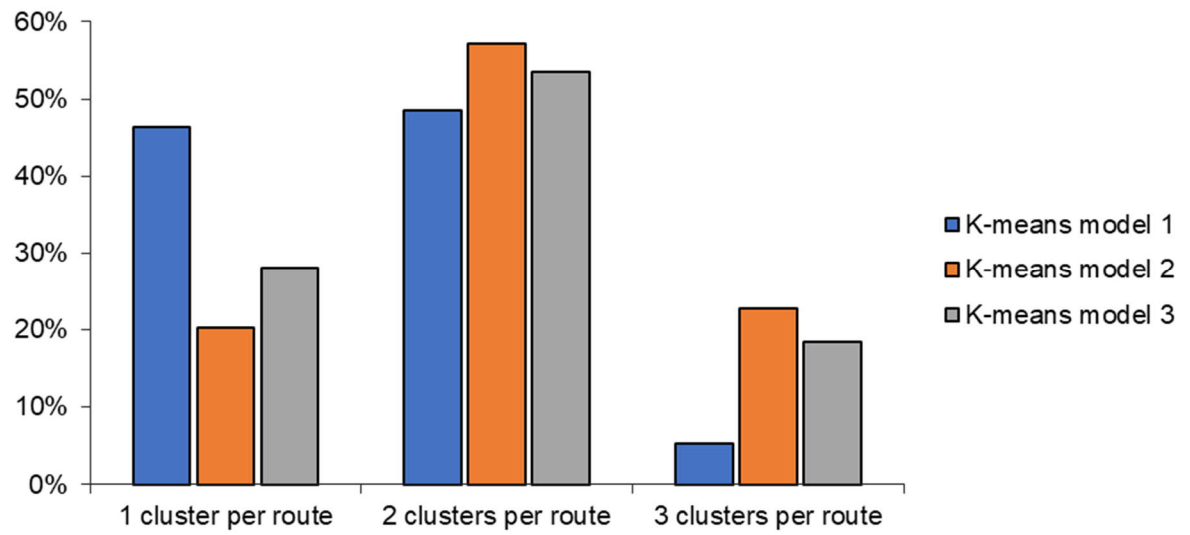


Figure 29: K-means model "goodness of fit" comparison

Another quantitative measure that was derived to determine the performance of the model was the “max sum proportion” which was determined by calculating the proportion of trips belonging to each route that were allocated to each cluster and summing the maximum of this value for each route for all routes. If, for example, route COCT001 had 40 trips of which 28 trips (70%) were assigned to cluster 2 and 12 assigned to cluster 3 (30%), the maximum proportion of for this route would be 0.70. By summing this proportion for every route in the dataset for each cluster model and dividing it by the total number of routes in the dataset, results in the *max sum proportion*. The max sum proportion calculated for km1, km2 and km3 were 0.90, 0.78 and 0.83 respectively.

Given the objective of deriving a method to use onboard data to classify MBT routes, km1 is considered a sufficiently descriptive grouping of clusters.

5.2.4 Classifying the Route Clusters

Of the clustering procedures tested, km1 yielded the most close-fitting results when evaluating the assignment of specific route's trips to a cluster group. The clusters, or classes, have not been given names yet as their underlying structure has not been discussed.

Table 8 shows the median and mean values for each one of the original continuous variables in km1.

Table 8: Median and mean values for the base variables in each cluster group (km1)

Cluster	Measure	Distance (km)	Travel Time (min)	No. of Stops	No. Of Passengers (pax)	Speed (km/h)	Stop Density (stop/km)	Passenger Turnover[7] (Pax/stop)
1	Median	22.18	37.00	6.00	15.00	37.92	0.26	2.50 (0.54)
	Mean	23.42	37.04	6.93	15.74	41.20	0.31	3.09 (0.81)
2	Median	13.14	29.00	7.00	15.00	27.35	0.51	2.00 (0.60)
	Mean	13.46	30.16	7.55	15.19	28.90	0.58	2.57 (0.79)
3	Median	6.07	15.00	5.00	15.00	23.25	0.89	2.33 (0.67)
	Mean	6.09	16.32	5.55	12.91	24.22	1.00	2.88 (0.86)

Since clustering was carried out on the trip level and since the objectives of this research included classifying the routes in terms of their operational characteristics and creating a spatial representation thereof, it was necessary to allocate a route's trips to a specific cluster based on their trips' cluster membership frequency. The "goodness of fit" approximation *max sum proportion* discussed in section 5.2.3 was developed specifically for this reason.

Table 9 shows that the re-allocation of trips to specific clusters based on their route does not have a noticeable effect on the median and mean values in Table 8.

Table 9: Median and mean values for the base variables in each cluster group (km1 routes allocated per cluster)

Cluster	Measure	Distance (km)	Travel Time (min)	No. of Stops	No. Of Passengers (pax)	Speed (km/h)	Stop Density (stop/km)	Passenger Turnover[8] (Pax/stop)
1	Median	22.01	36.00	6.00	15.00	37.92	0.25	2.50 (0.54)
	Mean	22.98	36.23	6.81	15.63	41.15	0.31	3.11 (0.81)
2	Median	13.12	29.00	7.00	15.00	27.43	0.51	2.13 (0.60)
	Mean	13.39	29.81	7.43	15.18	28.95	0.58	2.62 (0.77)
3	Median	6.16	16.00	5.00	15.00	23.23	0.88	2.25 (0.67)
	Mean	6.00	17.14	5.74	13.00	24.25	0.99	2.83 (0.87)

Based on the median and mean values the following definitions have been selected for the respective clusters:

[7] Values in brackets indicate the passenger turnover value when excluding the 1st stop's values. The values smaller than 1 indicate that passengers alighting en route exceed the passengers boarding en route

[8] Values in brackets indicate the passenger turnover value when excluding the 1st stop's values. The values smaller than 1 indicate that passengers alighting en route exceed the passengers boarding en route

- Cluster 1: Trunk routes (Long distance, high speeds, low stop density) – 2,111 observations (18.9% of observed trips)
- Cluster 2: Intermediate routes (middle distance, lower speeds, medium stop density) – 4,083 observations (36.6% of observed trips)
- Cluster 3: Feeder/Distribution routes (short distances, lowest speeds, highest stop densities) – 4,963 observations (44.2% of observed trips)

The survey project's scope required surveying each route for a finite and equal number of trips. The data extracted for the research, however, did not yet have complete sets of trips of trips for all routes and for some routes some of the observations were removed in the cleansing process.

In order to determine the relative proportions of each route type in the surveyed data, the number of trips per route needs to be taken into account.

Of the 555 routes, of which acceptable data remained after cleansing, the following proportions were identified to belong to the respective route types:

- Trunk routes: 20% of the routes surveyed (108 routes)
- Intermediate routes: 32% of the routes surveyed (179 routes)
- Feeder routes: 48% of the routes surveyed (268 routes)

5.3 Spatial Representation of the Clusters

5.3.1 The Three Route Types

One of the objectives of this study was to spatially represent the routes in terms of their service types. The previous section identified the classification method, k-means clustering, which was used to group the MBT routes surveyed into three distinct classes. These classes were, in an effort to stick to traditional public transport route hierarchies, termed trunk, intermediate and feeder routes. Trunk-feeder route systems typically only distinguish between trunk and feeder route types. A third type, called intermediate routes solely for the purposes of this study, was added as a high variation in operational characteristics was observed in the exploratory data analysis and principle component analysis.

5.3.2 Grouping the KML files

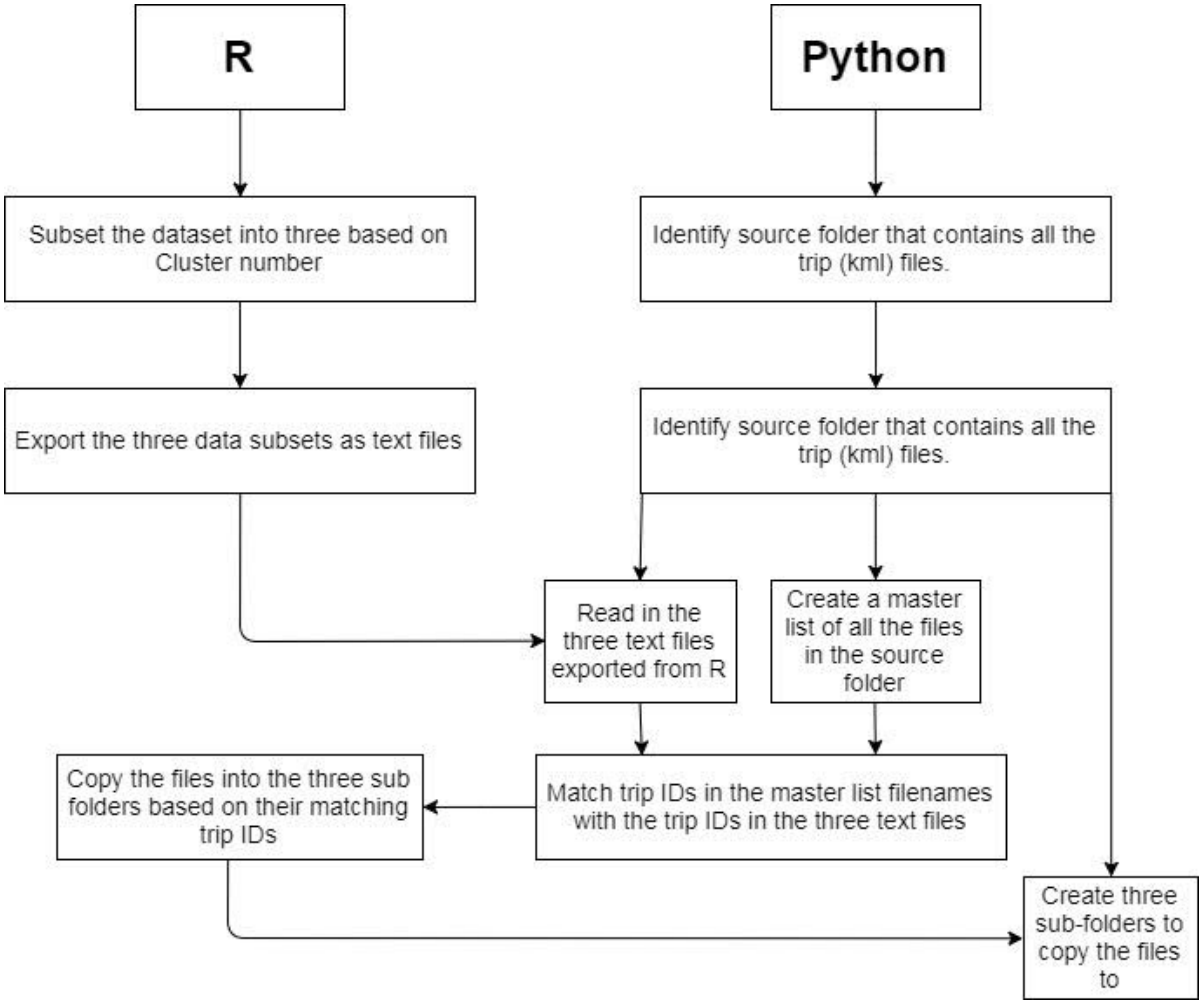
The early stage development of GoMetro's data platform was not yet fully operational at the time of extracting these data to carry out the research. As such, a raw data download was opted for which stored the LineString (GPS route trace) and point data for each MBT trip as separate KML files while the descriptive information (distance, travel time, etc.) was download as a CSV file. Each KML file is associated with a record in the CSV by sharing a unique identifier (Trip ID) as detailed in section 3.2.6.

During the clustering exercise, a cluster number was appended to each data record to identify which cluster (route type) it belongs to. Once it was identified which Trip IDs belonged to a specific route type (cluster number), it was necessary to associate the cluster number of each trip with the two KML files (route and stops) for each trip.

The objective of representing the identified route types spatially could be most simply achieved as displaying the routes using different layers on a map using GIS software. The

software chosen for this purpose was QGIS, a free and open-source desktop GIS platform. In order to generate different layers, the files needed to be grouped so that they could be imported as separate layers in to QGIS. Since the cleansed and clustered data consisted of 11,157 individual records, each of which have two KML files associated with them, manually copying the files into separate directories was not a feasible option.

A bespoke file copying programme needed to be developed to copy the files from their parent directory into three separate sub-directories, one for each cluster, based on the Trip ID value in their filenames. A procedure and programme was developed for this purpose using a combination of R and Python programming languages. The procedure worked as follows:



The code (with annotations) that carried out the process above, using R and Python syntax, is as shown in Appendix B.

This process was carried out for all trips according to their cluster allocations where each route was assigned a cluster based on its trip cluster membership as described in Section 5.2.4.

A trip belonging to each route's alignment was selected randomly for each route code to represent the route's alignment in the route type layers that were to be created in QGIS. This process has its shortcomings in that not each trip recorded for a specific route followed the

exact alignment. Displaying each trip recorded, however, defeats the purpose of this exercise of spatially representing the route type classifications.

For ease of layer manipulation and reduction in file size, the individual route files making up each one of the route type layers were combined using a free online tool, KML Merger[9], which combines separate KML files into one singular file or layer.

5.3.3 Spatial representation of the route classification

The result of classifying the routes and spatially representing the different route types in QGIS is given in the following Figure 30, Figure 31 Figure 32 respectively.

[9] https://kmlmerger.com/#

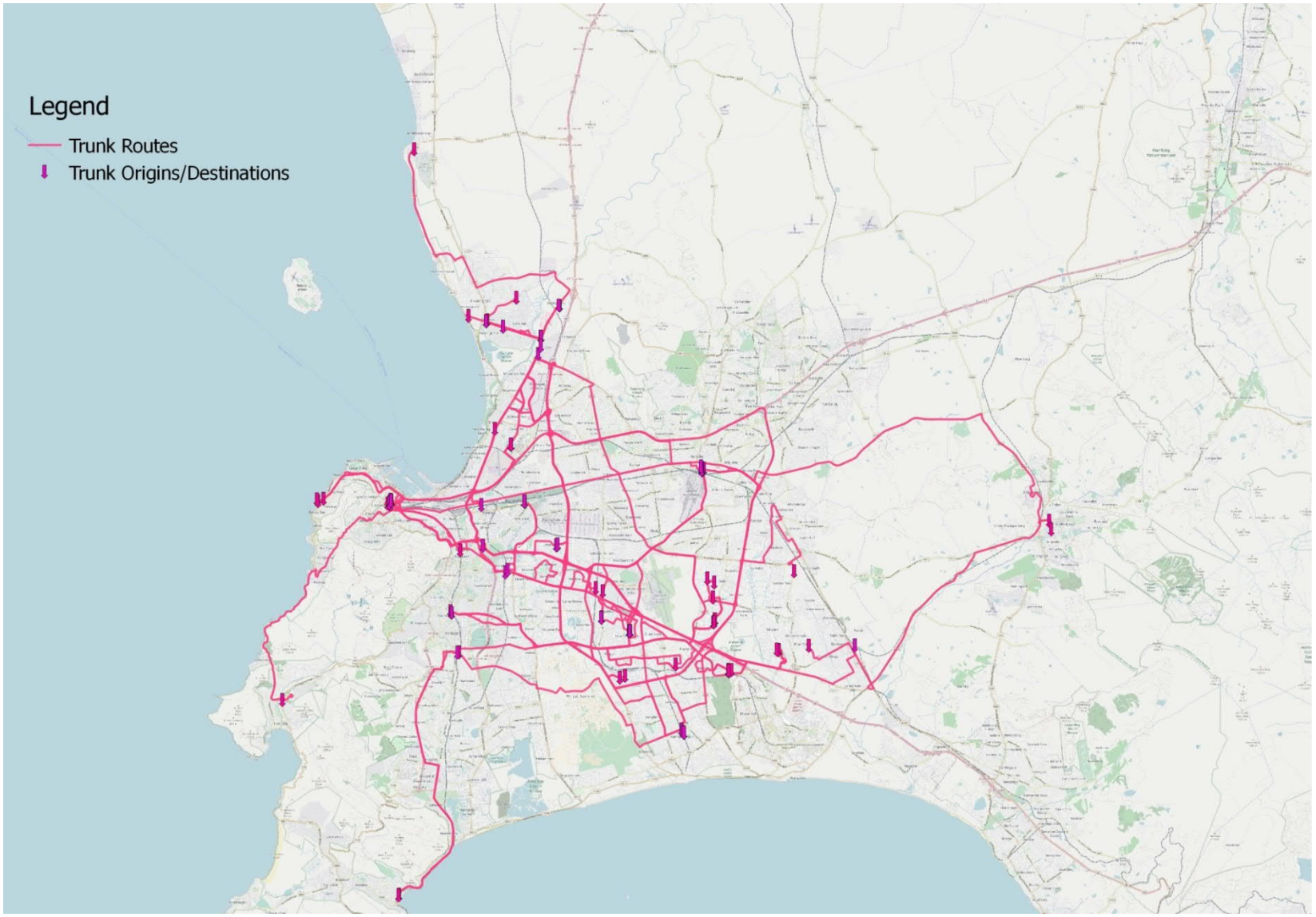


Figure 30: Trunk routes and connecting ranks

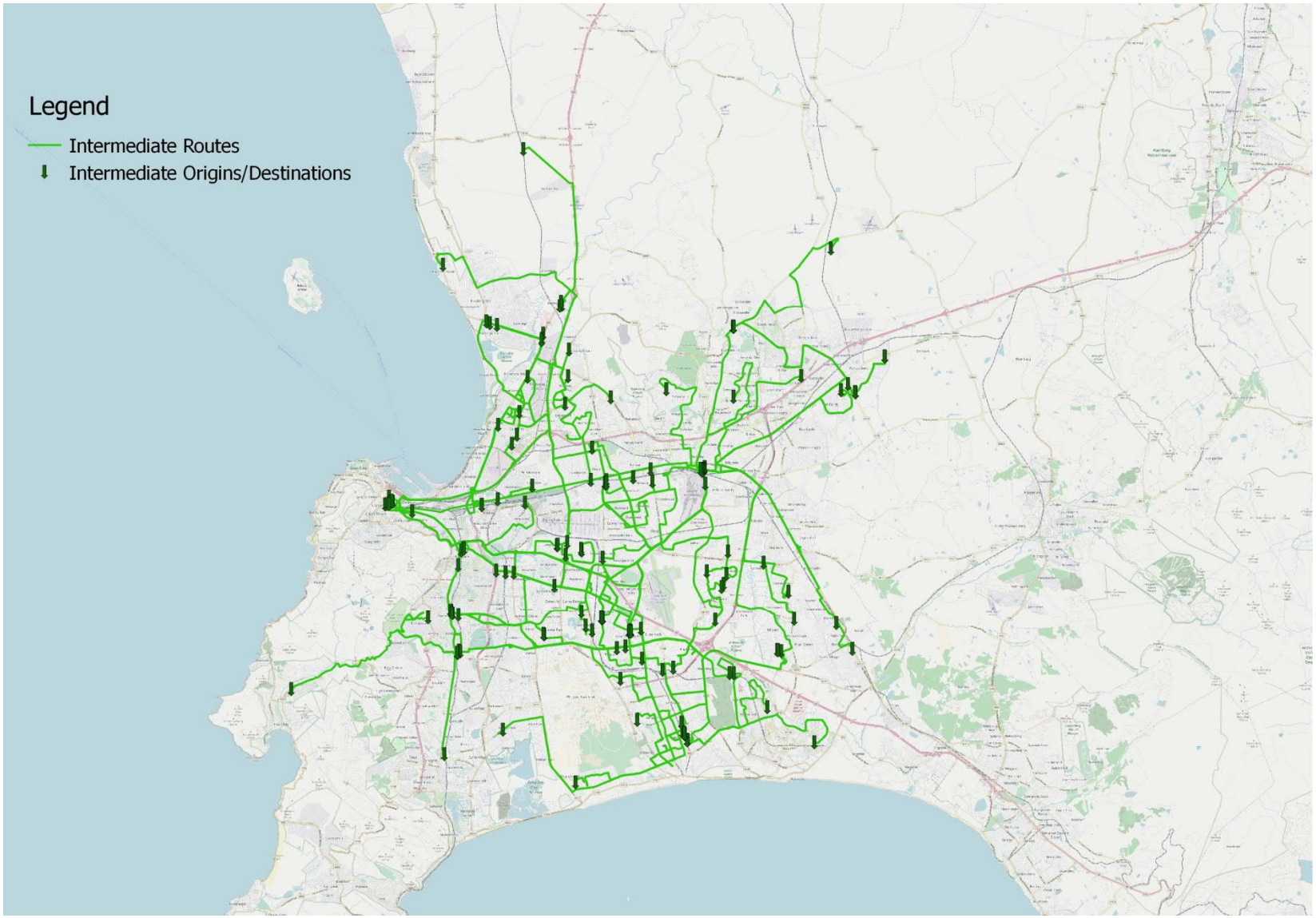


Figure 31: Intermediate routes and connecting ranks

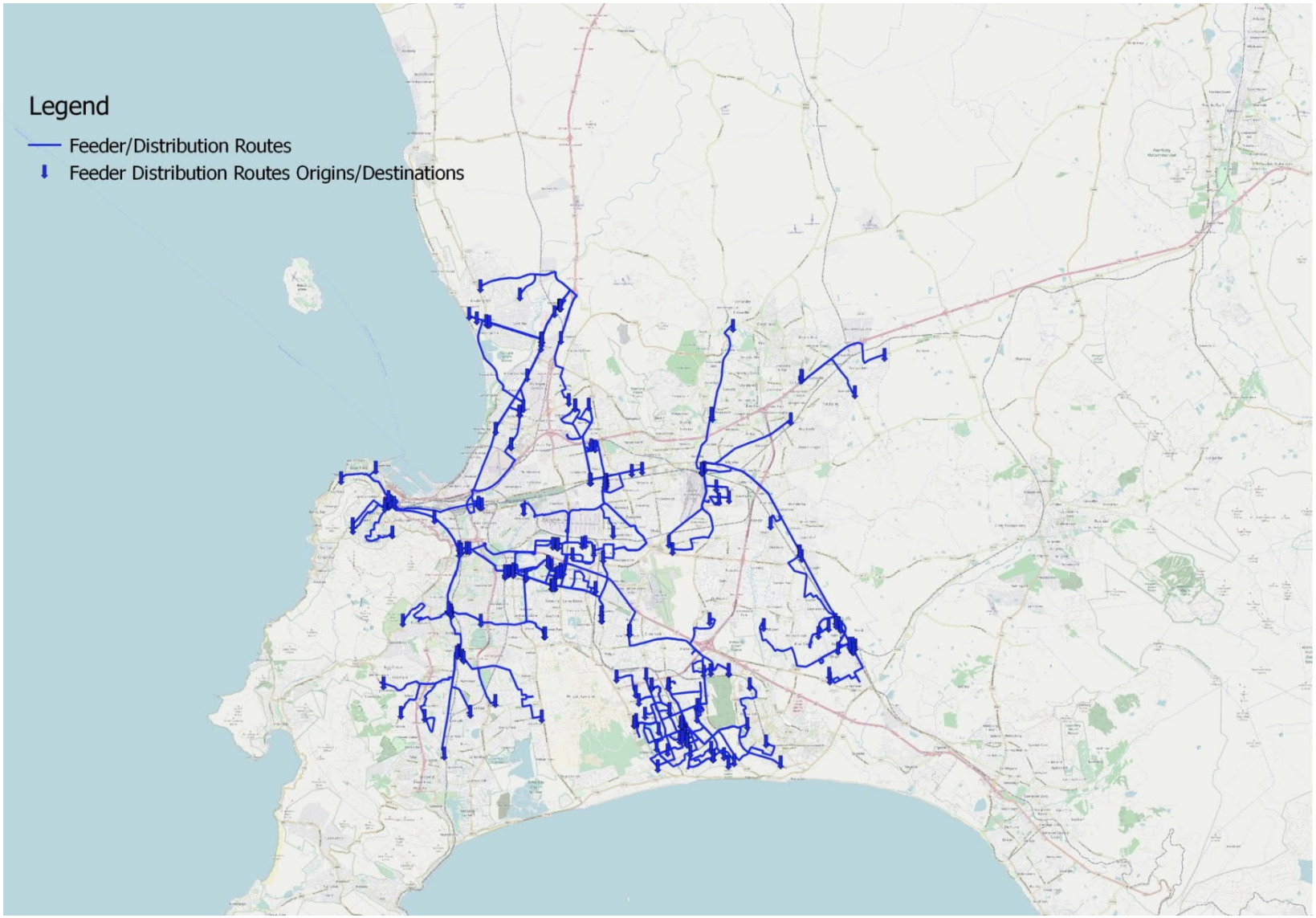


Figure 32: Feeder/distribution routes and connecting ranks

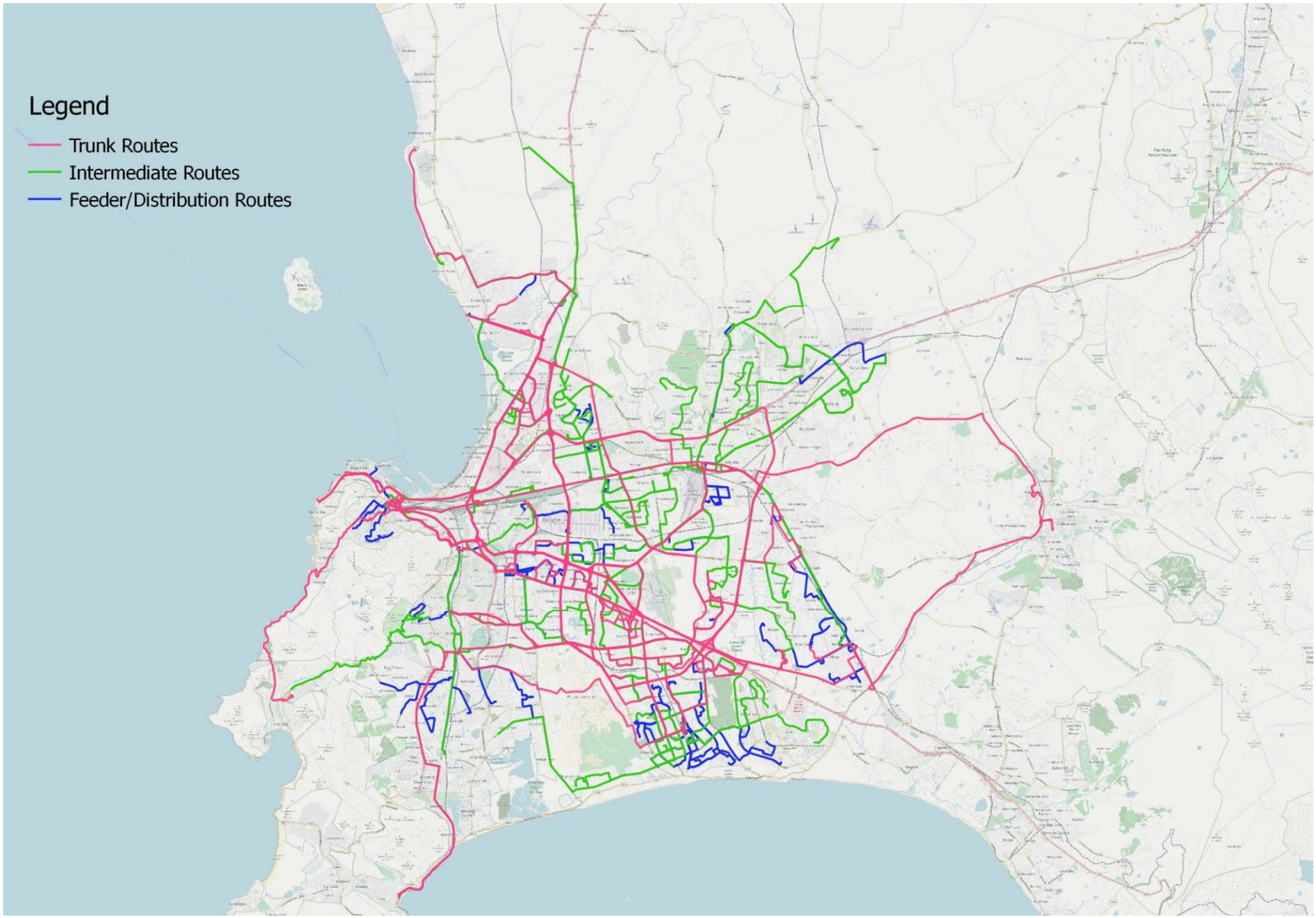


Figure 33: Trunk, intermediate and feeder/distribution routes

CHAPTER 6

6 Discussion

The purpose of this section of the report is to summarise the results of the analyses documented in the previous two sections and to discuss the key findings.

6.1 Distribution of operational characteristics

6.1.1 Distance, Speed and Travel Time

Distance

The analysis of the trip and route distances, speeds and travel times from these data for the various routes has revealed not only that routes differ in terms of operational characteristics from each other but that each trip for a given route may differ significantly from another in terms of its operating speed and the actual route, and therefore distance, travelled.

The MBT routes in Cape Town vary in distance from 1km through to about 50km. The bulk of these routes (70%) are between 5km and 20km in distance with the mean distance being 13.77km (median 12km).

Looking at individual routes, however, confirms the common knowledge of these services not operating on fixed routes. Figure 34 shows the different paths taken by vehicles for the route between Lower Crossroads and Claremont.

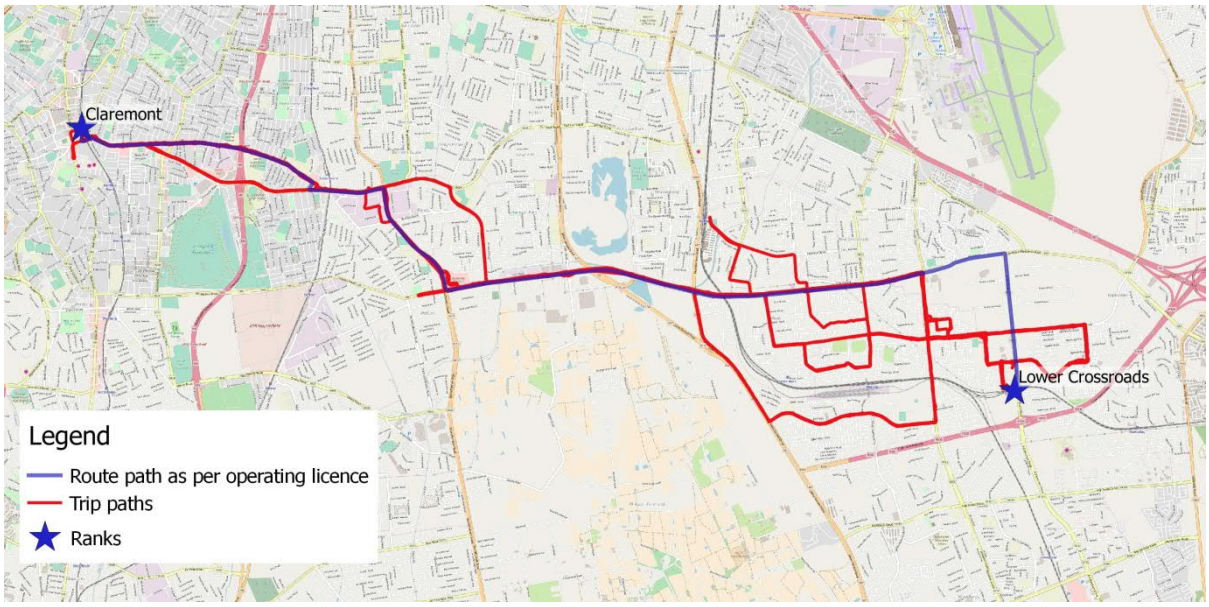


Figure 34: Different paths taken for the same route (Lower Crossroads - Claremont)

Figure 35 shows the variation in trip distance for the Lower Crossroads to Claremont route shown in Figure 34. The official route distance is 16.2km while the mean distance travelled on the surveyed trips is 16.14km and the standard deviation is 662m. The maximum distance travelled was 18.18km, 1.98 longer than the official route's distance. This phenomenon is common practice in the MBT industry with drivers using their specialist institutional knowledge to respond to situations, such as differing levels of demand.

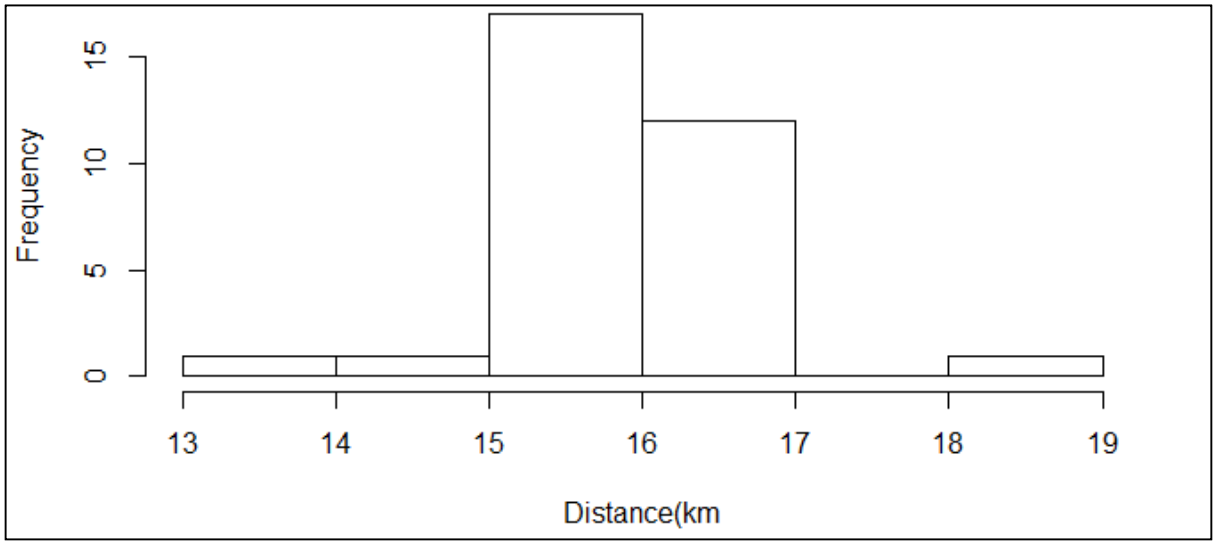


Figure 35: Variation in distance of the Lower Crossroads to Claremont route

The variation in route distances and the variation in trip distances for a specific route may seem a trivial concept but forms an important component of the overall understanding of the nature of these services, something that becomes even more apparent when considering the classification of routes in terms of their service types.

Speeds

The speed at which these routes operate, similarly, do not only vary from route to route but different trips are operate at significantly different speeds. The speeds recorded per trip were obviously subject to conditions such as congestion, weather, incidents and traffic enforcement that influences the route choices which in turn influences the average operating speeds. Figure 36 below shows the distribution of average speeds recorded for the same route, Lower Crossroads to Claremont, discussed in for variation in route distances above.

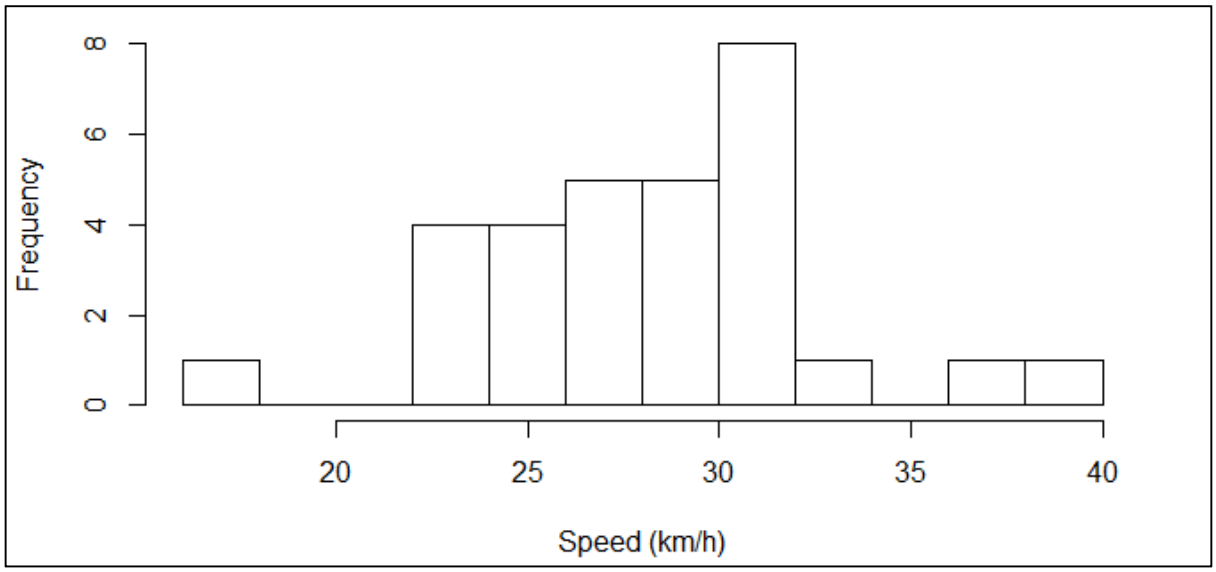


Figure 36: Average operating speed of the Lower Crossroads to Claremont route

There is also a clear difference in speeds, as shown in Figure 16 and Figure 17 depending on the direction of the route, forward or reverse, for a specific route origin – destination pair. It was determined that, on average, that routes were 27% faster in one direction than in the opposite. In general, the average route speeds determined from this study is comparable or slightly faster than reported for MBTs, and other modes, in the 2017/18 CITP.

Table 10: Average speed by mode (source CoCT CITP 2017/18)

Mode	Speed
Rail	23.4km/h
Contracted Bus (Golden Arrow, Sibanye)	18.1km/h
BRT (MyCiti)	12km/h
Minibus Taxi	21.5km/h
Minibus Taxi (this study)	29.2km/h

Travel Time

Travel time was less of an interesting measure to analyse as it is a direct function of the speed and distance discussed above. It is interesting, however, to note that the bulk of trips surveyed, and assuming the sampling margin of error approximated suggests it is representative of all routes, routes with a distance shorter than 40km, were between 5min and 30min in duration. The mean and median travel times were found to be 25.4min and 24min respectively which is significantly lower than the 53min average travel time for MBTs as reported in the 2017/18 CITP. It must, however, be added that routes between 40km and 52km (the longest route as per TRS) were not surveyed and that it can be deduced that the average travel times would increase if these were taken into consideration.

6.1.2 Stops per Trip, Passengers per Trip and Load Factor

The histogram shown in Section 6.1.2 shows that a large proportion of trips, 28.7%, surveyed transported 15 passengers, which is the passenger capacity of the vast majority of vehicles operating in Cape Town. It also shows that a large proportion of trips, 37.4%, transported more than 15 passengers. Approximately a third, 33.8%, of the trips surveyed transported fewer than 15 passengers. The 95th percentile value of number of passengers transported is 22.

Evaluating the number of stops made by the vehicles surveyed, it can be observed that the most common stop frequency was two stops which, in this survey data, represent the two rank stops at either end of the trip. A total of 1,378 of the 11,157 trips that comprise the analysed data, 12.4% of the trips, did not stop once along the way on their journey.

Of the trips that only made two stops, a total of 35.4% transported 15 passengers and 15% transported more than 15 passengers. A total of 50% of trips surveyed, therefore, only made two stops yet transported fewer passengers than most of the vehicles can legally accommodate. For the trips that only transported two passengers, it was also found that mean value of the maximum segment load of these trips was 11.16 passengers which, incidentally, is the same as the mean of the average segment load for these trips. Of these trips only making two stops, 500, 351 and 527 of the 1,378 trips were made during the morning,

midday and afternoon peak periods respectively which is distributed according to the volume of trips surveyed during each peak period vey consistently (3, 2, 3).

It can also be deduced from the histograms representing the average and maximum segment loads that the vast majority of trips had a maximum segment load of 15 passengers and that a very small percentage of trips (5.7%) had an average segment load of more than 15 passengers while a total of 26.5% had a maximum segment load more than 15. About a quarter of the trips surveyed, therefore, overloaded their vehicles for at least one trip segment – assuming of course that these vehicles were all 16 seaters (the driver counts as a person in the seating capacity ratings).

Assessing the above information from the opposite perspective, determining the number of stops made by vehicles transporting more than 15 passengers, it was found that on average number of stops made was approximately 8.8 stops per trip while the average number of stops made per trip for all trips was 6.6 stops. It is shown in Figure 40 that, as expected, a linear relationship between the number of stops and the number of passengers exists.

6.1.3 Fare and Revenue

Since one of the trip attributes captured during the onboard data collection was the cash amount paid by each passenger, the revenue earned by each trip could be determined.

The mean fare found for the Cape Town MBT routes was around 10 ZAR (standard deviation 2.8 ZAR) while the mean trip revenue was around 150 ZAR (standard deviation 73 ZAR), which could be expected with mean passengers per trip being 15.

6.2 Key Relationships

Key relationships between the aforementioned groups of trip properties that were determined after evaluating the distribution of values for each variable include the following:

- Fare and distance;
- Distance and number of stops;
- Distance and number of passengers;
- Number of passengers and number of stops;
- Number of stops and speed; and
- Speed and Distance.

6.2.1 Fare and Distance

Fare and distance, discussed in section 4.7 above, is a trivial relationship but has been included as it adds to the insight into what the range of services costs are. As expected, the fare increases with increasing distance.

6.2.2 Distance and Number of Stops

Given that MBT services are to some extent opportunistic and respond the passenger demand as it arises, often even if there is no capacity left onboard, there is**,** counterintuitively, **no clear relationship** between route or trip distance and the number of stops per trip. Figure 37 shows the spread of distance against stops observed from the survey data.

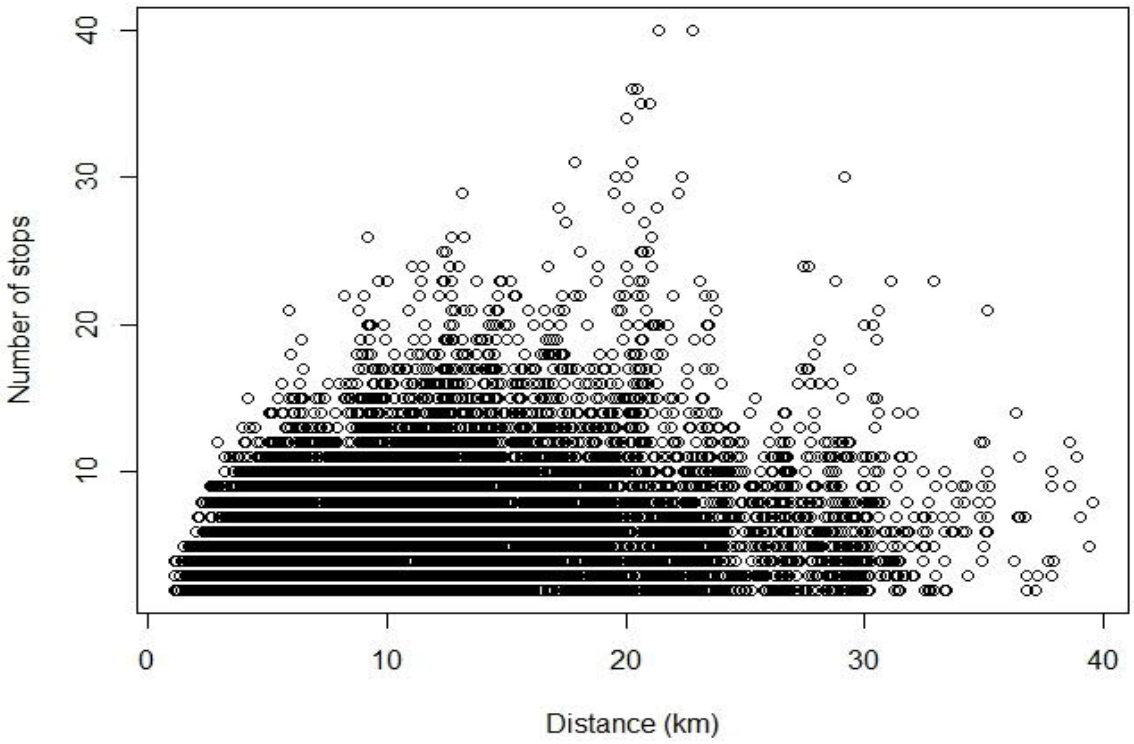


Figure 37: Distance and number of stops

Plotting the distance against the stop density (distance / number of stops), one would clearly expect a hyperbolic relationship ($f(x) = k/x$). Figure 38 confirms this relationship.

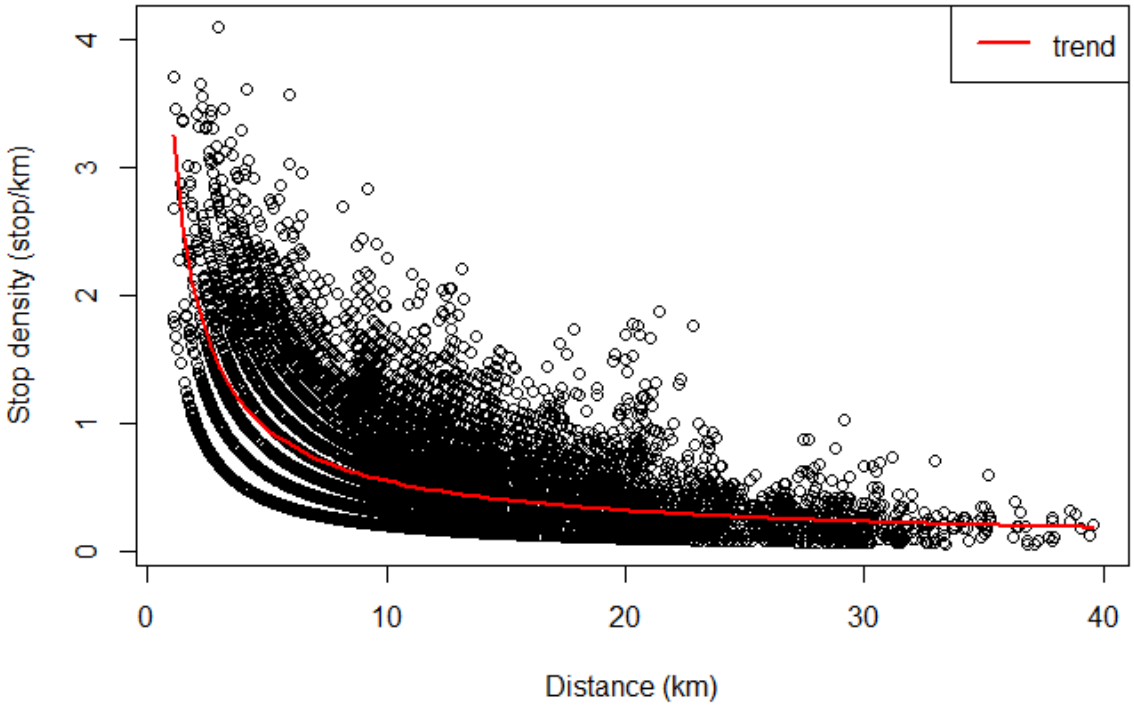


Figure 38: Relationship between distance and density of stops

The trend line shown in the graph ($y = 3.504x^{-0.798},\ x \in \mathbb{R} | 1 \leq x \leq 40 | R^2 = 0.42$) is shown to demonstrate the underlying relationship between these variables, albeit a relatively

week relationship with an R^2 of 0.42. The portioned appearance of the data plots can be attributed to the *number of stops* value being an integer.

6.2.3 Distance and Number of Passengers

Figure 39 shows that most trips, irrespective of their distance, have between 10 and 20 passengers. While this was already shown in distance histograms in section 4.1, Figure 39 shows that many of the shorter distance trips often transported lower passenger numbers. It should, however, be noted that these trip points are for both the forward and reverse directions during the commuter peak and inter-peak periods. Where vehicles on certain routes mostly transported full loads and emptier loads on the return trip (which often is the case) can unfortunately not be differentiated from these points. This figure therefore only serves to show most trips transported 15 or so passengers (median value 15, mean value 14.3) and most points are within the 5 – 15km distance range – information shown in the respective passenger numbers and distance histograms.

What could possibly be inferred from the relationship (or lack thereof) below is that vehicles might return to the busier rank with fewer passengers in order to make another full load in the peak direction for the shorter distance routes whereas for the longer distance routes, the vehicles would only return once their vehicles have filled up, thereby reducing their "dead passenger kilometres".

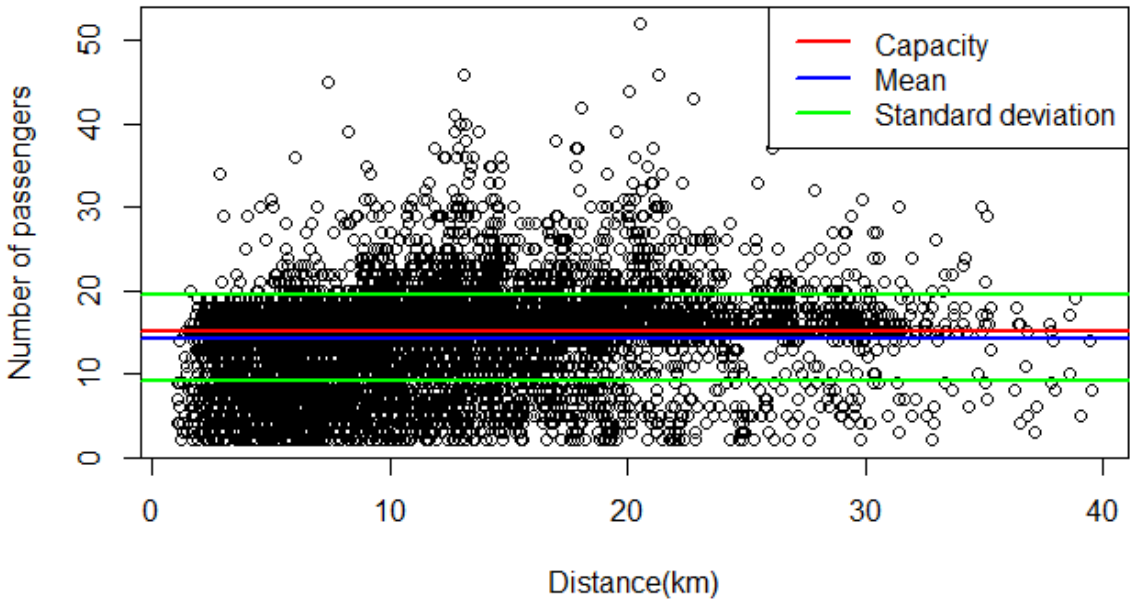


Figure 39: Distance and number passengers per trip

6.2.4 Number of Stops and Number Passengers

Figure 40 shows a positive linear relationship($y = 0.427x + 0.464 \mid x \epsilon \mathbb{R} \mid R^2 = 0.30$) between stops per trip and the number of passengers transported. The low coefficient of determination value, suggesting a relatively weak relationship between the number of stops and the number of passengers, shows how different the service types are, where some trips collect a full load of passengers at the first stop and have high passenger turnover along the route while others have no passenger activity along the way.

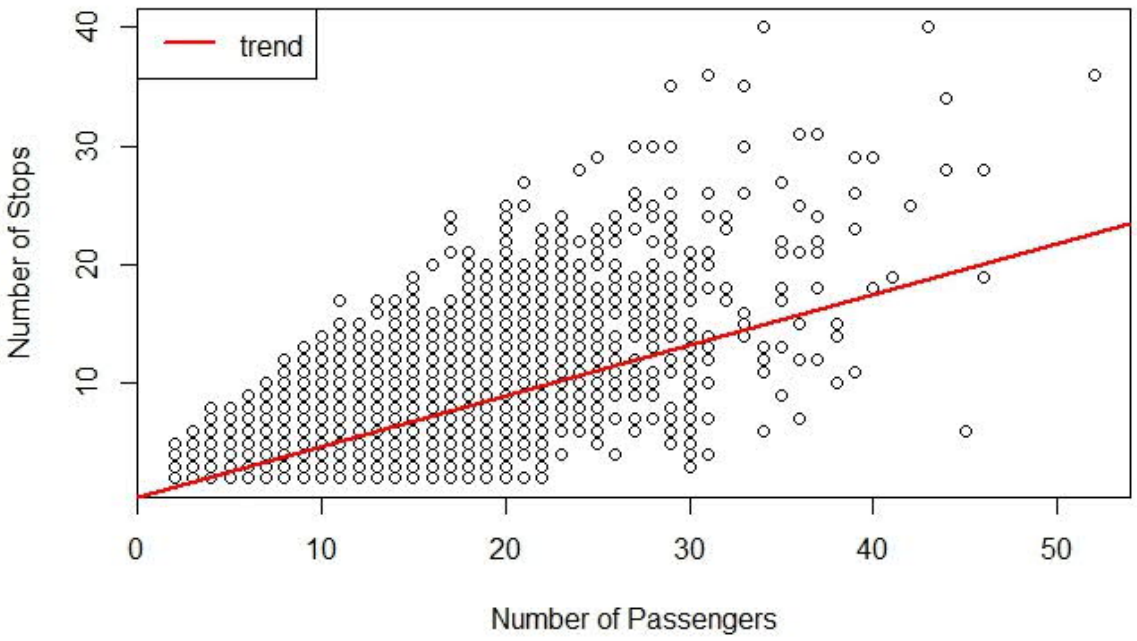


Figure 40: Relationship between stops per trip and passengers per trip

6.2.5 Number of stops and speed

As is the case with any form of passenger transport service, the more frequently a vehicle stops, the lower the average operating speed will be – where operating speeds is defined as the distance divided by the total trip time (including dwell times at interim stops). Figure 41 suggests that as the number of stops increase, the operating speed decreases, but that at the lower stop values, there is a high variance in observed speeds with almost no discernible relationship.

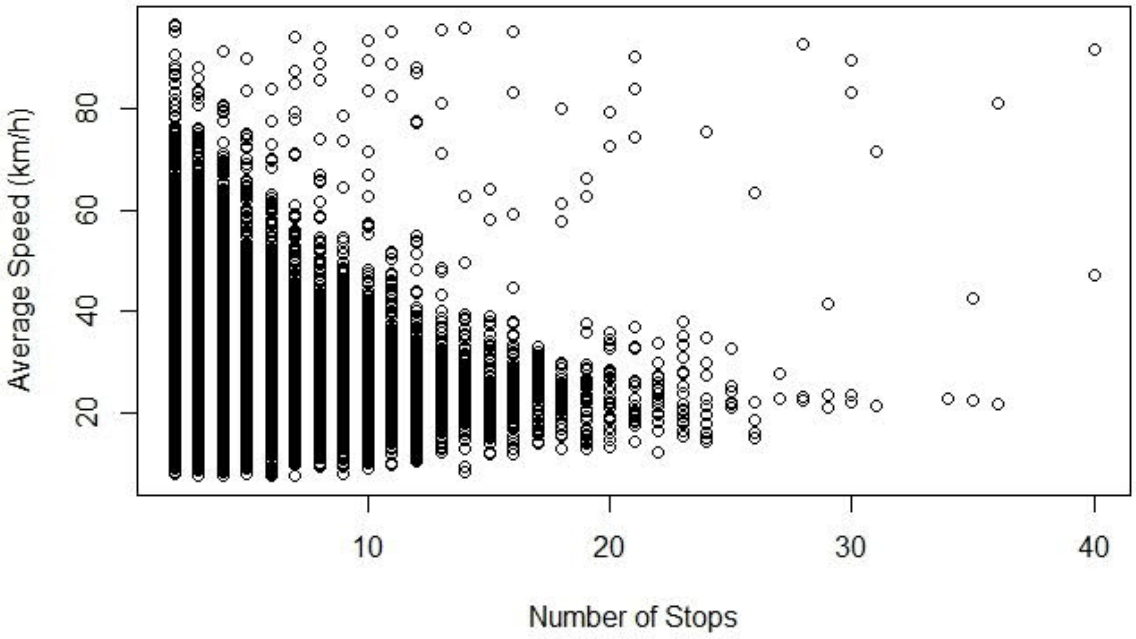


Figure 41: Number of stops and average operating speed

The relationship between stop density and speed (distance/number of stops), which can be inferred from the relationship between distance and stop density (Figure 38), is shown in Figure 42. This relationship is approximated by the function $y = 21.727 + \frac{2.964}{x}$ | ($\bar{R}^2 = 0.34$). While this relationship is still relatively weak,

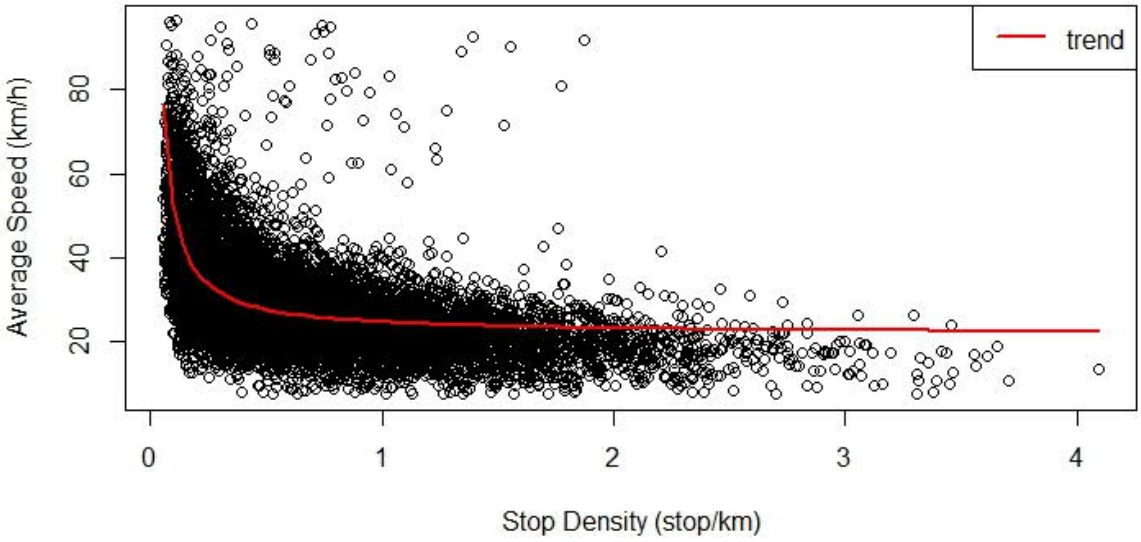


Figure 42: Stop density and average speed

6.2.6 Speed and Distance

The relationship observed between average operating speed and trip distance is shown in Figure 43 with both a linear and non-linear ($y = ln(x)$) approximation of the relationship.

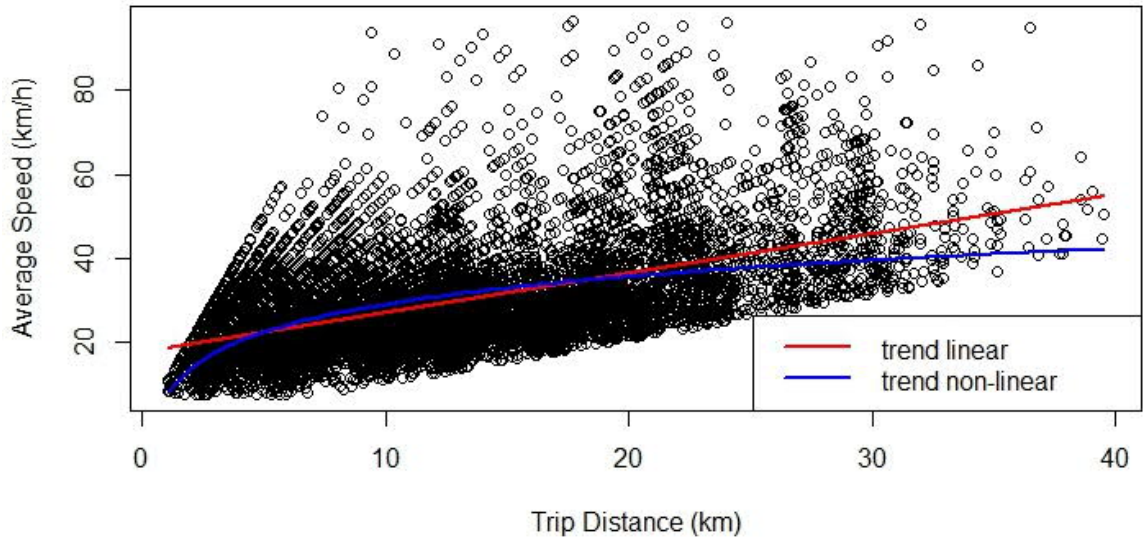


Figure 43: Relationship between trip distance and average operating speeds

The linear and non-linear relationships observed are given by $y = 0.941x + 17.77$ ($R^2 = 0.30$) and $y = 9.5537ln(x) + 7.0454$ ($R^2 = 0.26$) respectively.

While both the linear and non-linear relationships are relatively weak (low R^2 values), and cannot, as such, be used reliably as predictive models, they do give a general indication of the service characteristics and the variation therein across trips and routes.

6.3 Route Classes Observed

In section 5.2.4, the three resultant route classes determined through the clustering process were defined as:

- Feeder or distribution routes;
- Intermediate routes; and
- Trunk routes;

In order to explain the logic in the naming convention chosen, some of the characteristics of each one of these observed routes were identified and discussed in sections 5.2.4 and 5.3.1.

As summarised in section 5.2.4, the three route types have the following high-level attributes:

- Trunk routes – long distance, high speeds, low stop density – 2,111 observations (18.9% of observed trips)
- Intermediate routes – middle distance, lower speeds, medium stop density – 4,083 observations (36.6% of observed trips)
- Feeder/Distribution routes – short distances, lowest speeds, highest stop densities – 4,963 observations (44.2% of observed trips)

Further inspection and comparison of the operational characteristics of each one of these route classes demonstrates the differences and similarities in terms of the distributions and relationships previously discussed.

6.3.1 Detailed description of the intra-route operational attribute distributions

Whereas the mean and median values of the key operational metrics were discussed in section 5.2.4, the distributions of these values are provided below.

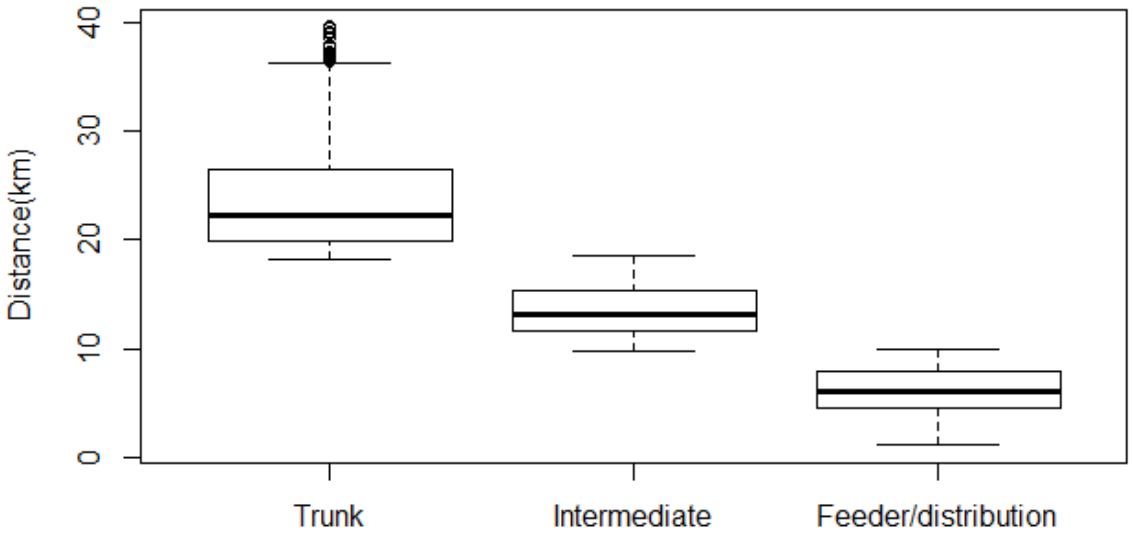


Figure 44: Boxplot of trip distances within each route class

Figure 44 shows that the trunk routes are longer distance routes between 40 km and 20 km, intermediate refers to the middle-distance routes between 20 km and 10km, while feeder/distribution refers to the routes shorter than 10km in length. As distance was one of the clustering variables, one would expect a difference in distance distributions between the three classes. The clear segregation between the three classes, each class effectively having a discreet range of distance values confirms the importance of distance as a classifying variable.

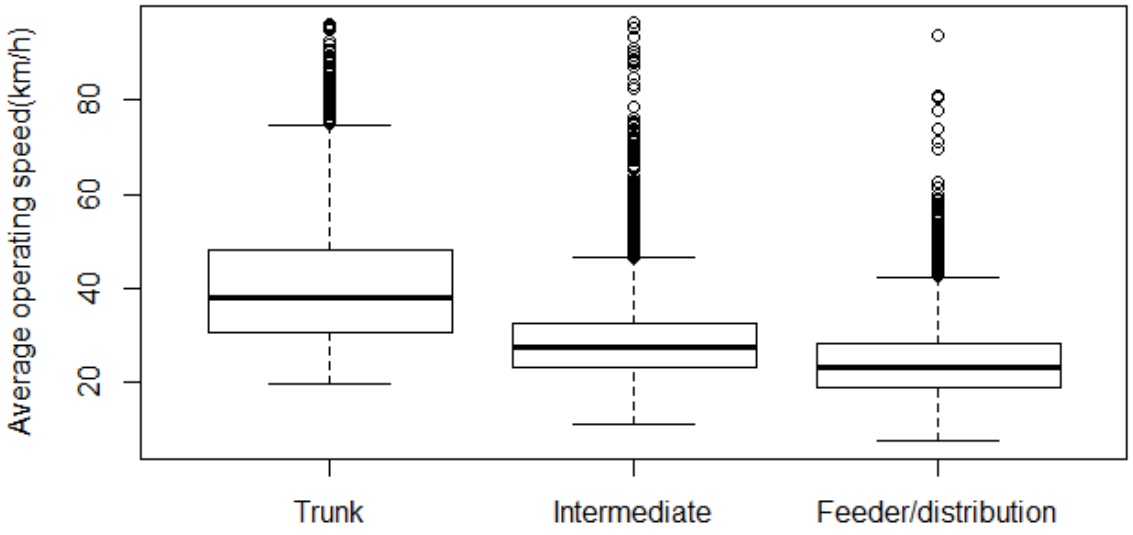


Figure 45: Boxplot of average operating speeds within each route class

Figure 45 confirms that the distances are directly related to the average operating speeds, that the higher distance trips typically operate at higher speeds. It also shows that for all three clusters that there are a relatively high number of outliers in the high-end of vehicle operating speeds – about 3% of the records, after carrying out the cleansing procedures, have average speeds in excess of 60 km/h and about 1% have speeds higher than 70 km/h. While these error percentages are in themselves not problematic, it should be borne in mind that these speeds were arbitrarily selected in order to gauge whether the speeds recorded are sensible. For the purpose of this study, however, the presence of these errors was considered acceptable as they do not significantly influence any of the findings.

Figure 46 shows the longer distance and higher speed trips are typically associated with the longer duration trips in terms of origin to destination travel time. It can be seen that the feeder/distribution class has a number of outlying trips in terms of high travel times. This could possibly be explained by roaming vehicles searching for passengers in low demand situations.

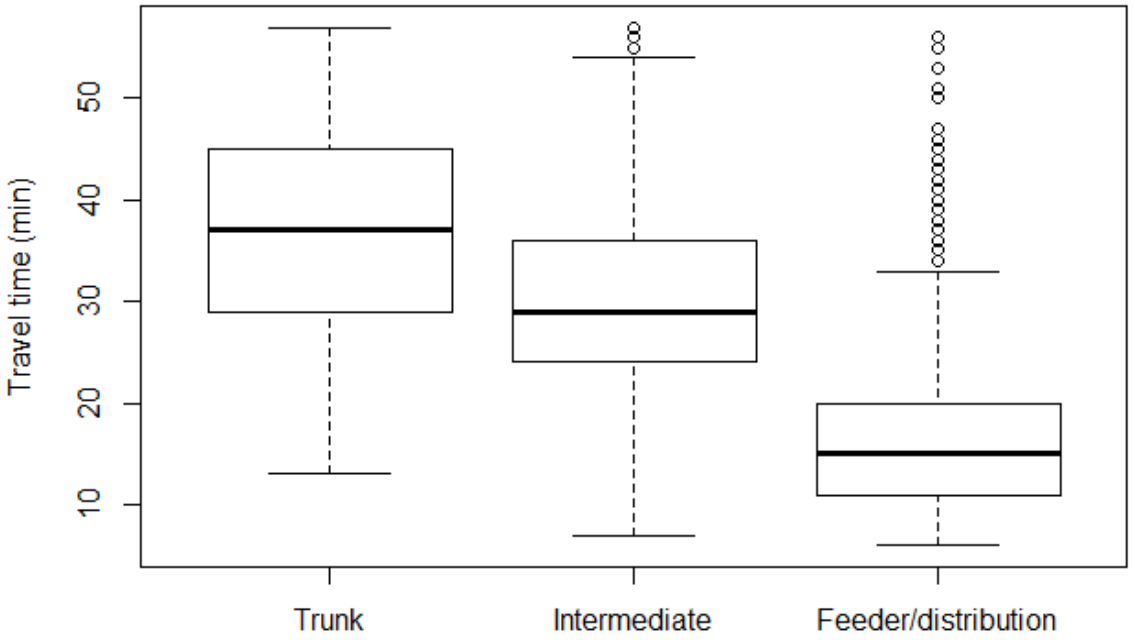


Figure 46: Boxplot of average operating travel times within each route class

Figure 47 shows that despite the clear differences in travel time, speed and distances between the three classes, all three groups typically make a similar number of stops. The median value for number of stops made by trips belonging to the three classes was found to be 6, 7 and 5 respectively.

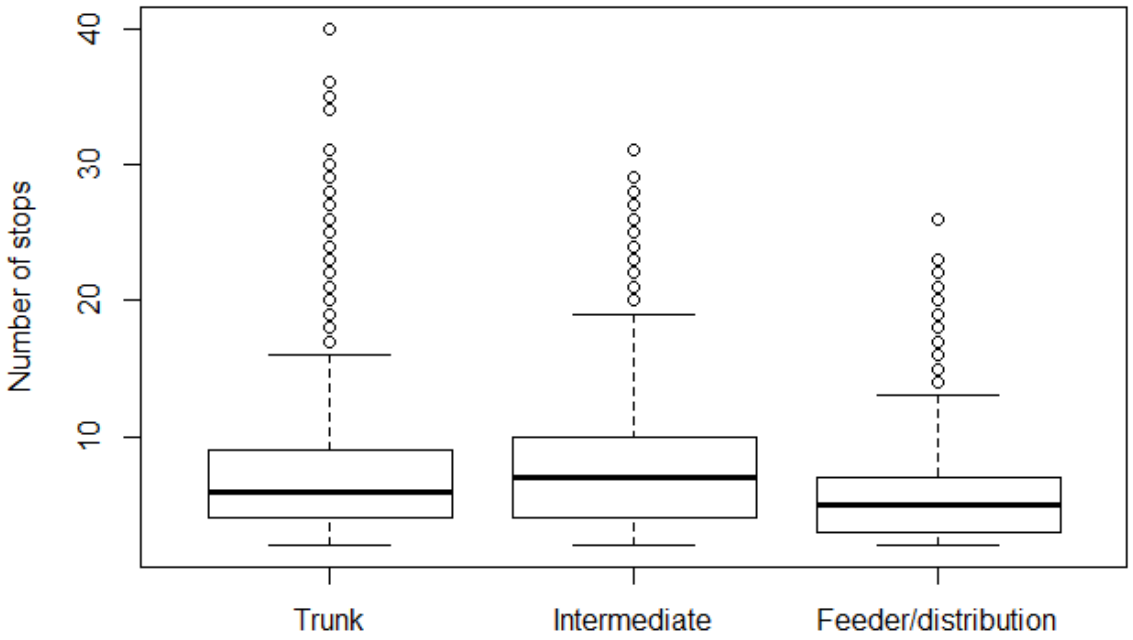


Figure 47: Boxplot of number of stops per trip within each route class

Figure 48 shows that although trips belonging to all three classes typically carried 15 passengers (capacity of most of the vehicles) per trip, trips in the feeder/distribution class have a large proportion of trips transporting fewer than 15 passengers per trip. All three classes, but especially trunk and intermediate, have a small number of outlying values in the high and low end.

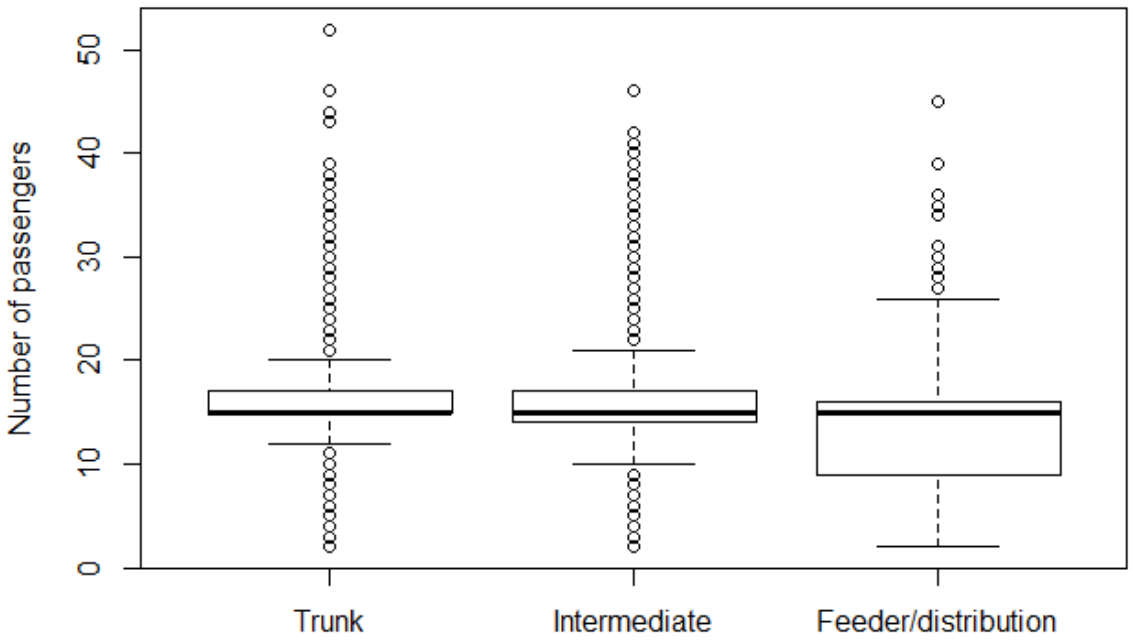


Figure 48: Boxplot of the number of passengers per trip within each route class

Figure 49 shows that for all three route classes, the vast majority of trips carried between 10 and 15 passengers but that the feeder/distribution class has a higher proportion of the low passenger number trips than the other two route classes.

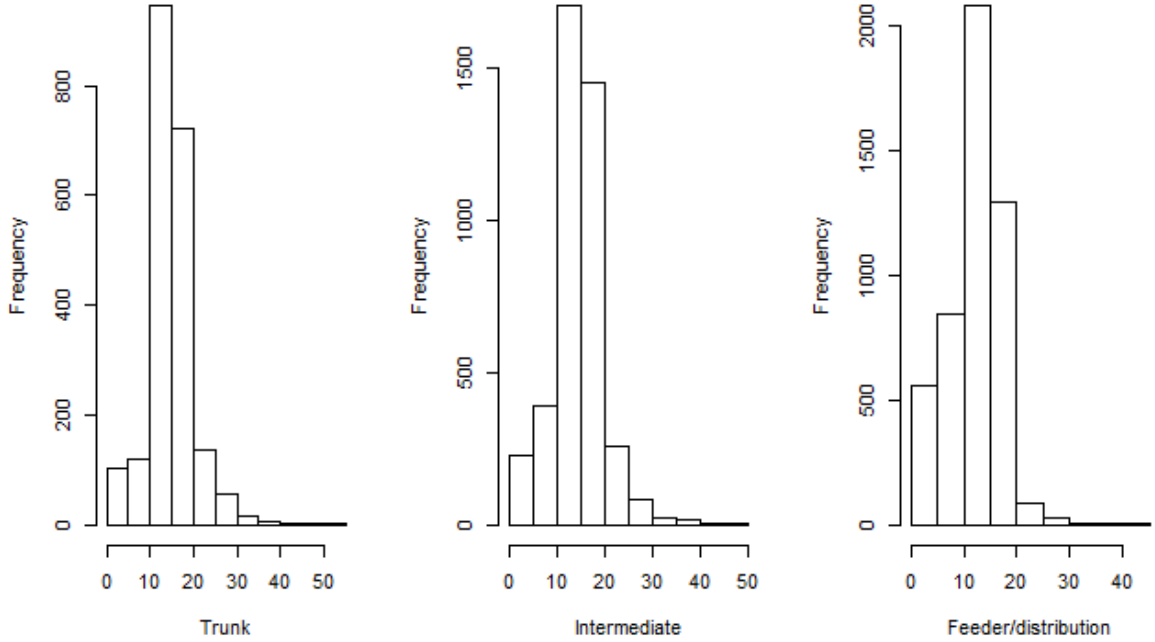


Figure 49: Histograms of passenger numbers per trip per route class

Figure 38 indicated that an inversely proportional relationship between distance and stop density exists. Figure 50 and Figure 51 provide an overview of this relationship within each of the three classes.

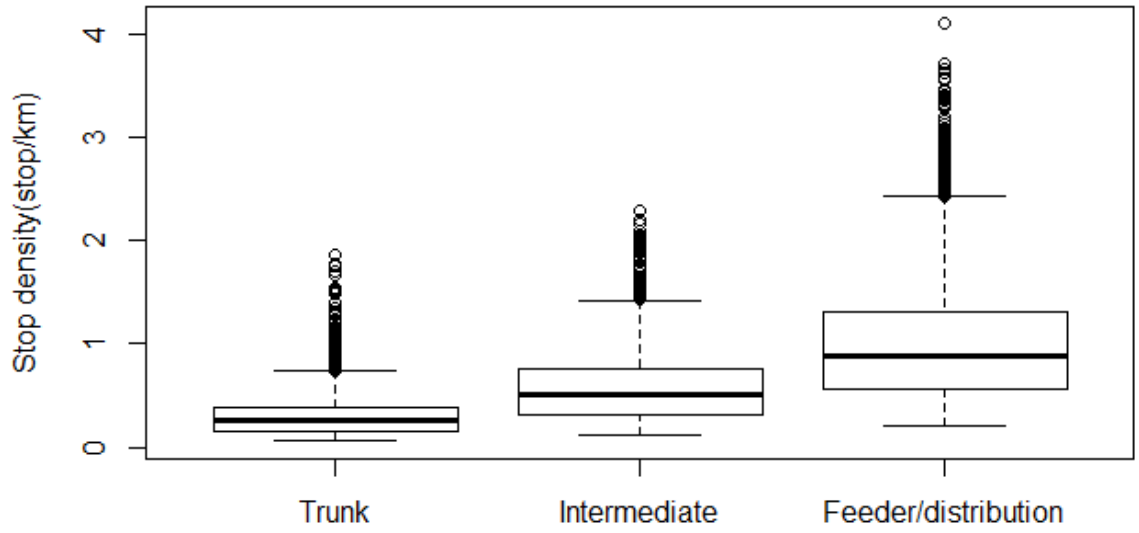


Figure 50: Boxplot of stop density within each route class

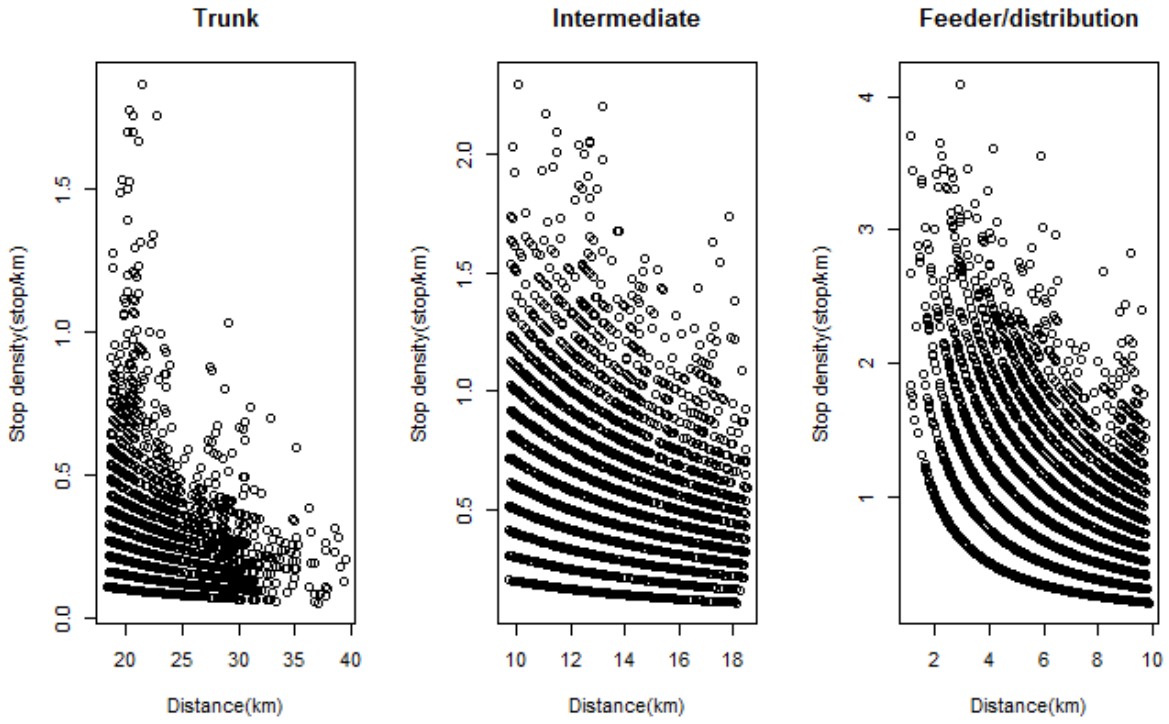


Figure 51: Relationship between trip distance and stop density per route class

Figure 52 shows that regardless of the route class, passenger turnovers are typically between two and four passengers per stop. Since for a large proportion (53%) of trips 15 passengers boarded at the 1st stop, the passenger turnover after the 1st stop is an interesting indicator of the activity along the route between the trip's origin and destination.

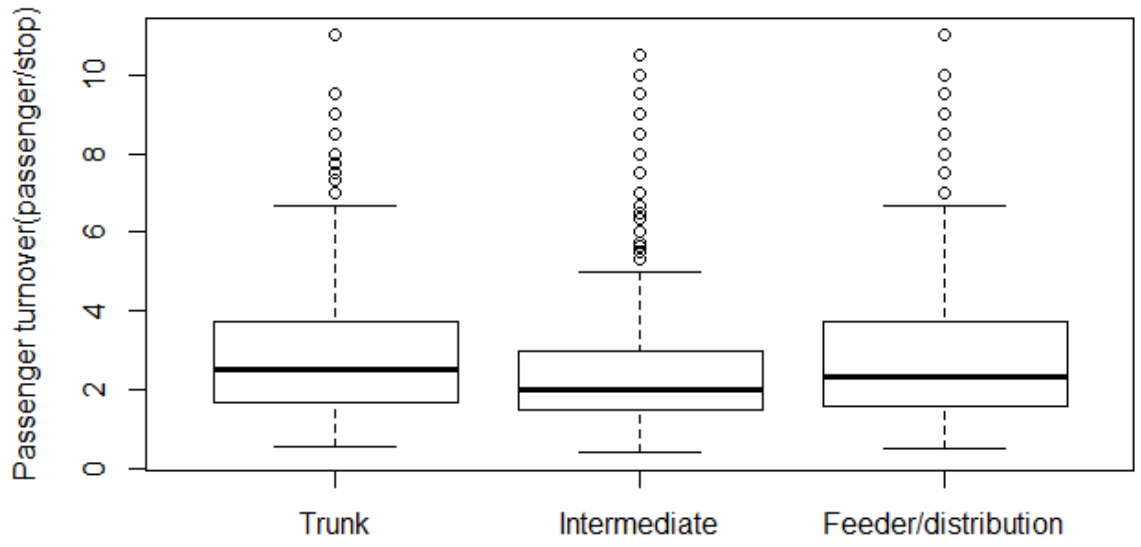


Figure 52: Boxplot of passenger turnover within each route class

Figure 53 shows, however, that when the passengers boarding at the 1st stop are not taken into account in the passenger turnover equation, the values diminish to between zero and two with the median values being very similar – 0.54, 0.6 and 0.67 for the three classes respectively.

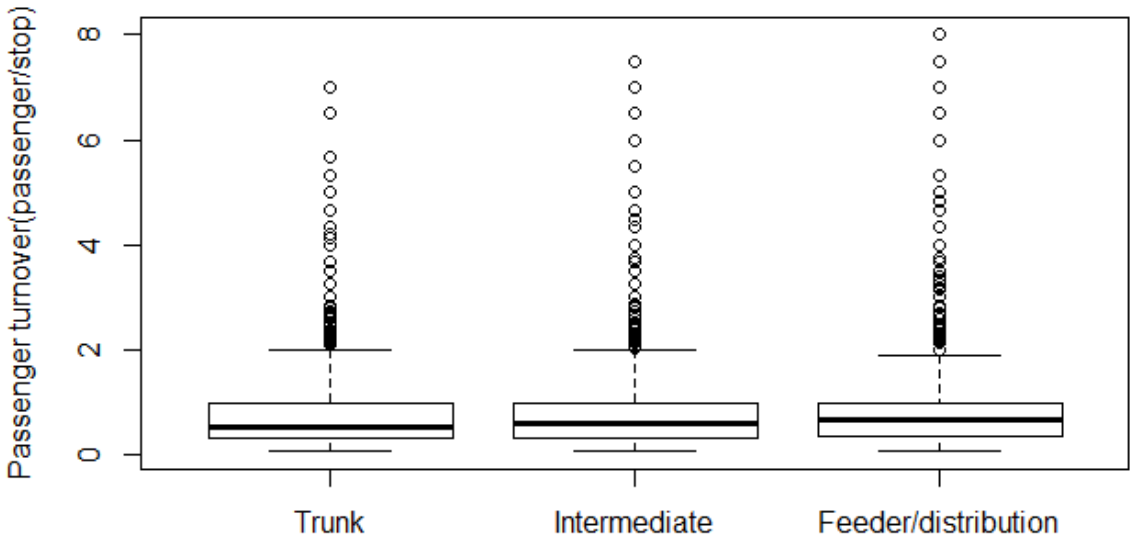


Figure 53: Boxplots of passenger turnover within each route class disregarding the 1st stop and its boardings

This indicates that for the trunk routes, about half of the stops made along the way are to pick up passengers while the rest are for dropping off passengers. For the feeder or distribution type routes, this proportion changes to about two thirds of the stops being made to drop off passengers.

6.3.2 Temporal variation of the route classes

At an aggregate level, there is very little difference, as shown in Figure 54, Figure 55 and Figure 56, in terms of the operational characteristics observed per trip for the different commuter peak periods, i.e. morning, inter-peak and afternoon.

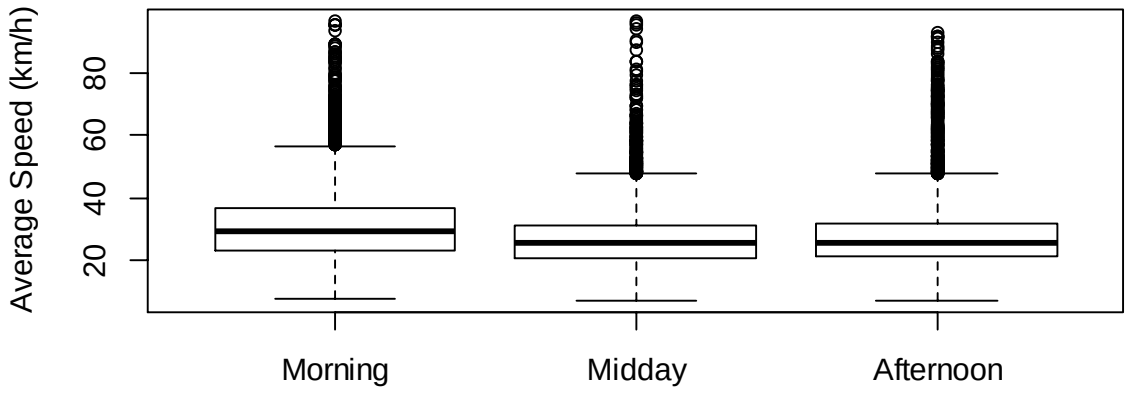


Figure 54: Boxplot of average operating speeds for the different periods of the day

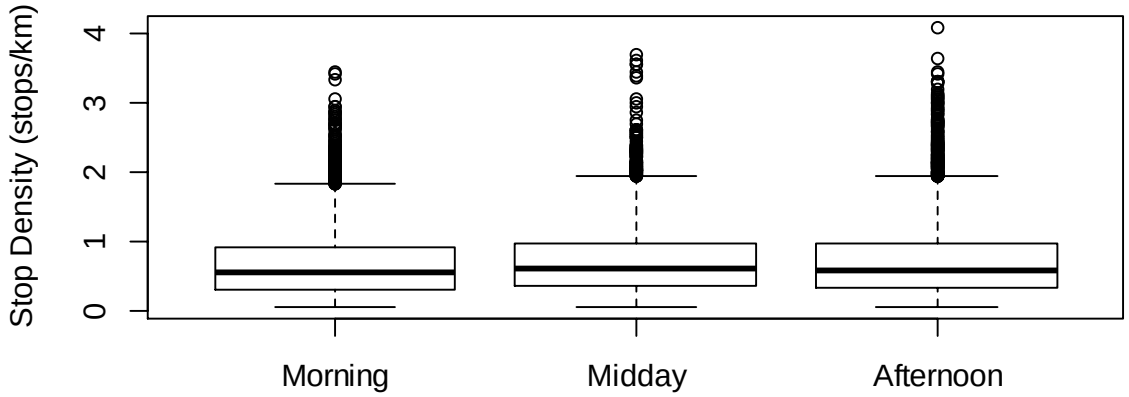


Figure 55: Boxplot of stop densities for the different periods of the day

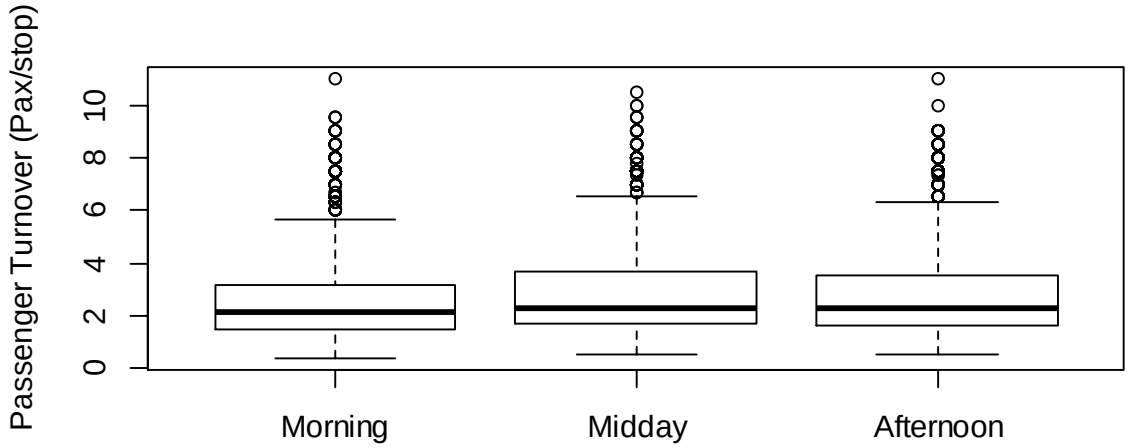


Figure 56: Boxplot of passenger turnover for the different periods of the day

It was, however, not known whether or not there is a difference in terms of the service type per route for the different peak periods, i.e. that a route operates as a feeder or trunk during the morning and operates as an intermediate type during the inter-peak or afternoon.

The clustering procedures were, therefore, repeated for the three time period subsets of the full dataset, sub-divided into trips that were completed within the three respective peak periods.

A comparison can then be made to the class assigned to each route for the full dataset and each one of the recorded periods to see which routes are reassigned when doing the temporally differentiated clustering.

The comparison of the service type classes assigned to each route for the three periods of the day can be summarised as follows:

- For 67% of the routes surveyed, all peak periods were assigned to the same class;
- For 31.7% of the routes surveyed, two of the three peak periods were assigned to the same class;
- For 1.3% of the routes, none of the three peak periods were assigned to the same class; and
- Of the routes for which not all three peak periods were assigned to the same class, 63% of these routes were assigned to two adjacent classes (i.e. Trunk or Intermediate or Intermediate and Feeder/Distribution). This comprises 21% of all routes surveyed. The remainder, 37% (12% of all routes) were assigned non-adjacent classes i.e. Trunk and Feeder/Distribution.

These values, and Figure 54, Figure 55 and Figure 56, confirm that, given the observed trips and routes, the majority of routes operate similar services during the morning, middle of the day and afternoon. Spatially representing the routes that do not exhibit this phenomenon does not show anything meaningful as, while it is a relatively small proportion of the routes, the overlap and coverage of routes is extensive; giving the appearance that this comprises all routes surveyed.

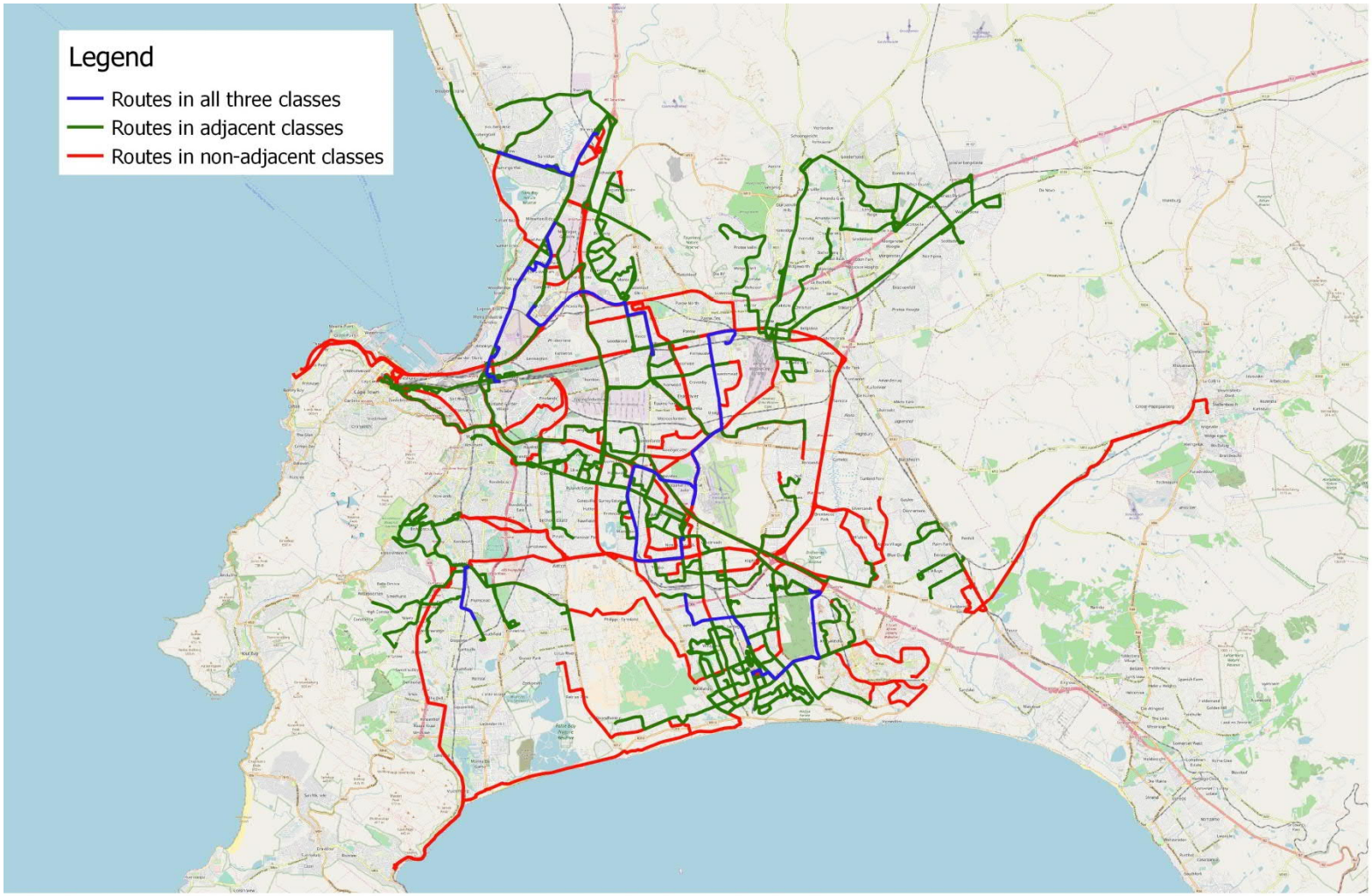


Figure 57: Temporally differentiated route class assignment

6.4 Further Spatial Analysis of Route Classes and Public Transport Interface

One of the objectives of this research is to determine whether the information extracted from these data could be useful in planning for the interface between scheduled and planned public transport services and the MBT services or designing hybrid services.

Hybrid public transport systems, integrating public transport and paratransit services, has been investigated and discussed by others. Golub, Behrens and Ferro (2012) evaluate different approaches to service integration and discuss examples of these arrangements in South Africa as well as internationally. Ferro, Behrens and Wilkinson (2013) argue that where the objective of many developing cities' public transport planning is to totally replace the 'informal services' with planned services, this total replacement would in most cases not be successful and a policy of integration should rather be pursued.

The trunk-planned and paratransit-feeder approach, identified by Golub *et al.* (2012) as the most suited approach to successful long-term integration, capitalises on the strengths of higher capacity vehicles (buses) offering higher speed services on the longer distance trunk routes and the economic advantage offered by the high coverage potential of smaller, more agile vehicles that typically operate paratransit services (Golub *et al.*, 2012).

In Cape Town, to some extent, the interaction of bus, rail and MBT services do operate as an integrated network where each mode, depending on area and corridors, occupy a different level in the network hierarchy.

Figure 58 shows an extreme level of service duplication and overlap across the various public transport modes (including paratransit) offered in Cape Town. Within this overlapping network, there are sub-networks that operate with hierarchical mode structure combinations, i.e. MBT – rail, bus – rail, MBT- bus, MBT – bus – rail.

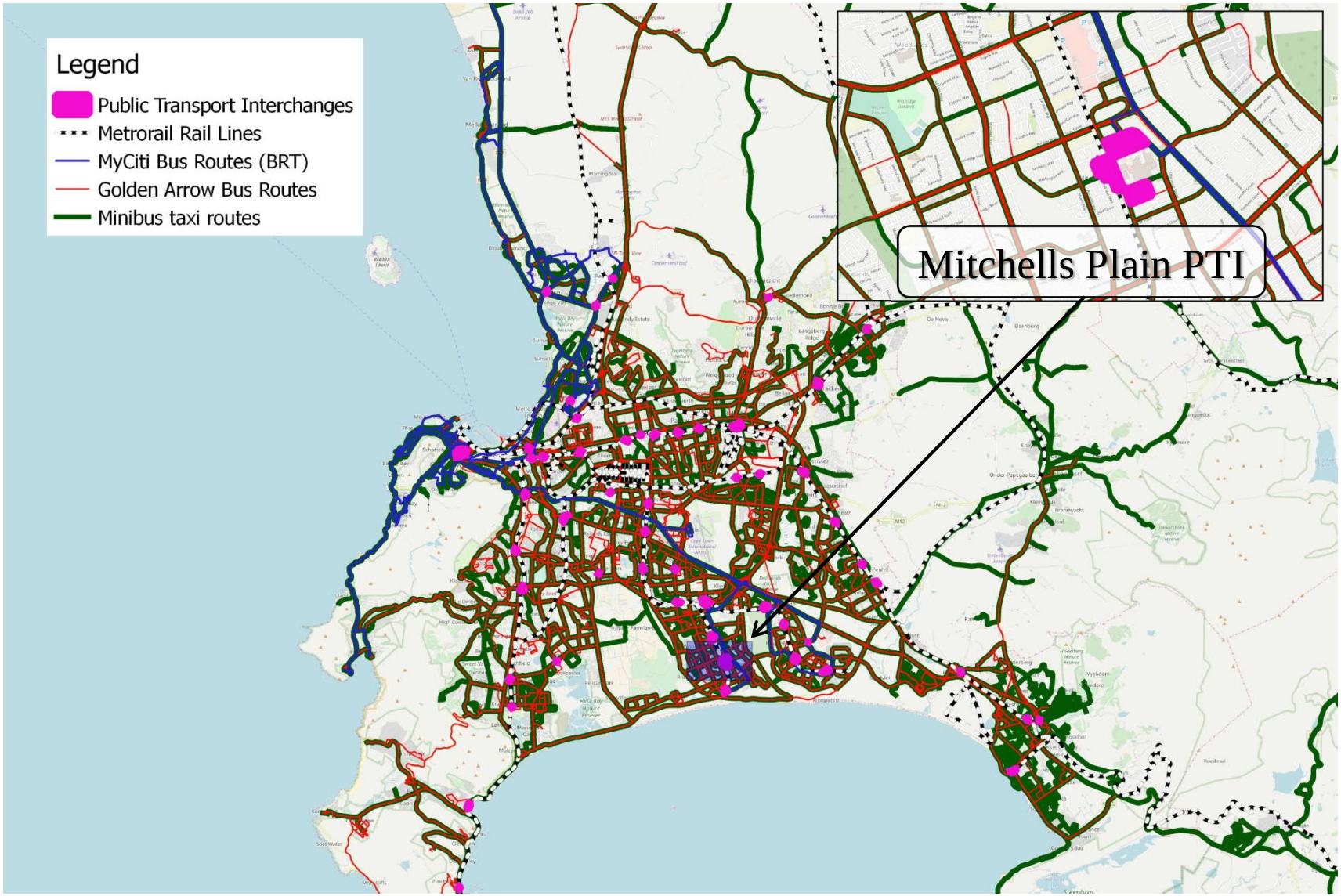


Figure 58: Public transport routes and interchanges in Cape Town

The South African Public Transport Strategy released March 2007 (Public Transport Strategy, 2007) documents the countries long-term vision of having "85% of the metropolitan city areas' population within 1km of a public transport feeder route". The implementation of Integrated Rapid Public Transport Networks, where the all modes operate seamlessly is given priority in the Public Transport Strategy with MBT and non-motorised modes serving as feeder networks.

Of the routes surveyed in this study, approximately 48% were classified as feeder or distribution type routes. Recognising where these services interact with planned or existing high capacity trunk bus or rail routes, and understanding their operational characteristics as part of the wider network is an essential element in planning for the effective service integration.

Public transport interchanges (PTIs) are key locations where multiple mode passenger transfers occur and are focal points of development in IRPTNs. Based on 2013 rank survey data, the 10 PTIs with largest MBT passenger volumes in Cape Town are as follows:

- Mitchells Plain Town Centre PTI;
- Khayelitsha Site C PTI;
- Cape Town Station PTI;
- Bellville Station PTI;
- Wynberg Station PTI;
- Nyanga Central MBT Terminus PTI;
- Retreat Station PTI;
- Parow Station PTI;
- Mfuleni PTI; and
- Du Noon PTI.

More recent, but incomplete (does not include Cape Town Station or Nyanga Station), data from 2017 surveys of rail transfer passenger activity per mode at PTIs indicate that Wynberg exceeds Bellville in terms of passenger activity and that Claremont replaces Du Noon in the top 10 MBT passenger activity PTI list. For the purposes of this study, the former of these two data sources was used in order to include Du Noon in the visualisation shown in Figure 59.

All of these PTIs service at least two modes with some serving three or four modes (BRT and bus considered separate modes for purposes of service differentiation). The Cape Town and Mitchells Plain PTIs, for example, serve MBT, Golden Arrow Buses, Metrorail and MyCiti Bus BRT.

Identifying the MBT routes that serve these areas, feeders and trunk MBT routes, and understanding their respective operational characteristics will be useful in prioritising service improvements and infrastructure planning to bolster the integration of PTI intermodality.

Of the routes that were classes as classified as feeder or distribution type routes and are connected to the 10 top busiest PTIs (in terms of MBT passenger activity) are shown in Figure 59 together with the Metrorail and MyCiti (trunk and feeder) lines.

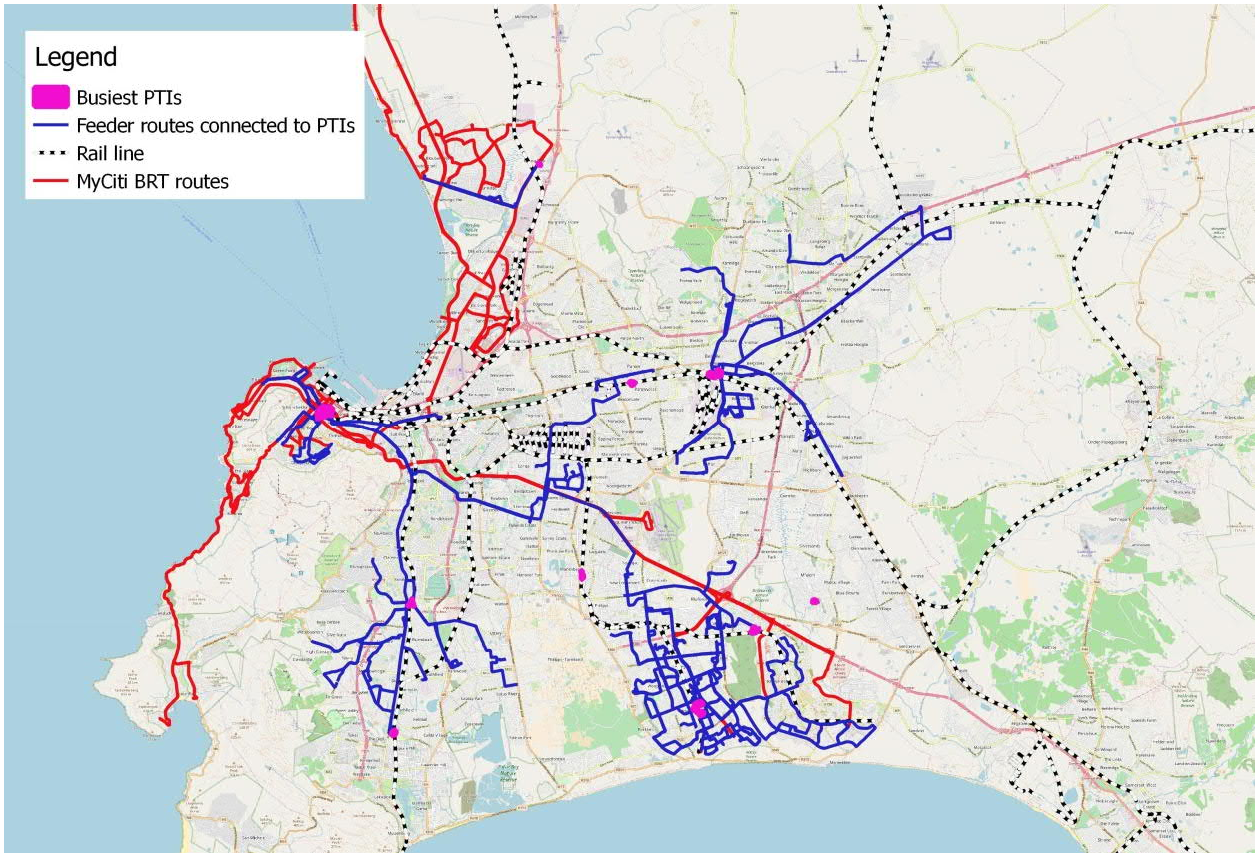


Figure 59: Top 10 busiest MBT activity PTIs and MBT feeder routes

Of all the routes that were surveyed, 56% of the routes either originate or terminate at one of the 10 PTIs listed above. In terms of the three classes, 74%, 63% and 46% of the trunk, intermediate and feeder/distribution routes are connected to these PTIs respectively.

Only the MBT feeder/distribution routes are shown in Figure 59 in order to demonstrate the modal network hierarchy exhibited at these PTIS for the route combinations. While the MBT services connecting to these PTIs do not only offer feeder type services, the author is of the opinion that by improving the operations of these services and providing infrastructure for efficient mode transfer facilities, the patronage of the trunk services will be greatly uplifted.

CHAPTER 7

7 Conclusion

7.1 Objectives

While the City of Cape Town's objective in collecting onboard MBT data, as per the advertised tender document, *Tender No: ITS 001/2016/2017 To Undertake Onboard Surveys on all Minibus Taxi Routes in the City of Cape Town* (CoCT, 2016), was to "clarify the actual extent of MBT services within the City", the objectives of this research are more wide-ranging and generic.

The core objectives of this research can be encapsulated into one overarching objective statement:

> To demonstrate how unscheduled (or informal) paratransit services can be better understood through onboard data collection and route mapping and how this knowledge can be used to, not only understand the individual services but their characteristics in relation to each other, as well as their role in the greater integrated public transport network.

7.2 Outcomes

As detailed in the literature review, there are various examples of research related to the gathering of data, especially spatial data, on paratransit services in urban areas in the developing world.

In many of the earlier examples of these studies, i.e. Ching *et al.* 2013, Klopp *et al*., 2014, Klopp *et al*., 2015 and Klopp *et al*., 2017, the aim was to gather spatial data and to develop maps of the paratransit networks, providing both the users, operators and authorities with a consolidated spatial dataset to better understand the network and the service offerings.

These studies paved the way for more comprehensive data collection initiatives, i.e. Saddier *et al.*, 2016, Saddier *et al.*, 2017, Saddier and Johnson, 2017 and Coetzee, Mulla and Oosthuizen (2018), that focussed more on collecting operational information in conjunction with spatial and temporal information element. These studies demonstrate the value that detailed operational information of these poorly understood services have for planning and operational authorities.

This research aims to build on these previous studies and to demonstrate how these data collection efforts can be enriched in terms of the information that can be collected and analysed at little or no extra cost.

By collecting the actual alignment of MBT trips and routes, the location and number of passengers boarding and alighting and the fare amount per trip, the operational information such as average speed, travel time, individual trip and route distances, stop frequencies and densities, load factors and operational revenue can be determined.

This operational information, calculated from the raw data collected per trip and aggregated on the route level, was used to determine average operational indicators for MBT services

and to classify the services in terms of the differences in their operational characteristics into different service types.

To demonstrate this capability, the MBT trips were clustered into three distinct service classes, namely, Trunk, Intermediate and Feeder/Distribution. The purpose of this exercise was to (1) demonstrate that these differences exist, (2) that this process can be executed with relative ease using onboard paratransit data and (3), that there is significant value to be garnered from understanding the services at this level of detail.

7.3 Results

As mentioned in the beginning of this study, one of the limitations of the research is that the sampling method implemented aimed to rather be cost-effective in terms of collecting data that provided the highest level of spatial service coverage, than to be statistically representative of the individual routes surveyed.

Nonetheless, the methods of analysing these data, and results thereof, may be applied to similarly structured data collection studies and, by employing more robust sampling techniques, yield information that is statistically representative and can be confidently used to describe the operational attributes of individual routes, operating associations or service types.

Through cleansing, visualising, analysing and clustering the onboard MBT data that was collected for 11,157 trips on 278 routes, it was found that there are distinct service types or classes within the MBT network in Cape Town (see Figure 30, Figure 31, Figure 32 & Figure 33).

The highlights of operational performance analysis per route class identified, as given in Table 9 (Section 5.2.4) and Figure 44, Figure 45 and Figure 55 (Section 6.3) are given below:

- The “Trunk routes” operate longer distance services ($\bar{x}$ = 22.98km) operate at higher speeds ($\bar{x}$ = 41.15km/h) and make the fewest number of stops of all the identified classes ($\bar{x}$ = 6.8 stops) at the lowest stop density ($\bar{x}$ = 0.25 stops/km);
- The “Intermediate routes” operate shorter distance services ($\bar{x}$ = 13.39km) than the “Trunk routes”, at lower speeds ($\bar{x}$ = 28.95km/h) and stop more frequently (higher stop density, $\bar{x}$ = 0.58 stops/km); and
- “Feeder/distribution routes” operate the shortest distance trips ($\bar{x}$ = 6km), at the lowest speeds ($\bar{x}$ = 24.25km/h) and make more frequent stops than the other two route types ($\bar{x}$ = 0.99 stops/km).

The clear difference in the distribution of these values per route class shows the type of services offered. Given the demand responsive nature of MBT services, however, passenger related indicators, total number of passengers transported, passenger turnover and load factors do not significantly differ between these route class types.

Linking the services to the busiest (in terms of MBT passenger volumes) public transport interchanges (PTIs) shows that 74%, 63% and 46% of trunk, intermediate and feeder/distribution type routes either terminate or originate at these PTIs, demonstrating the importance that MBTs play in the intermodality of the envisaged IRPTNs.

7.4 Summary

Despite the fact that each route was surveyed a limited number of times and that the data analysed in this study does not comprise the full dataset collected during the execution of the onboard survey project, this dataset is most likely one of the largest onboard paratransit survey datasets of its kind ever compiled and analysed in this manner.

The depth of information, much of which wasn't touched on in this study, extractable from this limited dataset alludes to the potential of the detailed analysis of large datasets collected using statistically sound sampling techniques.

The analysis presented in this research provides answers to the key questions posed by; (1) providing descriptive overviews of the means and distributions of MBT route operational attributes such as the spacing of stops and the operational speeds, (2) the identification of metrics that are central to the understanding of this type of passenger transport service such as passenger turnover and stop density, (3) showing that MBT services operate a range of service types and that their routes can be classified according to their operational attributes, and (4) providing of how the understanding of these classified services and their relationship with other modes may be harnessed in planning and policy making for the betterment of public transport in cities where they operate.

This study provides not only an insight into the operational attributes of the different classes of MBT services observable but, more importantly, provides a framework of evidence of how these data can be analysed, grouped, interpreted and applied to the planning of multi modal or hybrid public-trunk and paratransit-feeder networks.

CHAPTER 8

8 The Way Forward

8.1 Improved Sampling and Statistical Analysis

As mentioned in the list of limitations of this study, the sampling techniques employed in the data collection were not specifically designed to be statistically representative of the operations of each route surveyed but rather with the idea of surveying each route, back and forth, at least a minimum of three times per peak period and twice during the inter-peak period on two regular weekdays, a Friday and a Saturday.

It is recommended that detailed pre-survey planning be carried out or to use other sources of information to determine the level of operations on the route level. The key here is to determine the number of vehicles operated by each operator association on specific routes and to determine on the route level how the operations differ during the different times of the day and between weekdays and weekends.

Having knowledge about these intricacies allows for efficient planning of the surveys for each route and how many resources should be allocated to the data collection on each route.

It is then further recommended that the robust sampling is carried out on the route level and that a representative sample of vehicles on each route should be surveyed for the full duration of a number of days. This level of information for each route, for the inbound and outbound directions, will enrich and optimise the data cleansing processes and provide data that can be analysed, interpreted and incorporated into planning and policy making with confidence.

8.2 Expanding the Onboard Data with Static Rank Surveys

Given that improved sampling techniques are applied to the onboard survey method, it would be beneficial to carry out static, rank-based departure and arrival surveys, with passenger volume counts, over the same or a similar period as the onboard surveys. Marrying the onboard and rank surveys, as detailed by Saddier and Johnson (2018), capitalises on the strength of both survey methods by taking into account the passenger boarding and alighting activities and route adherence or deviation, for a sample of vehicles while capturing the headways, waiting times and passenger load factors of all vehicles departing from and arriving at the ranks.

8.3 Peak Commuter Direction Differentiation

While the direction of the route was taken into account by appending an R as a suffix to the route ID, as detailed in Section 1.6.3, the direction of peak commuter travel per route pair was not taken into account in the analysis of the data. More detailed analysis of this level of differentiation per route is expected to provide additional insight into the temporal and spatial MBT operations and passenger demands.

8.4 Focus on the Passenger Origin – Destination Information Collected

It was mentioned in Section 3.2.2 study that for each passenger boarding the vehicle, the following information was collected:

- The time and location they boarded the vehicle – all passengers boarding at the first stop (origin rank) were assigned the same timestamp and location coordinates as the trip data was only recorded once the vehicle pulled away from the origin rank;
- Age group, gender and ethnicity - subjective associative placeholders to assist the enumerators in identifying individual passengers. This is bolstered by seating positions in the survey app interface whereby the enumerator can shift passengers around in the vehicle if they changed seats en route.

While this information could be very useful in demand studies and for developing or estimating OD trip matrices for the mode, the information collected is largely subjective and requires extensive training to ensure that enumerators are using the app's function correctly and that the collected data is reliable. The risks related to the subjective element thereof should, nonetheless, be taken into account regardless of the level of training provided.

8.5 Route Adherence Studies

If the sampling of vehicles to be surveyed is carried out at the route level and each vehicle is surveyed for its full day's operations over a number of days, the path taken on each route can be traced and compared to the licensed route description to determine what level of route adherence is common.

In the project, from which this study's data was sourced, the objective, as detailed previously, was to collected data on each route for a specific number of trips. As the project was still ongoing at the time this research was carried out, the data collection for certain routes was not complete. For routes for which a "complete" or near complete set of trips were collected, it is possible to observe the route adherence mentioned above.

Figure 34 in Section 6.1.1 shows the different trip paths taken by vehicles between Lower Crossroads and Claremont with the designated route as per the Operating Licence (OL) shown in blue.

Services that exhibit poor adherence to their designated or licensed route paths can be identified through the analyses of the respective route's trip data by calculating the Coefficient of Variation (COV) of their trip distances. Routes that exhibit high COV for recorded trip distances can be examined spatially, as was carried out for the route/trips shown in Figure 34, to visualise the deviation from the designated route.

If a sufficient number of trips are recorded for different times of the day, analysing the temporal component of route adherence may reveal that deviations are more common or pronounced during the low demand periods. The opposite may, however, also be true in an effort by drivers or owners to reduce operational costs. Analysing the route adherence for the inbound as well as the outbound direction for a route may reveal whether the routes taken differ for the respective directions.

8.6 Optimum Hybrid Route Design

If the aforementioned improved sampling techniques are implemented, the passenger origin and destination data collected per route and route adherence analyses per route may contribute to a method of determining the most frequent and perhaps optimum routes connecting two locations. Combining this information with static rank surveys and applying

the processes to routes serving PTIs may be a good departure point in identifying which routes should be promoted and prioritised in terms of infrastructure improvements and funding allocations.

Since time and coordinates for each stop is recorded for every trip, it is possible to decompose the operating speed for each trip segment. Determining the speed on individual segments for different periods of the day and aggregated over a number of observations will allow for the identification of route or road sections where investment is needed most.

8.7 Policy Implications & Recommendations

Given the previously unattainable levels of insights into the operations of individual routes, the MBT network as a whole and the interaction with other public transport modes, it could be beneficial to review the policies affecting the planning and regulation of the MBT and public transport industries.

South Africa's long-term public transport vision includes placing "85 percent of a metropolitan city's population within 1km of an integrated Rapid Public Transport Network trunk or feeder corridor" (DoT, 2007b). In order to achieve this, a modal integration approach where non-motorised and other modes, including MBTs, will need to be implemented. The relatively low population densities in South Africa does not lean itself towards the successful operation of larger (30 to 40-seater) vehicles on the feeder and distribution network routes. The flexibility and higher frequency possibilities offered by smaller vehicles, i.e. MBTs, which can operate both scheduled and demand responsive type services offer a good opportunity for addressing this.

While the MBT mode has already been identified in the National Land Transport Transition Act (NLLTA) (DoT, 2000), the National Land Transport Strategic Framework (NLTSF) (DoT, 2007b) and the 2017 NLTSF (DoT, 2017) as a key mode of public transport, the policies relating to the MBT operating licences and Transport Register data collection requirements may require some revision in light of the findings of this research.

Operating licensing strategy:

- Change the operating licensing strategy pertaining to minibus taxis so that they are allowed to operate within a specific area, as they are already doing so to some extent (see Figure 34). Whereas previous policy documentation proposed that route-based operating licences be issued to all operators, reverting to area-based permits or licences but with a higher level of oversight and control (Operating Licence Administration System below) may be more appropriate and efficient in certain cases.
- Require all vehicles to collect data on their operations using technology such as was used in this study upon renewing of their operating licences.
- Optimise the area based 'routes', by analysing the continually collected data, in order to connect the maximum number of feeder and/or distribution routes to PTIs. New operating licences could therefore be approved and issued on this basis.

Transport Register (previously CPTR) data requirements:

- Data related to the operations and route coverage of the MBT industry should be collected in a manner similar to the data collected for this research but with more robust sampling techniques.
- The data collected should be supplemented with static surveys to expand the onboard samples with rank-side determined arrivals and departures – as described in Section 8.2.
- The collected data, onboard and static, should be stored digitally as in the City of Cape Town's Transport Reporting System (TRS)
- Establish Operating Licence Administration Systems within each transport authority's jurisdictional area. This system could be an interactive map indicating the location and extent of each route and containing all information as collected by vehicles operating the designated routes – similar to the existing TRS but with additional functionality.

9 References:

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach (Vol. 904). John Wiley & Sons.

Behrens, R., McCormick, D. & Mfinanga, D. (2016). Abingdon, Oxon, UK. Routledge.

Black, A. (1995) *Urban Mass Transportation Planning*. New York: McGraw – Hill, pp.178 - 179.

Cairo, O., Sendra Salcedo, J., & Gutierrez-Garcia, J.O. (2015). Crowdsourcing information for knowledge-based design of routes for unscheduled public transport trips. Journal of Knowledge Management, 19(3), 626-640.

Cervero, R & Golub, A. (2007) Informal transport: A global perspective, Transport Policy, Vol. 14, No. 6, pp. 445-457.

Ching, A., Zegras, C., Kennedy, S., & Mamun, M. (2013). A user-flocksourced bus experiment in Dhaka: New data collection technique with smartphones. Transportation Research Record: Journal of the Transportation Research Board.

City of Cape Town (CoCT). (2005) Cape Town Current Public Transport Record (CPTR) 2004/5. Cape Town.

City of Cape Town. (2013) *Integrated Transport Plan for Cape Town 2013 to 2018*, City of Cape Town, Cape Town.

City of Cape Town. (2016) *Tender No.: ITS 001/2016/2017- To Undertake Onboard Surveys on all Minibus Taxi Routes in the City of Cape Town.* City of Cape Town, Cape Town.

Classify. (2018). Merriam-webster.com. Retrieved from https://www.merriam-webster.com/dictionary/classifying

Coetzee, J., Krogscheepers, C. & Spotten, J. (2018). Mapping Minibus-Taxi Operations at a Metropolitan Scale – Methodologies for Unprecedented Data Collection Using a Smartphone Application and Data Management Techniques. South African Transport Conference, 2018.

Coetzee, J., Mulla, A. & Oosthuizen, N. (2018). Tools to Assist in Determining Business Values of individual Minibus-taxi Operations in Rustenburg, Northwest, South Africa. South African Transport Conference, 2018.

Department of Transport (2000). The National Land Transport Transition Act. Cape Town, South Africa, August, 2000.

Department of Transport (2007a). Public Transport Action Plan. Phase 1 (2007–2010). Catalytic Integrated Rapid Public Transport Network Projects South Africa.

Department of Transport (2007b). Public Transport Strategy, March 2007.

Department of Transport (2017). National Land Transport Strategic Framework, February 2017.

Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning (p. 29). ACM.

Dolan, D. (2014) *Inside SA's R40bn taxi industry*, Reuters, views 9 July 2017, Retrieved from <https://www.moneyweb.co.za/archive/sas-r398bn-taxi-industry/>

Ferro, P. S., Behrens, R., & Wilkinson, P. (2013). Hybrid urban transport systems in developing countries: Portents and prospects. Research in Transportation Economics, 39(1), 121-132.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association, 97(458), 611-631.

Friedman, V. (2008, January 14) Data Visualization and Infographics. Retrieved from <https://www.smashingmagazine.com/2008/01/monday-inspiration-data-visualization-and-infographics/>

Furth, P. G., Hemily, B., Muller, T. H., & Strathman, J. G. (2006). Using archived AVL-APC data to improve transit performance and management.

Gaibe, H. (2009). An Investigation into the Methodology of Mini-bus Taxi Data Collection as part of the Current Public Transport Record: A Case Study of Stellenbosch in the Western Cape. (Master's dissertation, University of Cape Town).

Gaibe, H &Vanderschuren, M. (2010) An Investigation into the Methodology of Mini-bus Taxi Data Collection as part of the Current Public Transport Record: A Case Study of Stellenbosch in the Western Cape. Proceedings of the 29th Southern African Transport Conference. Pretoria. 2010.

Golub, A., Behrens, R., & Ferro, P. S. (2012). Planned and paratransit service integration through trunk and feeder arrangements: an international review. SATC 2012.

Gschwender, A., Munizaga, M., & Simonetti, C. (2016). Using smart card and GPS data for policy and planning: The case of Transantiago. Research in Transportation Economics, 59, 242-249.Munizaga and Palma (2012)

Hellerstein, J.M. (2008) Quantitative Data Cleaning for Large Databases. White Paper. United Nations Economic Commission for Europe, 2008.

Hintze, J. L. (2007). NCSS User's Guide-IV: Multivariate Analysis, Clustering, Meta-Analysis, Forcasting, Time Series, Operation Research and Mass Appraisal.

Howe, J. (2006). The rise of crowdsourcing. Wired magazine, 14(6), 1-4. Retreieved from <https://www.wired.com/2006/06/crowds/>

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., ... & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization, 10(4), 271-288.

Klopp, J., Mutua, J., Orwa, D., Waiganjo, P., White, A., & Williams, S. (2014). Towards a Standard for Paratransit Data: Lessons from Developing GTFS Data for Nairobi's Matatu System (No. 14-5280).

Klopp, J., Williams, S., Waiganjo, P., Orwa, D., & White, A. (2015). Leveraging cellphones for wayfinding and journey planning in semi-formal bus systems: Lessons from Digital Matatus in Nairobi. In Planning support systems and smart cities (pp. 227-241). Springer, Cham.

Klopp, J. M., & Cavoli, C. M. (2017). The Paratransit Puzzle: Minibus Mapping and Transportation Planning in Maputo and Nairobi. Taylor & Francis Ltd.

Klopp, J.M., Orwa, D., Waiganjo, P., Williams, S., & White, A. (2017). Informal 2.0: Seeing and Improving Urban Informal Practices through Digital Technologies The Digital Matatus case in Nairobi. Field Actions Science Reports. The journal of field actions, (Special Issue 16), 39-43.

Klopp, J.M., (2017). Mapping and mobilization for public transport advocacy in African cities. (University of Cape Town Open Lecture).

Moodley, G. Y., Aucamp, C. A., & Wood, R. Developing the eThekwini Operating Licences Strategy: How Useful is the CPTR Information? In Proceedings of the 24th Southern African Transport Conference (SATC 2005) (Vol. 11, p. 13).

Moovit (2016). Moovit End User License Agreement. static.moovitapp.com/userguide/license.pdf

Muwanaula, P. (2013). *Adoption of an Automated Fare Collection System: City of Tshwane Taxi Owner's Perspectives* (Master's dissertation).

Mxolisis, M. (2006) *A Critical Analysis of the Process of the Taxi Recapitalization Policy.* Master Degree in Development and Management, North West University, South Africa.

Ndibatya, I, Coetzee, J & Booysen, M.J. (2016). Proceedings of the 35th Southern African Transport Conference. Pretoria. 2016.

Rousseeuw, P. J., & Kaufman, L. (1990). Finding groups in data. Hoboken: Wiley Online Library.

Saddier, S., Patterson, Z., Johnson, A., & Chan, M. (2016). Mapping the Jitney network with smartphones in Accra, Ghana: the AccraMobile experiment. Transportation Research Record: Journal of the Transportation Research Board, (2581), 113-122.

Saddier, S., Patterson, Z., Johnson, A., & Wiseman, N. (2017). Fickle or Flexible? Assessing Paratransit Reliability with Smartphones in Accra, Ghana. Transportation Research Record: Journal of the Transportation Research Board, (2650), 9-17.

Saddier, S., & Johnson, A. (2018). Understanding the Operational Characteristics of Paratransit Services in Accra, Ghana: A Case Study (No. 18-05537).

Sánchez-Martínez, G. E., & Munizaga, M. (2016). Workshop 5 report: Harnessing big data. Research in Transportation Economics, 59, 236-241.

Sayad, S. (2010). Principle Component Analysis [PDF Slides]. Retrieved from http://chem-eng.utoronto.ca/~datamining/Presentations/PCA.pdf

Sendra, S.J. and Cairó Battistutti, O. (2014), “Unscheduled public transport intelligent navigation system”, Proceedings of the 18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, Elsevier.

Statistics South Africa. (2014). National Household Travel Survey 2013, Pretoria.

Tamblay, S., Galilea, P., & Mu~noz, J.-C. (2015). Estimation of a zonal origin-destination matrix from observed public transport trips for Santiago de Chile.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. 1st.

Toronto Transit Commission. (2000). TTC Fare Collection Study. Toronto Transit Commission.

Trans-Africa Consortium. (2008). Overview of public transport in Sub-Saharan Africa. TransAfrica project, UITP, Brussels.

Transit Cooperative Research Programme (2013). Transit Capacity and Quality of Service Manual, 3rd Edition. Transportation Research Board (TRB). Report number TCRP Report 165. Washington D.C. National Academy Press, Chapter 3.

Transportation Reporting System (TRS), <http://trslive.aspdemo.co.za/Account/LogOn?ReturnUrl=%2f>

Tukey, J.W. (1977). Exploratory data analysis. Reading, PA: Addison-Wesley.

Van Zyl, J., & Labuschagne, K. (2008, July). Attractive methods for tracking minibus taxis in South Africa for public transport regulatory purposes. In 27th Annual Southern African Transport Conference (pp. 7-11).

Vuchic, V. (2005). *Urban Transit: Operations, Planning and Economics.* Hoboken, New Jersey: John Wiley, pp. 445 - 448; 469 - 470.

Williams, S., White, A., Waiganjo, P., Orwa, D., Klopp, J., 2015. The digital matatu project: using cell phones to create an open source data for Nairobi’s semi-formal bus system. Journal of Transport Geography. 49.

10 APPENDICES

APPENDIX A – DATA CLEANING SCRIPT

```
#set dataClean = to the basedataAll dataset

     dataClean=basedata_all

#detach basedata_all

     rm(basedata_all)

#remove trips with 1 passengers or less

     dataClean=subset(dataClean,dataClean$numPax>1)

#remove trips with 1 or fewer stops

     dataClean=subset(dataClean,dataClean$numStops>1)

#remove trips with distance shorter than 1km

     dataClean=subset(dataClean,dataClean$dist>0.99)

#remove trips with revenue 0

     dataClean=subset(dataClean,dataClean$rev>0)

#where speed is not equal to the distance/time, set it so

     dataClean$speed<-
     ifelse(dataClean$speed==(dataClean$dist/dataClean$travelTime*6
     0),dataClean$speed,(dataClean$dist/dataClean$travelTime*60))

#filter out trips 98th and 2nd percentile of trips based on speed

     dataClean=subset(dataClean,dataClean$speed<quantile(dataClean$
     speed,0.975)&dataClean$speed>quantile(dataClean$speed,0.025))

#filter out trips 98th and 2nd percentile of trips based on
travelTime

     dataClean=subset(dataClean,dataClean$travelTime<quantile(dataC
     lean$travelTime,0.975)&dataClean$travelTime>quantile(dataClean
     $travelTime,0.025))

#save the cleaned data as a ".txt" file extension

     write.table(dataClean,"cleanedData.txt",sep="\t",row.names=FAL
     SE)

#attach dataClean to the system memory in project in order to do
direct computation on the variables

     attach(dataClean)
```

APPENDIX B – FILE COPIER

Create three separate lists of Trip IDs, one for each cluster in R:

```
km1_clus1_tripIDs <- km1.cluster1[,c(1)]
km1_clus2_tripIDs <- km1.cluster2[,c(1)]
km1_clus3_tripIDs <- km1.cluster3[,c(1)]
```

Create three separate text files, one for each list above, containing the Trip IDs in R:

```
write.table(km1_clus1_tripIDs, "km1_clus1_tripIDs.txt", sep = "\t",
row.names = FALSE)
write.table(km1_clus2_tripIDs, "km1_clus2_tripIDs.txt", sep = "\t",
row.names = FALSE)
write.table(km1_clus3_tripIDs, "km1_clus3_tripIDs.txt", sep = "\t",
row.names = FALSE)
```

Identify parent directory in Python:

```
source_folder = 'C:\\Users\\dupreezd\\Anaconda2\\mbt_files'
```

Generate the *master list* of all filenames in the parent directory in Python:

```
master_list = (os.listdir(source_folder))
```

Read in the Trip ID text files for each cluster to generate the *trip lists* in Python*:*

```
km1_clus3_ids = [line.rstrip('\n') for line in
open('km1_clus1_tripIDs.txt')]
km1_clus3_ids = [line.rstrip('\n') for line in
open('km1_clus2_tripIDs.txt')]
km1_clus3_ids = [line.rstrip('\n') for line in
open('km1_clus3_tripIDs.txt')]
```

Create and identify the three destination subdirectories in Python:

```
dest_km1_1 = 'C:\\Users\\dupreezd\\Anaconda2\\mbt_files\\km1_clus1'
dest_km1_2 = 'C:\\Users\\dupreezd\\Anaconda2\\mbt_files\\km1_clus2'
dest_km1_3 = 'C:\\Users\\dupreezd\\Anaconda2\\mbt_files\\km1_clus3'
```

Match the Trip IDs in the *master list* with the *trip lists* to generate the file lists:

```
matching1 = [s for s in file_list if any(xs in s for xs in km1_clus1_ids)]
matching2 = [s for s in file_list if any(xs in s for xs in km1_clus2_ids)]
matching3 = [s for s in file_list if any(xs in s for xs in km1_clus3_ids)]
```

Iteratively copy the files based on the file names in the three *file lists* in Python:

Cluster 1:

```
for i in range(0,len(matching1)):
    copyfile(source_folder+'\\' +matching1[i],dest_km1_1+'\\'+matching1[i])
```

Cluster 2:

```
for i in range(0,len(matching2)):
    copyfile(source_folder+'\\' +matching2[i],dest_km1_2+'\\'+matching2[i])
```

Cluster 3:

```
for i in range(0,len(matching3)):
    copyfile(source_folder+'\\' +matching3[i],dest_km1_3+'\\'+matching3[i])
```