CIV4044F/S: RESEARCH PROJECT FOR THE DEGREE OF BACHELOR OF SCIENCE

O-D Matrix Estimation Procedure for Public Transport

Prepared by:

Ziphozihle Luvuno

Supervised by:

Mark Zuidgeest

Date of final submission:

27 November 2017

Project submitted in fulfilment of the requirements for the degree of

"Bachelor of Science in Civil Engineering"

Department of Civil Engineering

University of Cape Town, Private Bag Rondebosch, 7700

South Africa 7700

Plagiarism Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and to pretend that it is one's own.

2. I have used the Harvard Convention for citation and referencing. Each significant contribution to and quotation in this report form the work or works of other people has been attributed and has been cited and referenced.

3. This report is my own work

4. I have not allowed and will not allow anyone to copy my work with the intension of passing it as his or her own work.

Abstract

With current technological advances, the transport industry has been able to collect a vast amount of data. Smart card data is increasingly being used for transit network planning, analysis, and network demand forecasting. Thus, Automatic Data Collections (ADC) systems provide transit agencies with useful information that was previously expensive or improbable to obtain at the accuracy level ADC systems provide.

This thesis documents the development of a procedure to estimate a Bus Passenger Trip Origin-Destination Matrix (OD Matrix) based on ADC data collected from MyCiti, which consists of Automated Fare Collection (AFC) and Automatic Passenger Count (APC) data. The procedure will consist of four main steps: data preparation, estimation of a Single Route OD Matrices for all routes in the partial demonstration network, estimation of a Network Level OD Matrix using transfer flow information and validation of output against actual O-D data. The single route OD matrix estimation requires the development of a seed matrix constructed from AFC data (which is generated using the Trip-chaining method) and marginal control totals (boarding and alighting counts) derived from APC data. The Iterative Proportional Fitting (IPF) method is used to estimate the single route OD matrices, while the network level OD matrices are generated using two techniques namely Proportional distribution and Modified IPF method. The transfer flows for the network level OD matrix are provided at route level (transfer destinations at stop level are combined to be represented by the transfer route they occur on).

With the proposed procedure to estimate the O-D matrix of public transport, it will potentially be possible to apply the procedure to other PT modes which are flat fare based such as GABS in Cape Town.

Thesis Supervisor: Mark Zuidgeest

Title: Associate Professor, Department of Civil Engineering

Acknowledgements

This acknowledgment section goes to my supervisor, Mark Zuidgeest for the guidance he offered. Eddie Beukes for introducing me to the MyCiti dataset and providing insight for shaping the conceptual framework for this project. My family for getting me to remain focused and motivated in times of uncertainty. My friends and classmates for their suggestions on how to deal with some issues faced during this research project.

Table of Contents

Plagiarism Declaration **i**

Abstract **ii**

Acknowledgements **iii**

List of Tables **iv**

List of Figures **v**

List of Acronyms **vi**

Chapter 1

Introduction **1**

1.1 Background to the Study 1

1.2 Research project motivation and objectives 2

1.2.1 Motivation for research 2

1.2.2 Objectives of research 3

1.3 Scope and limitations 4

1.4 Research Workflow 4

1.5 Thesis Organization 6

Chapter 2

Literature Review **7**

2.1 OD data collection methods 7

2.1.1 Main OD data collection methods 7

2.1.2 Alternative OD data collection methods 7

2.2 Smart Card use in Automatic Fare Collection system 8

2.3 OD Estimation Using Automatic Fare Collection Data Sources 9

2.3.1 Trip-chaining method OD matrix estimation from AFC data 10

2.4 Full Expansion OD matrix estimation statistical techniques 12

2.4.1 Proportional Distribution 13

2.4.2 Iterative Proportional Fitting (IFP). 13

2.4.3 Maximum Likelihood Estimation (MLE) 13

2.4.4 Tsygalnitzky’s Fluid Analogy Method 14

2.5 Comparison of the methods 15

Chapter 3

MyCiti Data Description 16

3.1 General description of Dataset 16

3.2 MyCiti Automatic Data Collection System 16

3.2.1 Automatic Fare Collection System (AFC) 16

3.2.2 Automatic Vehicle Location system 18

3.2.3 Automatic Passenger Count system 18

3.3 Transfer Stop inference 18

3.4 General Analysis of dataset 20

3.4.1 Peak travel times 20

3.4.2 Average travel time in a MyCiti day 21

Chapter 4

General Approach to Estimating BUS OD Matrix Using ADC systems 22

4.1 General Data Preparation 22

4.2 Trip-chaining method 23

4.2.1 Trip-chain algorithm for O-D estimation 23

4.3 Single Route OD Matrix Estimation 27

4.3.1 Iterative proportional fitting 27

4.4 Network Level Bus Passenger Linked Trip OD Matrix estimation 28

4.5 Total transfer flows 30

4.5.1 Origin and destination distribution of the transfer flow 30

4.5.2 Modified IPF method: 30

4.5.3 Proportional Distribution 33

4.5.4 Comparison between proportional distribution and modified IPF 33

4.5.5 Non-transfer flows 34

4.5.6 Transfer flow OD matrices 34

Chapter 5

MyCiti Case Study Data Preparation 36

5.1 Validation OD matrices construction 36

5.2 Boarding Stop Identification for AFC Trips 38
5.3 Alighting Stop Inference 38
5.3.1 Trip-chaining method data preparation 38
5.3.2 Destination reassignment 39
5.3.3 Seed matrix formation 40
5.4 Boarding and Alighting Counts 41
Chapter 6
MyCiti Single Bus Route OD Matrix Estimation 43
6.1 Demonstration Corridor 43
6.2 Validation OD matrix 44
6.3 Iterative Proportional Fitting 44
6.4 Validation of the Estimation Results 46
Chapter 7
Network Level OD Matrix Estimation **48**
7.1 Transfer Flow Totals 48
7.1.1 Transfer trips in AFC data 48
7.2 Proportional Distribution 49
7.3 Modified IPF 51
7.3.1 Modifying the seed matrix 51
7.3.2 Modifying boarding and alighting totals 52
7.3.3 Modified IPF results 52
7.4 Validation of Results 53
7.5 Transfer flows Matrices 56
Chapter 8
Conclusions and Recommendations **57**
8.1 Supplementary Discussion of Main Results 57
8.1.1 Applications of single route OD matrix estimation 57
8.1.2 Application of a network-level OD matrix estimation 58
8.1.3 Discussion of potential Bias 58
8.2 Application of Procedure to GABS 59

8.3 Recommendations 60

8.3.1 Use of IPF method 60

8.3.2 Use of modified IPF method or proportional distribution for transfer flows 60

8.4 Potential for Future Research 60

8.4.2 Applying the methodology to other transit agencies 61

References **62**

Appendix A

(Additional OD Matrices Generated) 65

Appendix B

(Ethics Approval) 70

Appendix C

(Excel Procedures Used) 73

Appendix D

(Map of the Complete Myciti Network) 96

List of Tables

Table 2-1 Summary findings and possible gaps in existing literature. 12
Table 3-1 Sample of the raw MyCiti AFC data 17
Table 4-1: Transfer Flow OD Matrix 35
Table 5-1: Cleaned Sample data for trips of one passenger 37
Table 5-2: Sample data illustrating destination inference using Trip-chaining 39
Table 5-3: Sample data illustrating destination inference using Trip-chaining (1) 39
Table 5-4: Seed matrix for Route DO1 Kuyasa- Civic Center 41
Table 5-5: boarding and alighting counts 42
Table 6-1: Actual OD Matrix for Route DO1 Kuyasa to Civic Centre 44
Table 6-2: IPF OD Matrix for DO1 Kuyasa to Civic Centre 45
Table 6-3: Relative comparison of the estimated OD flows against actual OD flows 46
Table 7-1: Total OD flows Ending at the Transfer Point 49
Table 7-2: Total OD Flows Originating from the Transfer Point 50
Table 7-3: Transfer-only Flows with Origins Distributed 50
Table 7-4: Transfer-only Flows with Destinations Distributed 50
Table 7-5: Modified Route DO1 Kuyasa-Civic Center with Virtual-columns and Virtual-rows representing transfers to other routes 51
Table 7-6: Boarding and alighting Counts with Virtual-entries 52
Table 7-7: Modified IPF Method with Virtual Columns and Rows 53

List of Figures

Figure 1-1: Research Workflow 5
Figure 2-1: Smart card information system 8
Figure 3-1: Gates with card validators and On-board card validators 16
Figure 3-2: Inferred Destinations at Transfer Point vs without inferred destinations at transfer point 19
Figure 3-3: Number of trips per hour in the day 20
Figure 3-4: Number of boardings and alightings per hour in the day 21
Figure 4-1: Alighting estimation model 24
Figure 4-2: Implementation of trip chain method diagram 24
Figure 4-3: Work flow diagram for destination inference 26
Figure 4-4: Trip transfer inference example 28
Figure 4-5: General Set up for a Full network OD Matrix 29
Figure 4-6: Example Schematic Network 31
Figure 4-7: Example Schematic Network (OD matrices) 32
Figure 4-8: Scenario 1 of Proportional Distribution assumptions not holding 33
Figure 4-9: Scenario 2 of Proportional Distribution assumptions not holding. 34
Figure 5-1: Transfer point inference 37
Figure 5-2: Destination reassignment scenarios 40
Figure 6-1: Demonstration Corridor containing partial MyCiti Network 43
Figure 7-1: comparison of origin distribution of transfer flows 54
Figure 7-2: Comparison of destination distribution of transfer flows 55
Figure 8-1: Load along a single bus route. 57
Figure 8-2: AFC system (wayfare200 driver console) 59

List of Acronyms

ADC	Automatic Data Collection
AFC	Automatic Fare Collection
APC	Automatic Passenger Count
BRT	Bus Rapid Transit
CoCT	City of Cape Town
IPF	Iterative Proportional Fitting
GABS	Golden Arrow Bus
GPS	Global Positioning System
MLE	Maximum Likelihood Estimation
OD	Origin-Destination
PT	Public Transport
StatsSA	Statistics South Africa
SCAFC	Smart Card use in Automatic Fare Collection
TDP	Totally Disaggregate Approach

Chapter 1

Introduction

1.1 Background to the Study

Mr. Eddie Beukes from MyCiti approached Prof Mark Zuidgeest (referred to as the Supervisor) with a proposition of using AFC (Automatic Fare Collection) data collected by MyCiti to find an algorithm to predict and validate the OD (Origin-Destination) matrix for the MyCiti network when destinations are unknown, which could be subsequently used for other public transport (PT) routes.

Transit agencies have historically planned and managed their networks and services with limited knowledge of passenger travel patterns. Passenger-centric data, such as the origins, destinations and transfer flows of individual passengers have traditionally been collected mostly through costly surveys. The increasing adoption of ADC (Automatic Data Collection) systems is providing transit agencies with a vast amount of disaggregate data, which can be manipulated to reveal travel information from a passenger-focused perspective (Gordon, 2012).

A PT OD matrix has a wide range of applications in the transportation industry. It can be utilized for operations analysis, service planning and management, and impact analysis of a particular change in a network. For example, when considering potential express bus service system, it is required to know the number of passengers traveling between various stops to estimate the demand and the potential passenger travel time savings (Cui, 2006).

Thus, estimation of OD trip matrices is an intermediate step in many transport analyses. Each entry in such a matrix represents the volume of travel originating in a particular zone and intended for some other zone (Carey et al., 1981). Historically a commonly used method in estimating an OD matrix involves a survey of individual trip-makers. Trip volume forecast are usually obtained by sequential application of models, which initially estimate the total number of trips generated by or attracted to a particular zone and then distribute the expected trips among potential destination zones (Carey et al., 1981). The expense of such passenger surveys and counts is quite high and consequently made them infrequent practices.

Thus, the emergence of technologies such as AFC (Automatic Fare Collection), APC (Automatic Passenger Counting), GPS (Global Positioning Systems) has broadened the tools used for analysis and transportation planning. A combination of these technologies constitutes a new transport data source that can be used by transport authorities to improve the daily operations of public transport (Bagchi and White, 2004). Smartcard data is increasingly being used for transit network planning and network demand forecasting, consequently PT origin-destination (O–D) estimation has become a significant product of processing smart card data (Alsger, 2016).

The data collecting methods used by MyCiti has certain advantages over existing bus ticket machine data and sample data sources, allowing for a range of analysis applications that transport authorities may have found previously difficult to conduct. Therefore, there is an opportunity to use this advanced system to benefit other public modes of transport, by creating a procedure which can be utilized to obtain the OD matrix estimation and subsequently calibrated for other PT modes in Cape Town.

1.2 Research project motivation and objectives

To further understand the problem pertaining to O-D matrix estimation for Public Transport, the following section will discuss research motivation and objectives.

1.2.1 Motivation for research

The motivation for this research was initially sparked by the quality of AFC data which MyCiti accumulates daily. The potential uses of the dataset can provide transit authorities with insight into commuter travel patterns, and could lead to improvement of services delivered for not only MyCiti but other PT systems in the Cape Town area. This thesis will mainly focus on OD matrix estimation procedures using MyCiti AFC data.

Currently, the MyCiti AFC system validates boardings and alightings of individual passengers, thus allowing estimation of an accurate OD matrix directly. However, this cannot be done for other PT modes in Cape Town, which have a flat fare system. PT systems with flat fare only have boarding validation as the passenger pays only when boarding, and the alighting stop of a passenger is unknown. Generating the OD matrix estimation for a flat fare PT system is not a trivial task

By creating a procedure which can estimate the OD matrices of the MyCiti system (when all alightings within the MyCiti AFC data have been omitted to simulate a flat fare system), and then validating the estimated OD matrices against the Actual (validation) OD matrices of MyCiti. It would thus be possible to judge the accuracy and applicability of the procedure on other PT systems which are flat fare based.

PT systems which are flat fare in Cape Town is the Golden Arrow Buses (GABS), which in 2015, GABS identified the handling of money as a payment method on buses, kiosks and at depots as a high-risk factor in the company's business model. To mitigate this operational risk, GABS took a decision to invest in an AFC system at a cost of approximately R85 million (HCI, 2016). Thus, GABS system will be different to MyCiti, as Golden Arrow's routes form an open system in which passenger is not expected to tap out (interchange, 2016); consequently, leaving no record of the alighting location. Thus, the procedure (which involves inferring alighting destinations) can potentially be applied to GABS AFC data to estimate GABS OD matrices.

A number studies can be found on the use of AFC data to estimate the OD matrices for transit systems without alighting information, such as those found in New York, Santiago, Chicago and London (Barry et al.,2009; Munizaga and Palma, 2012; Gordon, 2012). This thesis will use aspects from aforementioned resources as well as various other papers to develop an algorithm which is best suited for PT systems in Cape Town.

1.2.2 Objectives of research

The immediate aim of this research is to create an exemplar procedure to estimate a single-bus OD matrix and network level OD matrix using raw data obtained from MyCiti transit authority. More specifically, this thesis seeks to fulfil the following objects:

- Datamining from the AFC dataset to find potential uses for transit authorities;
- Infer boarding and alighting locations to obtain seed matrix and control totals;
- Infer transfer flows between individual trip-legs;
- Apply expansion techniques specifically IPF, modified IPF and proportional distribution;
- Validate the estimated matrices and;
- Demonstrate the procedure by application to the MyCiti bus network. Therefore, this thesis intends to establish a methodology to estimate to OD matrices which is semi-automated (as computational analysis was completed using excel). A fully automated process to create a software for regular production of OD matrices forms potential for future research.

1.3 Scope and limitations

The investigation will attempt to estimate the OD matrix for PT, which entails utilizing AFC data gathered by MyCiti to create a procedure which can be used to estimate a single route OD matrix and a network level OD matrix. Subsequently, validating the estimated OD matrix against the actual (validation) OD matrix generated directly from MyCiti data.

Therefore, the project will primarily be an extensive desk study, with the AFC data limited to what is provided by MyCiti authorities. This investigation requires statistical skills and computer programming skills in order to successfully generate and apply the algorithm. While the investigator may not currently possess such expertise, it is required that a basic and appropriate understanding of such topics is acquired during the period of research. Consequently, the outcomes of the project are limited by the expertise of the investigator, and secondly, the time period over which this investigation will occur.

1.4 Research Workflow

The research process was exploratory in nature in terms of finding potential uses of AFC data, however it has also produced new data that extended the existing knowledge of OD flows for MyCiti cardholders. The research intentions were initially based on creating a software, which could be used to generate for OD Matrix of an AFC system, so that transit authorities could use it on a daily basis to analysis. The software was going to be built on platforms such as python, however the limited time and an improvement of understanding regarding current literature on the topic and research methodologies the focus change. The nature of the project required extensive research and developing analytical skills.

Thus, the focus of the thesis was narrowed to find a procedure which could be used to generate OD matrices for flat fare systems using AFC data. Consequently, relatively large portions of this thesis is dedicated to documenting the methodology and the results of the OD matrix estimation procedure.

Figure 1-1 below illustrates the work flow process which was completed for this thesis. A number of activities are interconnected and are an iterative process.

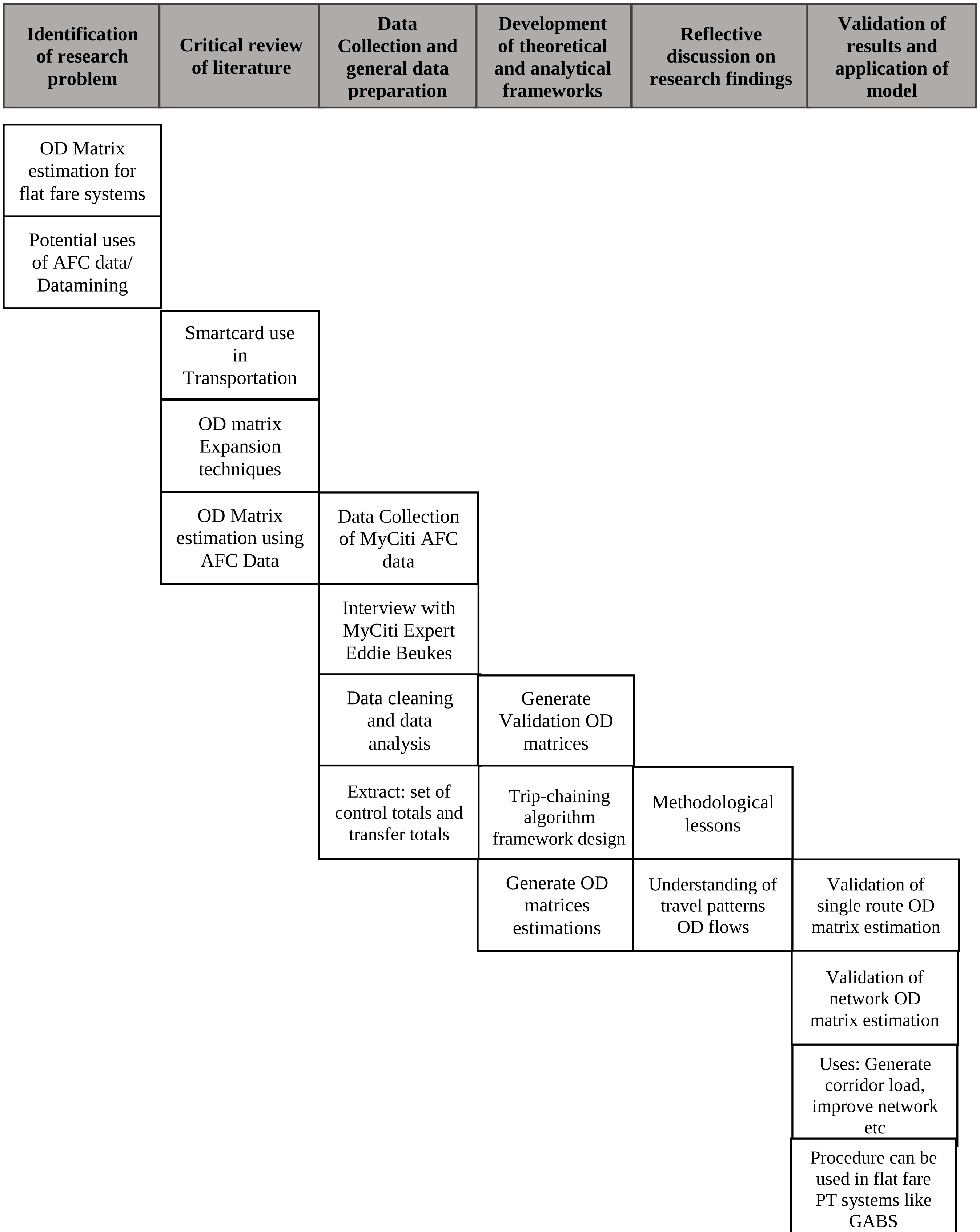


Figure 1-1: Research Work-flow

1.5 Thesis Organization

The thesis is divided into 8 chapters. Chapter 2 reviews related literature, followed by a description of the dataset in chapter 3. Chapter 4 presents a general approach to estimating a bus OD matrix using ADC systems, while Chapter 5 introduces the MyCiti case study and the data preparation required. Subsequently, Chapter 6 describes the construction of a Single-Bus OD Matrix Estimation, followed by the Chapter 7 illustrating the construction of a Network level OD Matrix Estimation. Finally, Chapter 8 provides a conclusion and recommendations for future research.

Chapter 2

Literature Review

This chapter introduces previous studies conducted with regards to OD estimation for PT. Section 2.1 discusses the different OD data collection methods, while section 2.2 discusses Smartcard use in AFC systems. Section 2.3 focuses on OD estimation using ADC data sources and section 2.4 reviews the different OD matrix expansion statistical techniques. Finally, Section 2.5 compares the different techniques.

2.1 OD data collection methods

2.1.1 Main OD data collection methods

There are numerous methods to calculate the OD matrix, with the underlying difference being the way the data is collected. The main method proposed for calculation of OD matrix estimation is using origin and destination data from AFC, obtained directly from the transport service provider MyCiti. The AFC system automatically counts the number of passengers on boarding and alighting from a bus.

2.1.2 Alternative OD data collection methods

Other methods of data collection include the use of a checker notes a seating diagram the boarding stop as the passengers choose a seat and note their destination stop as they alight. This method has some clear draw backs, the first being the fact that some people stand on the My Citi Bus. The second method involves a survey which asks questions such as the origin and destination of the passenger. According to Simon and Furth (1985), the response rate tends to be low and biased in favor of more literate passengers and those traveling longer distances who are more likely to get a seat. The third method requires cards coded with a unique origin stop signature be distributed to entering commuters, who are then instructed to hand back the cards when leaving the bus. Simon and Furth (1985) argue that this data collection method has a near 100% response rate with little bias. However, it is also recommended to have a checker at each door who can quickly communicate instructions to the boarding/alighting passengers. The only flow to the method might occur is the transit vehicles have insufficient capacity to carry demands, such that passenger queues form at stops (Cassidy & Li, 2006).

2.2 Smart Card use in Automatic Fare Collection system

Smart card technology is by no means a new technology, first invented by German researchers Dethloff and Grotrupp (Shelfer and Procaccino, 2002) and commercialized in 1977 by Motorola (Trepanier et al, 2007). In the past two decades, the technology has been growing rapidly in the transportation planning industry. Smart card application in the transit industry fall into the two following categories:

- Used for transactions, as credit is deducted from the card and;
- Personal identification and authentication. Information stored on the data system verifies whether a person can use the transit system or not.

Smart card use results in a generally more efficient system with lower risk of shirking from payment as Smart card automatic fare systems (SCAFC) instantly validates a commuter's identity on boarding.

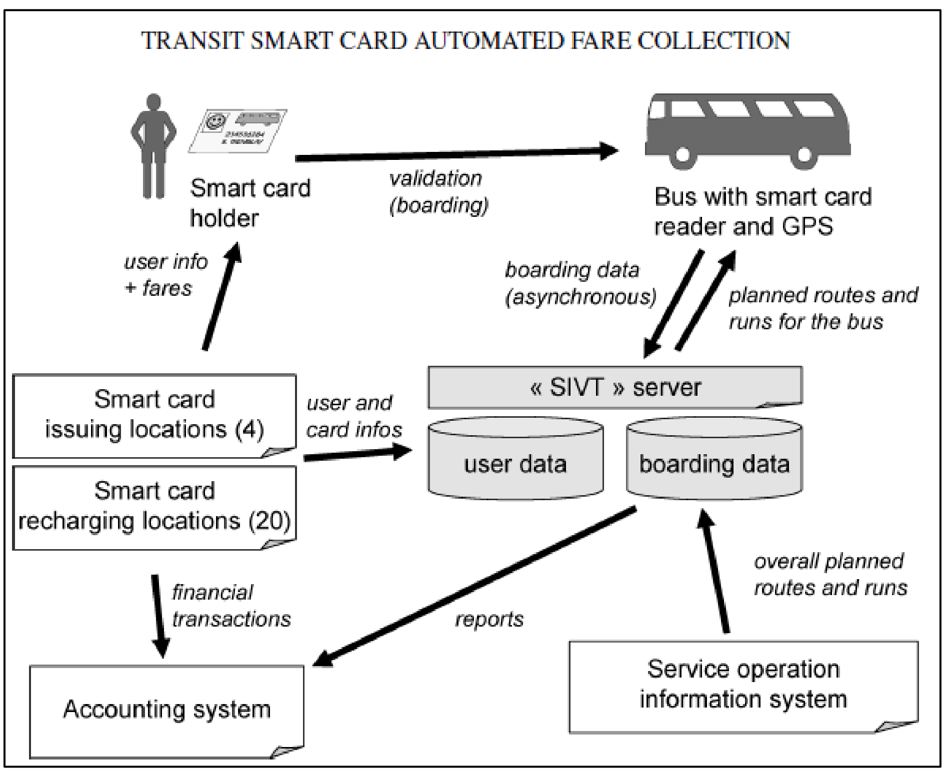


Figure 2-1: Smart card information system (Trepanier et al, 2007)

The Automatic Fare Collection system is comprised of multiple subsystems as illustrated in the figure 2-1 above. The smart cards are bought at issuing locations similar to those of the My Citi Bus and passenger information is captured at these points in most cases (Trepanier et al, 2007). Other points can be usually established to recharge credit in the smartcards such as the My Citi virtual kiosk. When commuters board the bus, the smart card's fare is validated as follows:

i. The bus system has planned runs for daily operation (a run is a sequence of stops in one direction). The GPS reader in the bus identifies where the commuter boards.
ii. The system instantly verifies that the user's card fare is accepted depending on the customer's fare and bus route.
iii. Card number, date, validation status and stop number are stored at each boarding. Subsequently, this data is downloaded to the central server at end of a business day depending on how it was set up.

(Trepanier et al, 2007)

In conclusion, a smartcard is a great tool for transit planners, however in flat fare systems its use is limited as the system do not have alighting validation, such as the case in Chile, Gatineau, New York. In the previously mentioned examples it was required to infer the destinations, using algorithms suited for the particular network and its AFC dataset,

2.3 OD Estimation Using Automatic Fare Collection Data Sources

Transport planning in the modern age has been aided by the inclusion of technology and consequently research in OD matrix estimation for public transport systems has become possible. However, the focus of the research tends to be different depending on the type of data available.

The differences in the type of data available usually depend on the fare collection system put in place. A large number of transit agencies outside of South Africa use a flat fare system, which only validates the boarding of the passenger and thus the alighting location remains unknown. An example of such as system is the Chicago Metro Bus services and locally Golden Arrow Bus. Systems with a more complete dataset of a passenger's movement usually have a Distance Fare system, thus requiring validation of both the boarding and alighting locations. The MyCiti bus system has a Distance Fare system in place, resulting in relative ease of establishing an OD matrices for the network (MyCiti, 2017).

The first section reviews the trip reviews the trip-chaining method of destination inference used for entry-only bus system estimation (section 2.2.1). Section 2.2.2 will review the method utilized by the Caliper Corporation to estimate bus system OD for New York City Transit.

2.3.1 Trip-chaining method OD matrix estimation from AFC data

The trip-chaining method is used in entry controlled systems to obtain the destination stop of a passenger trip, which is an essential step in obtaining a seed matrix. To infer the destination, Trip-chaining was used based on the following three assumptions:

(1) Passengers start their next trip at the destination station of their previous trip, or at another station in nearby, a maximum allowable walking distance is 1000m (Barry, 2002).
(2) Passengers end their last trip of the stop where they started the first trip of the day (Barry, 2002).
(3) If a passenger changes routes, the transfer point is the destination of the first trip leg (Cui, 2006)

Not all passengers can have their destinations successfully inferred, the resulting OD flows need to be scaled to control totals (total boardings and alightings) to estimate a system-wide OD matrix.

Although Trip-chaining method is a method generally used for construction of a seed matrix (sample OD flows), the high sampling rate it achieves can still provide transit agencies useful information. Thus, the seed matrix generated using Trip-chaining method should be similar to the actual OD matrix in terms of relative weights of the OD flows, without applying any scaling statistical techniques (Cui, 2006). This method can be used in situations in which the control variables are difficult to obtain.

The following sections will review other literature regarding Trip-chaining method.

2.3.1.1 Studies conducted to estimate OD matrices using AFC data

Research to generate OD matrices from smart card data has been undertaken by numerous researchers, all of whom had slightly different ideas. For example, Trepanier et al. (2007) proposed a method to estimate the alighting point of commuters along a route where users are only validated when boarding; Chapleau and Chu (2007) propose a method to identify and replace suspicious observations from the AFC systems; Zhao et al. (2007) implemented a method to estimate the alighting point for rail boarding transactions in the Chicago CTA system, by focusing on rail boarding followed by a bus boarding transaction. They assigned the nearest station to the next boarding bus stop within a 400m radius of the alighting station. Zhao et al. (2007) reported a 71% success rate in estimating alighting stations for rail boarding, while Trépanier et al. (2007) obtained a 66% success for the bus-only Gatineau system. According to Munizaga and Palma (2009), their research adds more dimensions and complexity to the previously mentioned studies, in that their method can be implemented in large scale multimode public transport system.

2.3.1.2 South-East Queensland, Australia case study

Alsger et al. (2016) evaluated the common trip-chaining method assumptions using a smart card fare dataset obtained from TransLink, the public transport authority of South-East Queensland (SEQ), Australia. The main advantage of their unique dataset (which is like the MyCiti dataset) for evaluation of the trip-chaining method assumptions was since their dataset included both boarding and alighting times and locations for each passenger of the public transport services that comprise buses, trains, and ferries. The study comprised of the usual individual assumptions (allowable transfer time, allowable walking distance and last destination of a given day being the same as the first origin of that day) of the trip-chaining method, in a situation where actual boarding and alighting information were known (Alsger et al, 2016). Alsger (2016) further claims that none of the studies mentioned in the previous paragraph have implemented and validated the whole estimation algorithm with are reliable dataset.

Table 1 below provides a summary of studies regarding Trip-chaining method and the possible gaps in existing literature.

Table 2-1 Summary findings and possible gaps in existing literature.

Component		Studies	Findings	Gaps
Estimation Assumptions	Walking distances (buffer zones)	Cui (2006), Wang (2010), Nasser et al. (2011), and Munizaga and Palma (2012)	Different walking distance were chosen to infer alighting stops (eg 400m-1100m)	Different values were used for the assumptions of the O–D estimation. None of these studies validating the whole estimation algorithm with a reliable dataset
	Transfer time	Bagchi and White (2004), Nassir et al. (2011), Kieu et al. (2013), Ma et al. (2013), Hofmann and O'Mahony (2005)	Different transfer times were chosen to connect trip-legs to infer O–D trips (e.g., 30, 60 and 90 min)	
	Last destination assumptions	Barry et al. (2002), Nassir et al. (2011), Munizaga and Palma (2012), and Gordon et al. (2013)	Assumed the last destination as the first origin, where others assumed it as the closest stop to the first origin	
Validation attempts	Additional data requirement for validation	Farzin (2008), Barry et al. (2009), Devillaine et al. (2012), Munizaga et al., (2014) and Chow (2014)	Additional data (e.g., travel survey, personal interviews, and passenger counting) were used for validation	The accessibility and quality of the additional data required for further evaluation of the trip-chaining method are usually a concern

2.4 Full Expansion OD matrix estimation statistical techniques

There are several expansion methods of OD estimation for public transport, with some methods possessing higher accuracy in certain conditions than others. Some of the more popular methods of PT OD matrix estimation are as follows:

i. Proportional Distribution.
ii. Iterative Proportional Fitting (IFP).
iii. Maximum Likelihood Estimation (MLE).
iv. Tsygalnitzky's Fluid Analogy Method.

The following section will provide an overview of what each method entails in order to gain an understanding of which method would generate the best OD matrices for the provided MyCiti dataset.

2.4.1 Proportional Distribution

This method distributes transfer flows to eligible boarding stops proportional to the total flow from all eligible alighting stops to the transfer stop (Cui, 2006). Proportional distribution assumes that the passenger OD behavior is the same for transfer passengers and non-transfer passengers. Cui (2006) states that this method has the advantage of simplicity and it will be further discussed in section 3.3 and demonstrated in section 7.3.

2.4.2 Iterative Proportional Fitting (IFP).

According to Navick and Furth (1994), this method is usually used in the estimation of a single route OD matrix by combining a seed matrix with the total boarding and alighting counts. The method requires a high-quality boarding and alighting counts as the control totals, and a small but representative seed matrix. In addition, Navick and Furth (1994) state that the accuracy of the estimation depends on the reliability of the control totals. However, the quality of the seed matrix is also very vital to the final estimation result, with a biased seed matrix being worse than a null seed matrix or a gravity-based trip distribution model (Navick and Furth, 1994).

According to Cui (2006), the main difficulty with using the IPF technique is the probability that a significant number of potential OD pairs in a network have little or no travel between them. Cui (2006) further explains that matrix cells having zero OD flows because of the structure of the matrix are referred to as structural zeros, and those due to the low sampling rate of low O-D flows are termed non-structural zeros. If there are no structural zeros and if all data are consistent, IPF will yield a unique solution (Akiva, 1987), but if the potential travel matrix has a lot of O-D pairs with no observed trips, the IPF method will result in non-zero flows for only a portion of the O-D pairs which do have flow (IPF will be further discussed in the following chapters).

2.4.3 Maximum Likelihood Estimation (MLE)

MLE can be applied to estimate a single route OD matrix, a network level OD matrix or transfer flows. In addition, it can be utilized to combine data from different sources with different reliability (Cui, 2006). For MLE only a sample of the passenger OD flows is needed (the seed matrix), and a sample of the boarding and alighting counts rather than total boarding and alighting counts (Hsu cited in Cui 2006). A modified MLE formulation can introduce time period variables to represent the different sampling time periods for the different data sources. More specifically, the boarding and alighting counts obtained from Automatic Passenger Count (APC) data only represent observations during the time period equal to the sum of the vehicle headways with valid APC data (Cui, 2006). Though theoretically, this headway should be the actual headway, the actual headway varies along the route. For simplicity, the sum of scheduled headways for all vehicle trips with valid APC data is used instead.

2.4.4 Tsygalnitzky's Fluid Analogy Method

Using on-off counts (Tsygalnitzky cited in Simon and Furth 1985) was able to find a method for estimating a route O-D matrix. Tsygalnitzky's assumes that at a given stop, all qualified passengers are likely to alight. According to Simon and Furth (1985), a ‘qualified' passenger is one that has their boarding stop at least a certain minimum distance from the given stop. Essentially the system treats boarding passengers which meet the minimum distance as one fluid system, from which a representative cross-section is drawn at each alighting stop.

Furth and Simon (1985) claimed that the fluid analogy makes the computation of a unique O-D relatively simple. The fluid analogy was tested by the Southern California Rapid Transit District, using an on-board survey of 5 bus trips traveling a segment of a network. O-D data for the trips were collected by having checkers hand commuter's cards that had points of origin coded on them (similar data collection method described in section 2.1.2). The cards were punched per the alighting stop of the passenger. In addition, because the passengers were simply holding there was no bias present in the data, only two people per trip were omitted from the analysis (Furth and Simon, 1985). Thus, using the on-off counts an estimated O-D matrix was generated and used as a validation matrix. Furth and Simon (1985) further argued that compared to the methods presented by Akiva et al (1985) their current test is an advance, as it measures the estimate against actual O-D data and rather than alter previous base O-D matrices theirs could generate O-D matrices from on-off counts.

2.5 Comparison of the methods

Due to time constraints, it is estimated that only two methods can be tested on generating a network level OD matrix. Thus, the chosen statistical techniques are required to be computationally simple to apply, in order for transit agencies to easily apply the method on a daily basis if required.

First, comparing the three methods Tsygalnitzky's Fluid Analogy, IPF and MLE. Akiva et al (1985) found that when using a relatively larger dataset compared to the average on-board survey data, the differences between the results of IPF and MLE is negligible with the opposite being true for small dataset MLE results in more accurate results. In a study conducted by Cui (2006), IPF method was found to have a significantly simpler computational process compared to MLE, thus applying IPF on a daily basis would be simpler for the transit agencies and is the deciding factor as results of the two methods have similar results. The Tsygalnitzky's Fluid Analogy has only been tested for on-board surveys and conducting surveys on large networks tends to be extremely expensive and time-consuming rendering the process improbable. In addition, Fluid Analogy relies on the strong assumption of passengers travelling a certain minimum distance. However, in South Africa due to informal planning of settlements and Apartheid segregation, there is a large variance on the distance passengers of different ethnicities travel to reach their workplaces or CBD. Consequently, there are numerous factors which need to be considered when applying the minimum travel distance assumptions, making the Fluid Analogy method difficult to use.

Second, comparing IPF and Proportion distribution method. Per Cui (2006) there are two critical differences between the two methods, the first being the difference in basic assumptions, and secondly, the effect of the seed matrix. Cui (2006) states that the proportional distribution method assumes that the passenger O-D behavior is the same for transferring between stops and non-transfer passengers, while the modified IPF method does not depend on this assumption. The IPF method can fully utilize seed matrix information to obtain an OD matrix that is closer to the actual passenger behavior compared to Propositional distribution. However, Cui (2006) states that the use of IPF method requires a large sample size to avoid the potential biases in the modified IPF method that are due to non-structural zeros. In addition, proportional distribution is computationally easier to apply than the modified IPF method when dealing with large networks. Both these methods have specific benefits over each other, thus IPF and proportional distribution will be tested and the results compared.

Chapter 3

MyCiti Data Description

This chapter aims to introduce the MyCiti AFC dataset. Section 3.1 describes the MyCiti system and section 3.2 focuses on the Automatic Data Collection the system used to gather the data used in the case study.

3.1 General description of Dataset

The ADC data analyzed in this study was obtained from MyCiti, the public transport authority of the City of Cape Town, South Africa. It consists of 164 000 taps (boardings and alightings) and 72 000 trips all completed in a 24-hour period on the 9 May 2017.The network consists of 41 different single-bus routes and a total of 452 stops. In the MyCiti network, a transaction record is generated and stored each time a passenger boards and alights. Each transaction contains information comprising: the operation date, route, ticket number, smartcard ID, boarding time, alighting time, boarding stop and alighting stop (MyCiti, 2017).

3.2 MyCiti Automatic Data Collection System

The MyCiti Automated Fare Collection system (AFC) is integrated with two other systems, Automatic Vehicle Location (AVL) and the Automatic Passenger Count (APC).

3.2.1 Automatic Fare Collection System (AFC)

The AFC System consists of permanent gates with card validator units at each of the stations and on-board card validators on each of the buses which will require a person to tap their EMV compliant card upon entering and exiting the bus system (MyCiti, 2017), as shown in figure 3-1.

Figure 3-1: Gates with card validators and On-board card validators

The relevant fare will be calculated based on the distance traveled and will be automatically deducted from the value previously loaded onto the card. All users on the system require a MyCiti card, thus there are no situations of cash transaction, all fare transactions produce AFC records. The MyCiti cards can be purchased at MyCiti kiosks and loaded with credit either at the kiosks or Auto Vending Machines which are available at most stations. A sample of the raw AFC data for MyCiti transactions is illustrated in table 3-1

Table 3-1 Sample of the raw MyCiti AFC data

CARD_NUM	**DEVICE_ID**	**ROUTE_ID**	**ROUTE_NAME**	**STOP_NAME**	**Transaction Type**	**Hour**	**Min.**
5759842851889720000	19705818	290	109 Civic Centre - Hout Bay	Glen Beach	1st Boarding	15	45
5760630154770400000	19689443	9999	9999 (COCT)	Adderley	Alighting	16	22
10055429	19705337	900	214 Dunoon - Table View	St Johns Wood	1st Boarding	7	3
5760630154770400000	19705834	892	261 Omuramba - Civic Centre	Mansfield	1st Boarding	6	15
5761458780519890000	19693789	881	260 Woodbridge - Summer Greens	Namer	Alighting	15	30
2800705707409090000	19732990	330	213 West Beach - Table View – Sunningdal	Tryall	Alighting	17	1

The data fields available in the AFC data play a vital role in the type of analysis that can be completed to generate an OD matrix. The data consists of the following fields:

i. Card number is a unique number which identifies a specific MyCiti card, which can be used to identify transit trips of a passenger. Even though the owner can borrow their card to other people for use, it will be assumed that each MyCiti card is used by one person (the owner).
ii. The Device ID is a unique number to identify the fare box where the MyCiti card transaction occurred and can be mapped to stop location. Therefore, the farecard transaction can be linked to a stop (location), depending on the Equipment ID.
iii. Route ID provides the route on which the transit trip occurred for the particular passenger.

iv. Route Name provides the route-direction combination required to generate a single OD matrix. Thus, the single OD matrix is generated according to direction of the bus along a particular route.
v. Stop Name act as the origin or destination location in the OD matrix. Hence, one OD flows is between two different stops.
vi. The Transaction Type column indicates whether a passenger was boarding or alighting when they tap their MyCiti cards at the entrance and exit gates.

3.2.2 Automatic Vehicle Location system

The AVL and APC are at stop level. Therefore, the AFC data can be matched with the AVL data identify the transaction at stop Level.

3.2.3 Automatic Passenger Count system

The APC system data is used for boarding and alighting counts. Most common APC systems have sensors which use double infrared beams, depending on the order the infrared beams are broken, the sensor determines whether passenger is alighting or boarding. The APC system for the MyCiti system is not located on the bus but at the entry and exit gates.

3.3 Transfer Stop inference

The MyCiti data provides a fully connected OD flows for passengers. Thus, regardless of the number of trip-legs a passenger goes through their final destination is given. However, when analyzing a transport network system, transit planners need to be able to focus on a route level. Hence, they need to be able to track movement of a passenger through the network system and that requires transfer points, or the OD flows within a route will be underestimated as illustrated in Figure 3-1.

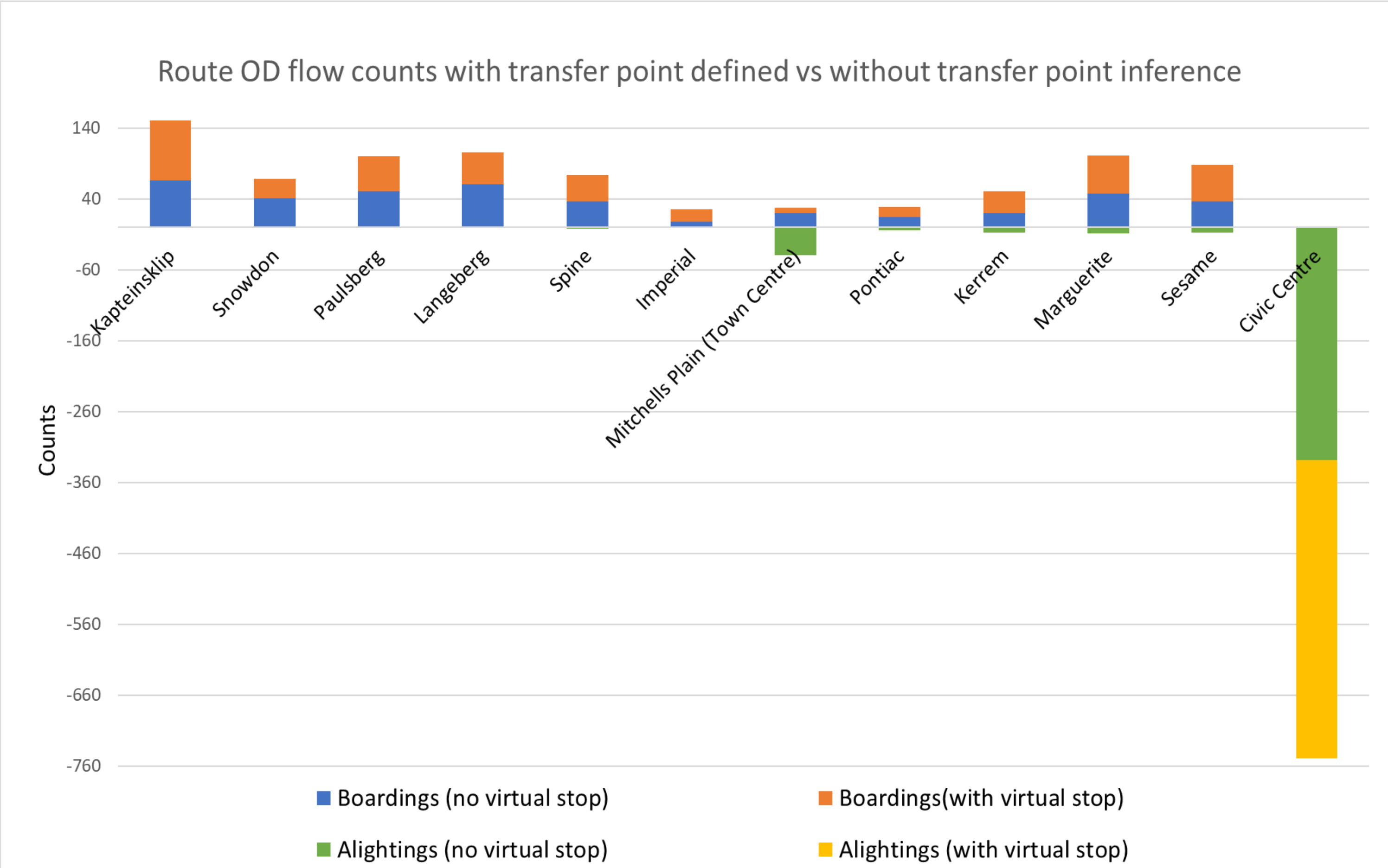


Figure 3-2: Inferred Destinations at Transfer Point vs without inferred destinations at transfer point

Figure 3-2 illustrates the boardings and alightings where Virtual stops act as destinations for transferring passengers on a particular route. It can be observed that a route without the transfer stops defined results in an underestimation of OD flows and when the transfer stop is defined at Civic Center that increases the amount of OD flow counts for the route direction. Therefore, after defining the transfer point, the is an increase in both the boardings represented by the orange and an increase in the alighting represented by the yellow. This process gives a better representation of the route load and how the routes interact through the network, enabling transit planners to the flow of commuters through the system.

3.4 General Analysis of dataset

The AFC dataset has multiple analysis applications it could serve, this section aims the various output which can be generated the by transit agencies in order to aid network planning.

3.4.1 Peak travel times

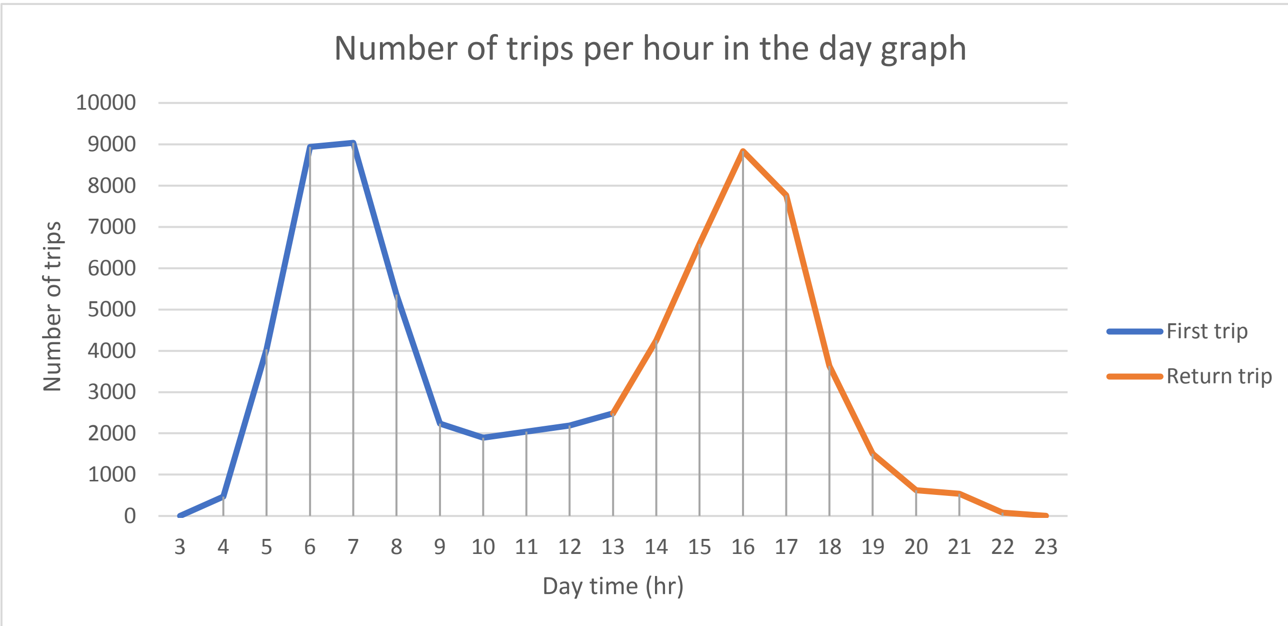


Figure 3-3: Number of trips per hour in the day

Figure 3-3 illustrates the number of trips per hours in the day, the graph has several applications which transport planner can use it for. The graph has two distinct peak points, the first is the morning peak, which is between 05:00-08:00 and the second peak is between 15:00-18:00. Use of such a graph can help indicate to planners the number of buses required to cater for the demand for the specific time periods and thus prevent congestion of passengers inside the system. In addition, transit planners at MyCiti use this information in their pricing policies, in which it is more expensive to use the buses during the peak time. This is done in the attempt to distribute the load to less congested time periods and to rise revenue as a high demand is justified by higher prices. In addition, with AFC data from other days, transit planners would be able to tell the changes in trip demand according to the day of the week and even the effect of weather on travel patterns.

3.4.2 Average travel time in a MyCiti day

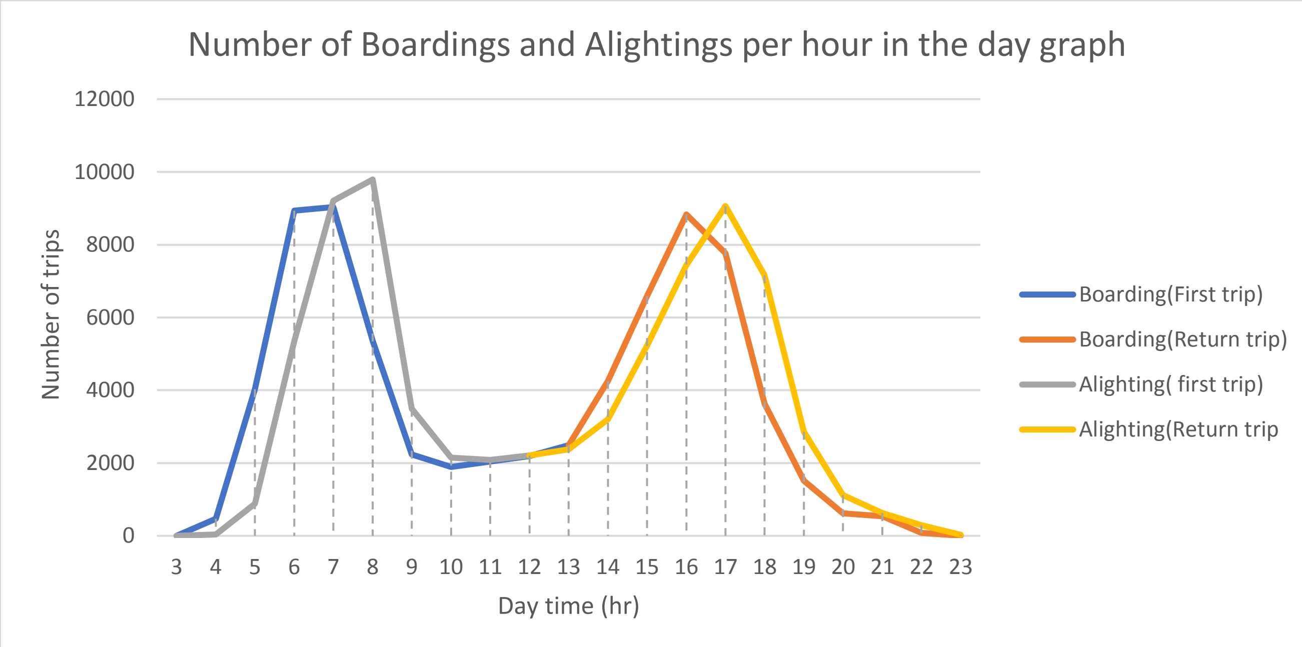


Figure 3-4: Number of boardings and alightings per hour in the day

Figure 3-4 shows the boarding and alighting per hour in the day. The graph can be interpreted as the time difference between boardings and alightings or the travel time. For example, considering point A on the graph, if passengers boarded bus at approximately 14:00, transit planners would know that at that specific time there are roughly 4000 passengers using the MyCiti system, and that the average travel time for those passengers is the time difference between the two black dots on the graph (approximately 30min). In addition, it can be observed that the largest average travel times occur at peak hour points of boardings and alightings. Therefore, from this graph, transit planners would be able to predict average travel time throughout the day and thus update timetable schedules accordingly. With data from other days, transit planners would be able to track average travel time according to the day, or even whether (e.g. rainy days would result in greater time differences between the boardings and alightings).

Various other analysis can be can be conducted through AFC data, to help transit planners, such as dwelling time information according to the time of the day or overcrowding during peak hours; time delays can be tested and speed of vehicles per route can be tested. AFC systems, make it possible to continually improve the MyCiti network making AFC data a powerful tool for transit planners.

Chapter 4

General Approach to Estimating BUS OD Matrix Using ADC systems

The best suited procedure to estimate the OD matrix is dependent on the type and quality of data available; however, there is a general approach which will be discussed in this chapter. The general approach assumes that AFC data is sufficient to generate a set of sample boarding and alighting counts for all stops (referred to as control totals). In addition, a sample of passenger OD flows (referred to as the seed matrix). The general approach consists of the following steps:

i. Process the raw data from AFC systems to obtain a sample set of boarding and alighting counts. This step is discussed in Section 4.1
ii. Use the trip-chaining method to obtain a representative seed matrix which consists of a sample passenger trip OD matrix. This step is discussed in Section 4.2
iii. Combine the seed matrix and control totals using a statistical technique (IPF) to produce a single route OD matrix for every route direction. This step is discussed in Section 4.3.
iv. Use transfer flows totals to link all the single route OD matrices to form a network level OD matrix. This section is discussed in Section 4.4 and 4.5.

It is noted that the main objective is to produce the estimation of a network level OD matrix, however, the intermediate products such as single route OD matrices and the total transfer flows by route-direction combination have significant planning value to transport authorities.

4.1 General Data Preparation

In data preparation, the first factor to consider is the type of data available, which largely depends on ADC systems the transport authority is utilizing. According to Gordon (2012), the lack of certain data may compromise the quality of the solution, for example with AFC data integrated with AVL it is possible to locate the boarding stop for a passenger trip. Without AVL data, it may still be possible to determine the general location of the boarding location of a passenger by identifying the location of the bus according to the schedule data, however, stop level accuracy might be compromised (Gordon, 2012). In addition, relying on schedule data would require more processing effort.

To obtain a set of boarding and alighting counts will require the bus system to have an Automatic Passenger Count (APC) system integrated with the AFC system. The APC system can count the number of passengers entering and exiting a bus, subsequently the counts are categorized according to the stop location using AVL system. However, in flat fare systems passengers are only expected to tap in and not out, therefore even though the APC is counting the number of passenger's alighting at a certain stop it doesn't record the passenger's Card_Num. Hence, these counts can be aggregated to obtain the complete set of boarding and alighting counts at a stop.

In processing the data, some entries will not have the required characteristics to be used in completing the OD matrix since the destination cannot inferred and thus will be dropped from the seed matrix as per Barry (2012) destination inference assumptions. To counter the loss of trips scaling factors will be implemented with the use of IPF method. Since the scaling factors are different for each route-direction combination and time periods it is required to estimate a scaling factor for each route and time period. In such cases, the potential for bias when working with data is something to consider, both in the raw data and because of the processing methods used

4.2 Trip-chaining method

4.2.1 Trip-chain algorithm for O-D estimation

The trip-chaining method develops a list of PT passengers' trips, by connecting the corresponding trip-legs for each smart card holder, when criteria for a transfer between the trip-legs are met, refer to figure 1 below for visual illustration (Alsger et al., 2016). These criteria are often based on the assumption which is explained in the next section.

4.2.1.1 Assumptions regarding Trip-chain method

Munizaga and Palma (2009) propose the method to estimate alighting is similar to Barry et al. (2002) in assumptions which are users will return to their previous trip's destination station and at the end of the day, users return to the first boarding station of the day. However, this study will use the more realistic assumption made by Gordon cited in (Alsger et al., 2016) which is to choose a stop on the last trip-leg's route which is the closest stop to the first origin in a given day as the last destination of the day). Additional constraints are added to improve the estimation procedure such as the addition of walking speed. The basic idea is to follow the trip chain of a smartcard on a network line refer to figure 5 and identify the alighting position by looking at the position and time of the next boarding.

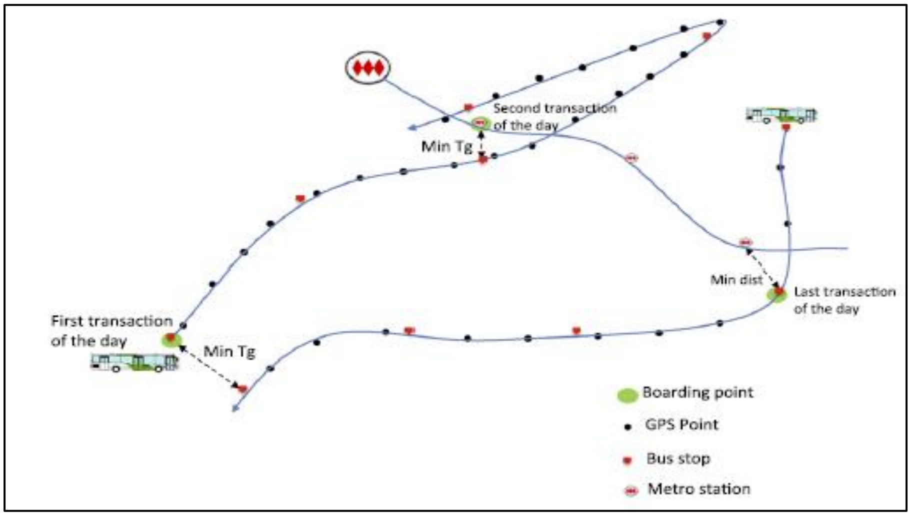


Figure 4-1: Alighting estimation model (Munizaga and Palma, 2012)

4.1.1.2 Implementation of assumptions in Trip-chain method

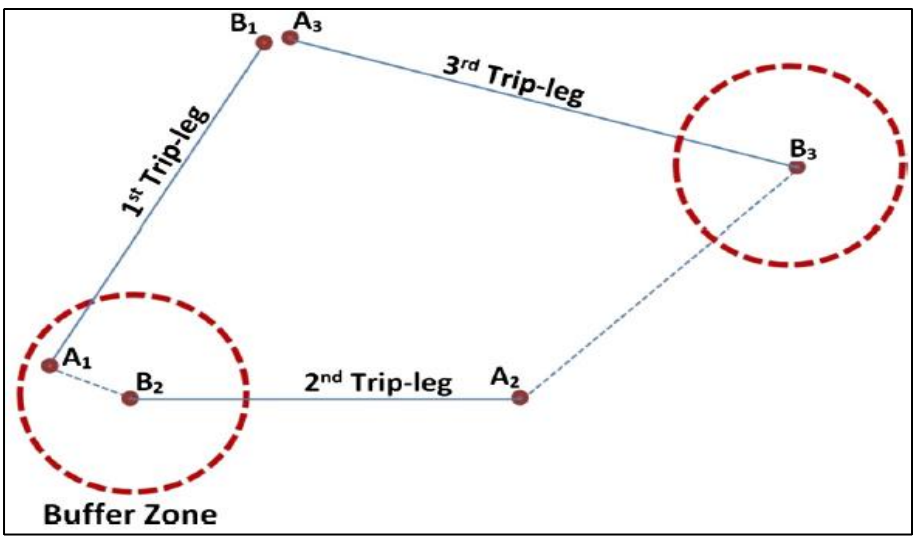


Figure 4-2: Implementation of trip chain method diagram (Alsger et al., 2016)

Consider B1 as the boarding stop and A1 as the alighting stop, in the 1st trip-leg. Following which the passenger walks to the next boarding stop B2 to their second trip leg. A buffer zone is included around A1 and B2 which makes use of all the above assumptions to decide whether a passenger is alighting or transferring. As alighting information is usually not available for other PT modes, such as GABS (which is not the case for the dataset provided by MyCiti in this study, however in an attempt to simulate other PT modes we assume alighting information is not available and validate the estimated O-D matrix with the correct/original O-D matrix). Thus, a theoretical alighting stop A1* is used instead of the actual unknown A1, where A1* is inside the buffer zone and satisfies trip chaining method assumptions. If the time the passenger took to walk to B2 is less than the allowable transfer time, the 1st and 2nd trip legs are chained to create an O-D where B1 is the origin and A2 destination (Alsger et al., 2016).

4.1.1.3 Proposed Trip chaining algorithm

The algorithm aims to basically estimate the alighting stops and chains the trip-legs wherever there is a transfer along the total trip length to form a network level OD matrix, as described in the previous paragraph.

For each Card_Num ID's trip-leg (which could potentially be multiple routes), a search is performed through the schedule of the corresponding route to estimate the best-fitting alighting stop according the assumptions previously mentioned (Alsger et al., 2016). The Trip-chaining algorithm was applied using the following general steps:

i. Pick a smart card ID from the dataset and find the first trip-leg of the corresponding passenger for a given day.
ii. Search through the current route-direction combination to find the nearest stop to the start location of the passenger's next trip-leg (referring to figure 1, the area with the buffer zone).
iii. A) If the distance between the inferred alighting stop and the next boarding stop is greater than the specified allowable walking distance (use 800m), the alighting stop is labeled as the trip's destination.
iii. B) If the distance between the inferred alighting stop and the next boarding stop is less than the specified allowable walking distance, label the trip-leg as a transfer.
iv. If the currently processed trip-leg is the last trip-leg of the smart card holder, check the assumption that the last alighting is within allowable walking distance to the first boarding of that day.
v. The search process continues for all card ID holders for the given day and hence generates a seed OD matrix (Alsger et al, 2016).

The algorithm can be completed using excel with a series of queries as shown in figure 4-3, which illustrates the workflow of queries completed.

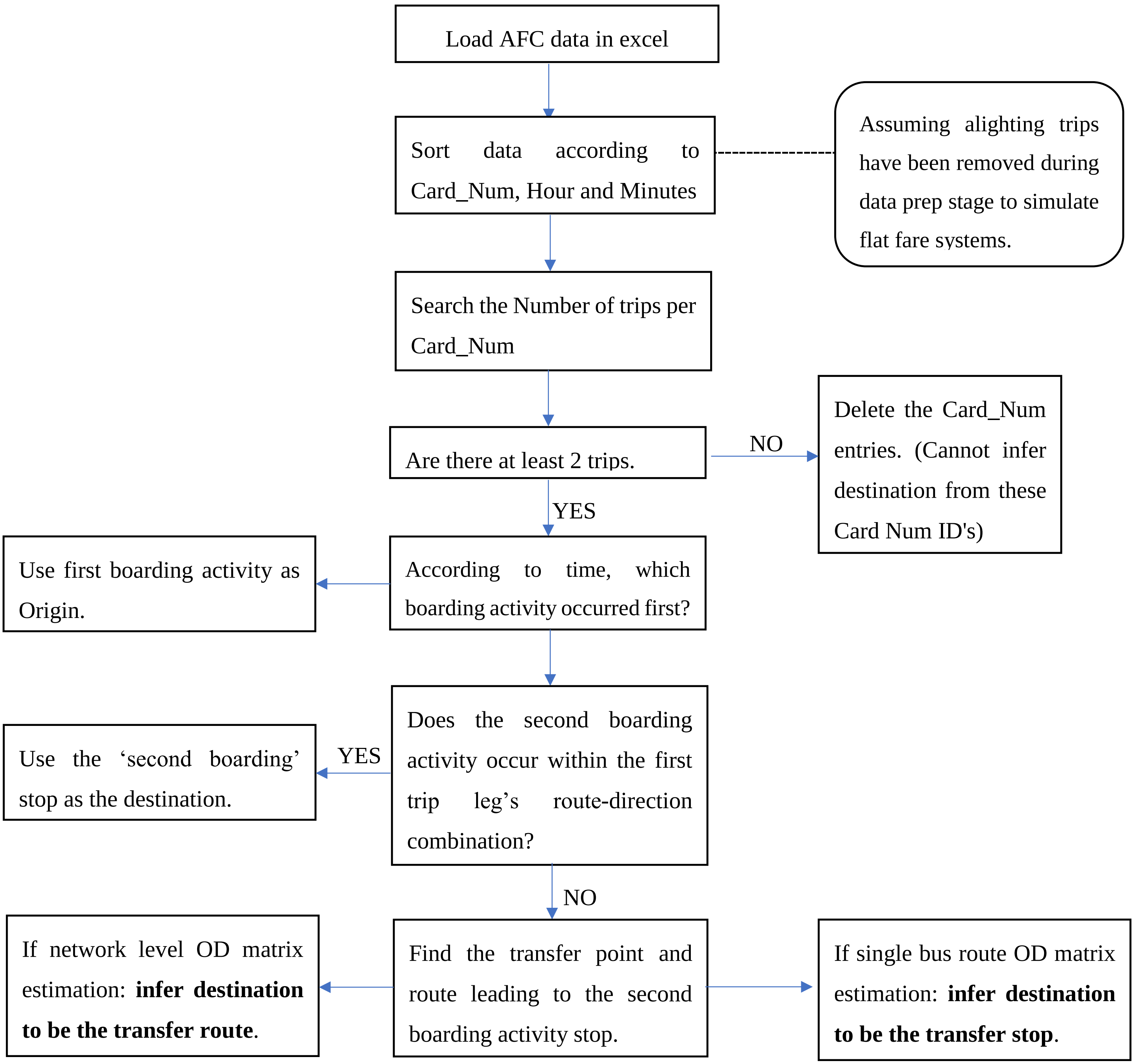


Figure 4-3: Work flow diagram for destination inference

4.3 Single Route OD Matrix Estimation

With the output from the previous steps, the control totals and the seed matrix can be combined to form a single route OD matrix for each route-direction combination. The statistical techniques applied in this thesis to do that is Iterative Proportional Fitting (IPF) or Proportional Distribution.

4.3.1 Iterative proportional fitting

The IPF method has been found to be the most reliable OD estimation technique by several authors. The marginal values will have to be the total counts rather than just a sample of boarding and alighting counts. The total boarding and alighting data for each stop are obtained from the AFC data (which has APC capabilities).

IPF requires the extraction of sample data from the population data to form a seed matrix, generally, this is obtained from the output of the Trip-chaining method. The seed matrix to be put through an iterative process until the total boardings equal the control boarding and total alightings equal the control alightings. (Cui, 2006) argues that IFP method is one of the simpler O-D matrix estimation methods can be applied using Excel however for a large data source use of Structured Query Language (SQL) is recommended. The underlying goal is to find the row and column factors to solve the simultaneous equation of equation 2 and 3 below.

$$\mathrm{T_{ij}} = \mathrm{a_i b_j t_{ij}}, \forall_{i,j} \tag{1}$$

where $\mathrm{T_{ij}}$ is the total passenger flow between origin segment i and destination segment j and

$\mathrm{t_{ij}}$ is seed matrix flow between i and j, $\mathrm{a_i} \geq 1$ is row factor, and $\mathrm{b_j} \geq 1$ is the column factor

$$\textstyle\sum_i T_{ij} = M_j, \forall j,\ \ where\ M_j\ is\ the\ total\ alighting\ count\ for\ segment\ j \tag{2}$$

$$\textstyle\sum_j T_{ij} = M_i, \forall i,\ \ where\ M_i\ is\ the\ total\ boarding\ count\ for\ segment\ i \tag{3}$$

4.4 Network Level Bus Passenger Linked Trip OD Matrix estimation

Network level bus passenger trips can be defined in various ways, mainly through “linked trip” and “unlinked trip”. An “unlinked trip” is a trip between a passenger boarding and alighting on a single bus, thus a passenger moving along one route. A “linked trip” means one or more unlinked trips that occur when a commuter changes buses/ route to move from the origin to the final destination. Therefore, referring figure 4-4 a linked trip would be one where the passenger starts at origin moving along route A to the transfer point to route B and alighting at the destination.

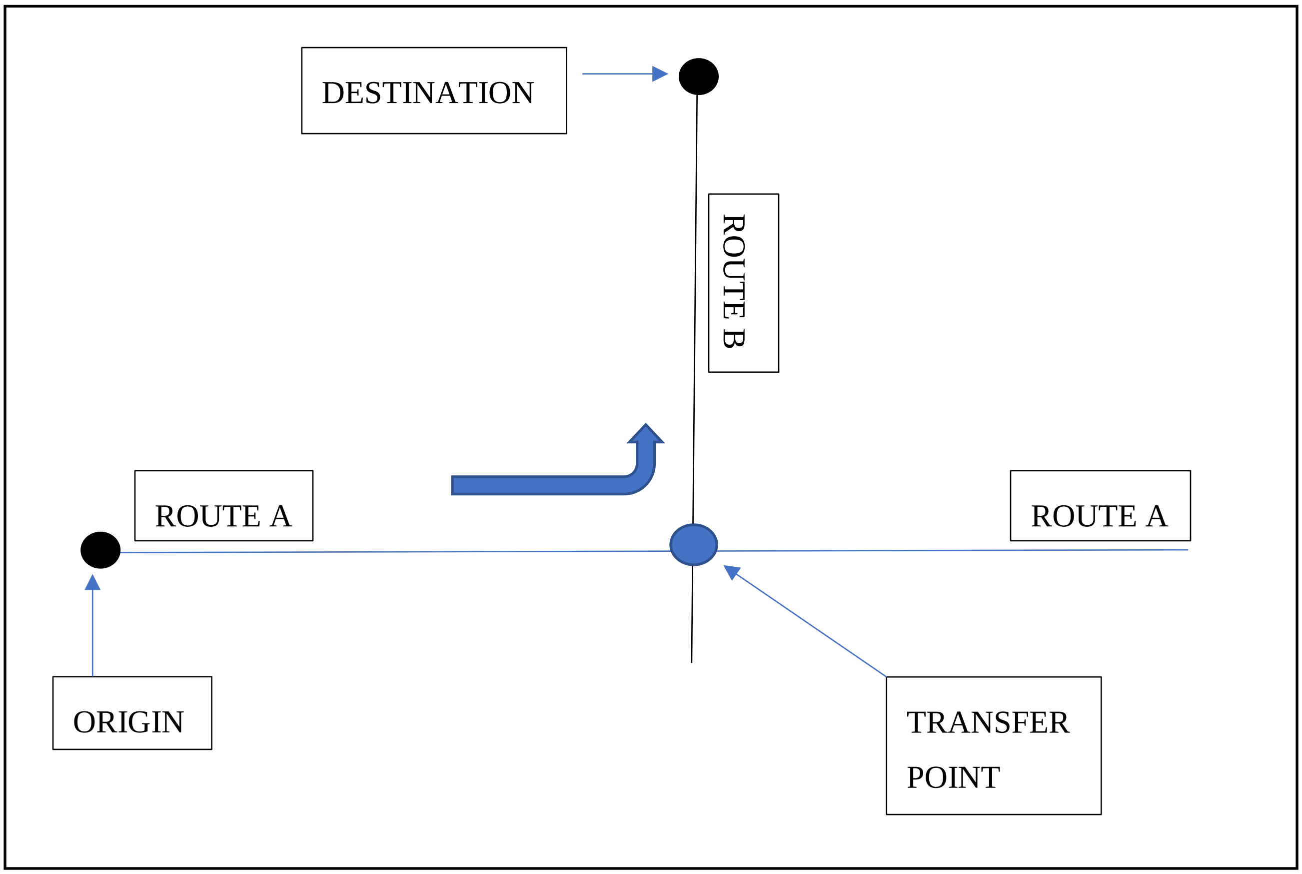


Figure 4-4: Trip transfer inference example

Transfer trips seldom involve more than two unlinked trips, more than that usually results in an OD matrix with a large number of non-structural zeros. For a transfer trip to occur the transfer between the unlinked trip must be within a specified time and distance to each other. In addition, two trips on the same route, will not count as a transfer trip.

Thus, the total transfer counts are a vital aspect of generating a network level OD matrix. The basic steps to generate a network level OD matrix are as follows:

1) Obtain the origin distribution for the first leg trip or the destination distribution for the second leg of the transfer trip.
2) Separate transfer flows from single-bus route matrices.
3) Distribute the transfer flows to eligible origins and destinations.

Figure 4-5 illustrates how the non-transfer OD matrices of single routes and transfer matrices between single-bus routes are combined to form a network level OD matrix. The steps mentioned above are further explained in the following subsections.

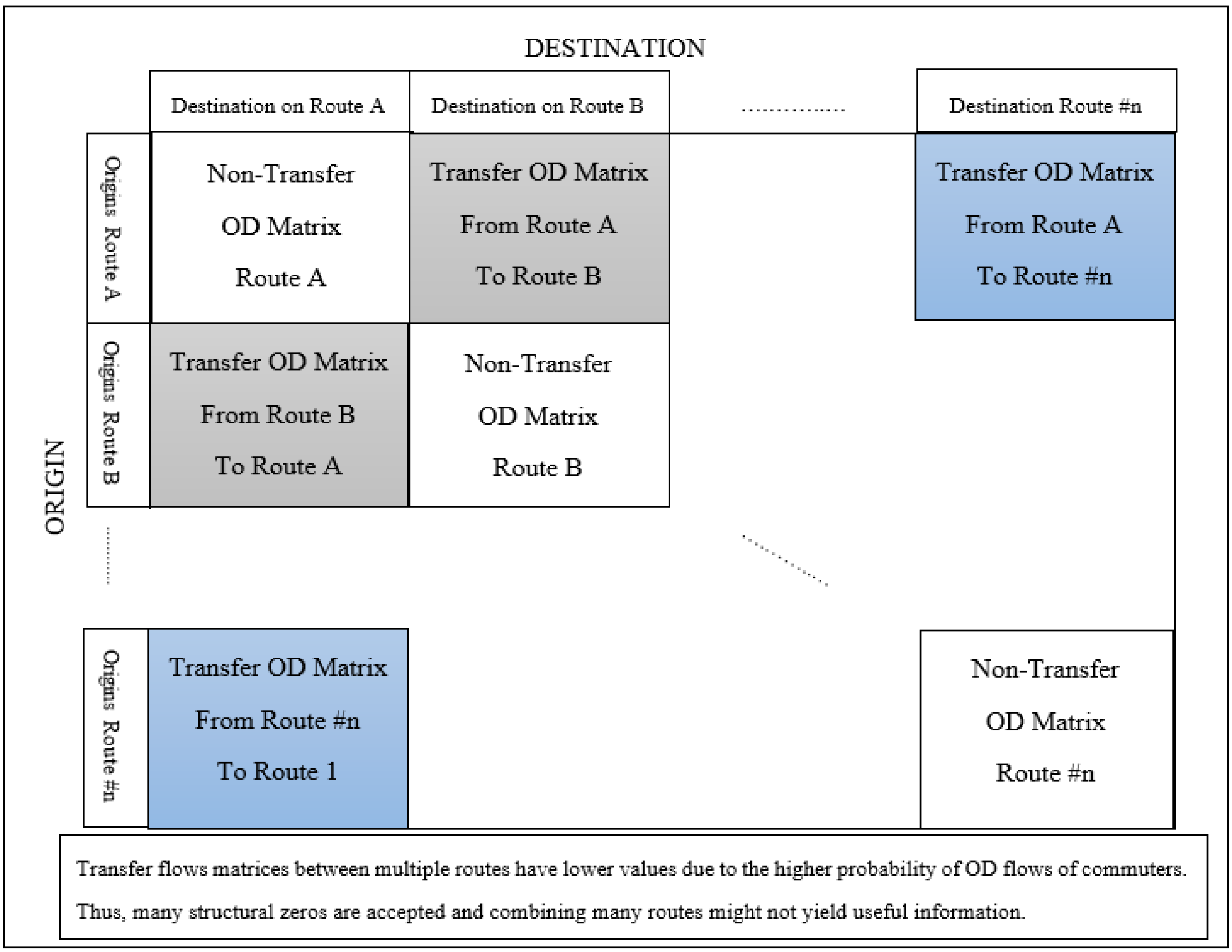


Figure 4-5: General Set up for a Full network OD Matrix

4.5 Total transfer flows

According to Cui (2006) AFC data is the basis for estimating transfer flow totals, and depending on the quality OD data available, it may be required to either scale the aggregate transfer flows from AFC data to obtain the transfer flow totals, or use other data sources to estimate the transfer flow totals.

From Figure 1, it can be observed that the direction of transfer flows is an important in generating the network OD matrix. Thus, the direction of the first leg and second leg of the transfer must be known, this information can be obtained from either AFC data or AVL data. If AVL data is not available the vehicle schedule can be used to obtain the direction.

4.5.1 Origin and destination distribution of the transfer flow

The first leg of the trip consists of a known destination stop which is the corresponding transfer stop, however, the origin of the particular transfer flows is unknown. Depending on the quality of data, it is possible to extract the origin of the transfer flows directly but most transit systems might not have such systems in place. Thus, two other statistical techniques of probability are used to find the transfer flows among eligible origins. The first being the Modified IPF method and the second Proportional Distribution.

4.5.2 Modified IPF method:

This method involves modifying the IPF method to include virtual stops representing a destination stop for transfers to a particular route (Cui, 2006).

Example:

Considering the network illustrated in figure 4-6, it is known that there 14 transfers from Route A east to Route B north and Route C Northeast at stop 3. Figure 4-7 shows the partial OD matrices with the virtual stops for Routes A east, Route B North and Route C Northeast. Thus the 12 transfer flows are split up between the two routes, and to be more precise the flows can be further split between four different stops. Transfers in other direction combinations between the two routes are not considered for this example.

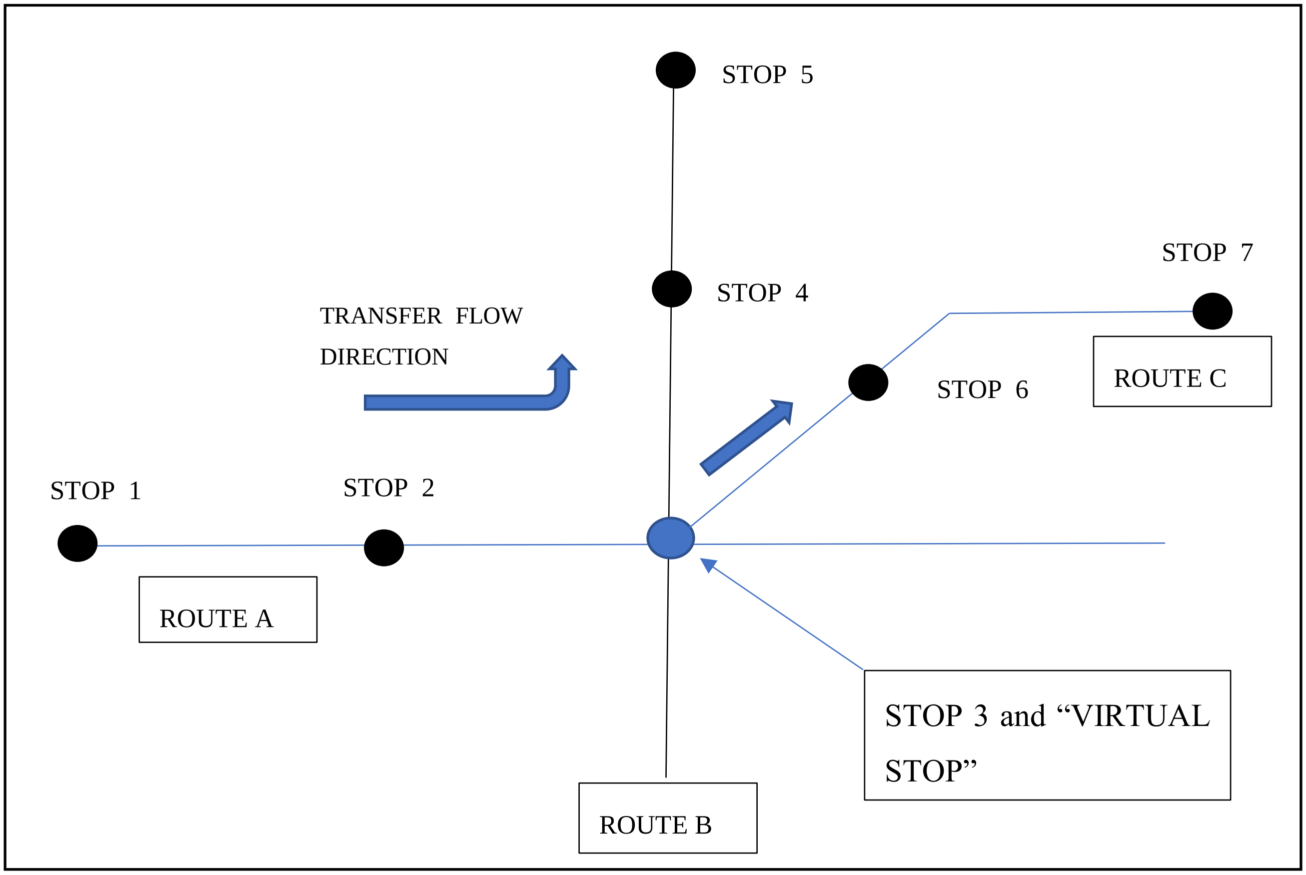


Figure 4-6: Example Schematic Network

The initial step in using the modified IPF is to modify the seed matrix used for single route OD matrix estimation and add the Virtual Stops. For Route A East, two virtual-columns are added representing transfers to Route B North and Route C Northeast as shown in Figure 4-6. The top table in figure 4-7 shows a standard OD matrix (with the IPF results). The OD flows at destination stop 3 contain both transfer flows and non-transfer (flows with stop 3 as the final destination of a passenger).

ROUTE A	Stop 1	Stop 2	Stop 3	Total Boarding
Stop 1	0	5	15	20
Stop 2		0	9	9
Stop 3			0	0
Total Alighting	0	5 *(non-transfer)*	24 *(transfer and non-transfer)*	29

Virtual Stops

Route A Modified IPF	Stop 1	Stop 2	Stop 3	Transfer to Route B	Transfer to Route C	Total Boarding
Stop 1	0	5	6	4	5	**20**
Stop 2		0	4	2	3	**9**
Stop 3				0	0	**0**
Total Alighting	0	5 *(non-transfer)*	10 *(non-transfer)*	6	8	**29**

Total Transfer counts (Distributed from stop 3 counts)

Figure 4-7: Example Schematic Network (OD matrices)

The lower table in figure 4-7 shows an OD matrix with two columns added called Transfer to Route B and Transfer to Route C, which are the virtual stop. Thus, all passengers transferring to Route B North and C Northeast will treat these virtual stops as their destination even though their intermediate destination is the transfer stop 3. Therefore, the transit planner can be aware of the transfer flows from one route to the next, hence forming a partial network OD matrix at route level. The seed matrix and marginal values used for IPF will be adjusted with these virtual stops in mind.

4.5.3 Proportional Distribution

This method aims to distribute transfer flows to eligible stops proportional to the total flow from all origin stops to the transfer stop. Continuing from the previous example, the top table in figure 4-7 shows the partial single route matrices of Route 1. It can be observed that the eligible origin stops are Stop 1 and Stop 2. Therefore, the transfer flows from stop 1 at stop 3 are $14 \times \frac{20}{29} = 9.65$, and the transfer flow from stop 2 at stop 3 are $14 \times \frac{9}{29} = 4.35$. From this point, it is possible to infer the destination of the transfer flows to obtain a transfer flow OD matrix.

4.5.4 Comparison between proportional distribution and modified IPF

Both these methods in the previous example resulted in relatively similar outcomes in terms of transfer flow distribution from the eligible origins. However, the two methods are different in assumptions, which can cause significant differences in large networks. Proportional distribution assumes that passenger OD behavior is the same for non-transfer passengers and transfer passengers. This assumption is not required for modified IPF. Cui (2006) argues the assumptions proportional distribution method makes are not practical. For some networks, the OD pairs for transfer trips may be less than for non-transfer trips. First reason being, if a passenger is going on a trip with two trip legs to reach the destination stop, the passenger will only take the second trip depending on the distance from the transfer point to the destination stop, refer to figure 4-8 for an illustration. Thus, if the utility to walk (hence saving money for example) is greater than the utility to take the bus on the 2nd trip leg, the assumption would not hold.

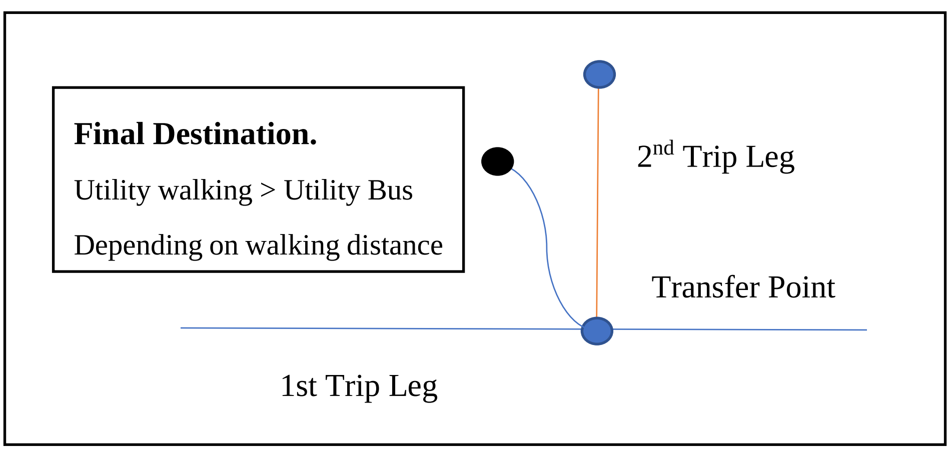


Figure 4-8: Scenario 1 of Proportional Distribution assumptions not holding

The second reason, refer to Figure 4-9, in a portion of a network system sharing two routes. The passengers can freely transfer anywhere between point B and C. However, it would be impractical for a passenger alighting at C to transfer from the current bus they are on which moves from "FBCE" to the bus moving “ABCD”. Thus, in this scenario the would be less OD flow pairs for transfers, hence the assumption would not hold

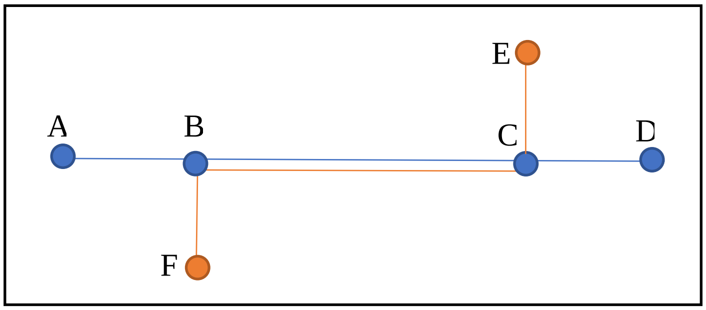


Figure 4-9: Scenario 2 of Proportional Distribution assumptions not holding.

4.5.5 Non-transfer flows

In the process of generating a network level OD matrix, it is required to separate the transfer flows from the non-transfer flows in single route OD matrices, thus leaving trips which can go directly into the network level OD matrix.

The implementation of this step depends on whether the modified IPF method or the proportional distribution method was used. If the proportional distribution method was used, transfer flows are subtracted from the full OD flows to get the non-transfer OD matrix. If the modified IPF method is chosen, then only the origins and destinations involving the virtual stop are removed from the OD matrix to get the non-transfer OD matrix (Cui, 2006).

4.5.6 Transfer flow OD matrices

Continuing from the previous example, showing the distribution of the origin and destination of transfer trips if adequate for numerous application. However, there is an additional step required to obtain a network level passenger trip OD matrix. Thus, in this step, the distribution of the origin of the first leg of the transfer is linked with the destination of the second leg of transfer to obtain the OD flows. The transfer flows are combined with the non-transfer flow single route OD matrices to form the network level OD matrix as shown in the bottom table of figure 4-7.

Table 4-1: Transfer Flow OD Matrix

Destination **(Route A and B)** / Origin **(Route A)**	**Stop 4**	**Stop 5**	**Stop 6**	**Stop 7**	**Total Boarding**
Stop 1	1	3	2	3	9
Stop 2	0	2	1	2	5
Stop 3	0	0	0	0	0
Total Alighting	1	5	3	5	14

A Transfer Flow OD Matrix can be generated as illustrated in table 4-1. With the marginal value (total boarding and alighting counts) and seed matrix, it was possible to estimate the OD flows using the IPF method. The advantage of obtaining such a matrix are outweighed by the analysis which is required to produce them. Since the transfer of flows is usually low and thus the resulting seed matrix tends to be of low quality, with many non-structural zeros. Thus, the benefits of producing a transfer flow OD matrix are usually not enough to justify is use (Cui, 2006). Hence, this these will focus on network level OD matrix estimation at route level instead of at stop level.

Chapter 5

MyCiti Case Study Data Preparation

Data preparation is a crucial step in the aim of estimating OD matrices and validating against actual data. Thus, the MyCiti AFC data preparation will consist of:

1) Generating the actual OD matrices from the MyCiti data. The provided MyCiti data has both alighting and boarding, thus it is possible to simply generate the matrices without implementing trip chaining method to infer destination or apply expansion techniques such as IPF to scale the boarding and alighting counts. OD Matrices generated from the complete dataset will be used as validation OD matrices for the estimated OD matrices.
2) A set of boarding and alighting counts to develop marginal values.
3) Removal of alighting destination from the raw dataset (to simulate PT systems with no alighting validation) and implementation of trip chaining method to infer destinations.
4) Generating a seed matrix from the inferred destinations and boarding OD flows.
5) Obtain transfer flow totals between all route-direction combination in the demonstration corridor.

This chapter will focus on the 4 steps mentioned above and transfer flows will be further discussed in chapter 7 when implementing the network level OD matrix estimation. These data preparation steps will be divided into the following tasks: Validation OD Matrices construction in Section 5.1 boarding stop identification in Section 5.2, alighting stop inference Section 5.3, boarding and alighting counts in Section 5.4.

5.1 Validation OD matrices construction

The following steps detail the procedure that is required to generate the actual OD matrix for single routes. The following steps were completed in excel and they are as follows:

- Isolate the trips of the route of interest, e.g. (referring to figure 5-1) 109 Civic Centre- Hout Bay. This is to find all the Card Num ID's that used the particular route-direction.
- Using the found Card Num ID's, search the entire network (dataset) to see if these card numbers occur elsewhere. This is to find all the potential transfer points.

- After finding the potential transfer point, create virtual stops (for when implementing the modified IPF method) or change actual alighting destinations beyond the route to be the end point of the route of interest. Therefore, referring to figure 5-1, if Stop 4 and Stop 5 had passengers alighting, they would be inferred to alighting at stop 3 which is along Route:109 Houtbay - Civic Centre (for a single OD Matrix).

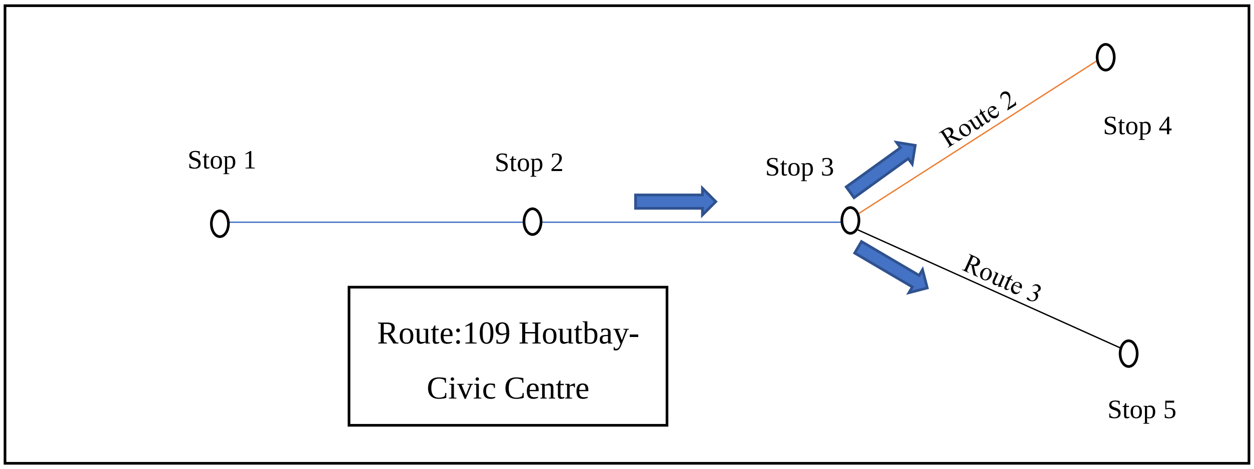


Figure 5-1: Transfer point inference

- Sort the data according to Card_Num, hour and minutes. At this point, it is possible to connect OD flows between stops, and subsequently create an OD matrix for the route. A cleaned sample of the expected final input data, contributing to construction of the validation matrix, for one MyCiti passenger is illustrated in table 5-1.

Table 5-1: Cleaned Sample data for trips of one passenger

CARD_NUM	DEVICE_ID	ROUTE_ID	ROUTE_NAME	STOP_NAME	Transaction Type	Hour	Min.
5759842851889720000	19705818	291	109 Hout Bay - Civic center	Imizamo Yethu	1st Boarding	7	45
5759842851889720000	19689443	291	109 Hout Bay - Civic center	Glen Beach	Alighting	8	20
5759842851889720000	19705337	290	111 Civic center-Hout Bay	Glen Beach	1st Boarding	16	15
5759842851889720000	19705834	290	112 Civic center-Hout Bay	Imizamo Yethu	Alighting	17	5

5.2 Boarding Stop Identification for AFC Trips

In situations, where the AFC data does not contain the boarding information of a passenger Gordon (2012) suggests a methodology to obtain the boardings. The initial step is to link the transaction fare box identification with the bus number, after which it is possible to select the most likely records from the AVL records for that bus. With the AFC and AVL data combined it would be possible to identify the boarding stop (see Gordon 2012 and Cui 2006 for a more detailed description).

5.3 Alighting Stop Inference

The destination stop inference is based on the trip chaining method as described in section 4.2.1. The simple form of trip chaining method identifies the destination stop to be the origin next trip. The next trip may not be on the same trip leg (route-direction), in which case the destination of the first trip is assigned to be the stop on the current route closest to the origin of the next trip. Application of the Trip-chaining method will be further discussed in section 5.3.1 and destination assignment in section 5.3.2.

5.3.1 Trip-chaining method data preparation

Application of Trip-chain algorithm to route 109 Civic Centre-Hout Bay:

1) Isolate the trips of the route of interest, eg (referring to figure 5-1) 109 Civic Centre- Hout Bay. This is to find all the Card Num ID’s that used the route-direction.
2) Using the found Card Num ID’s, search the entire network(dataset) to see if these card numbers occur elsewhere. This is to find all the potential transfer points.
3) Sort the data according to Card Num, hour and minutes.
4) Identify Card Number ID’s that have one trip in the day (the passenger making one transit trip using their MyCiti card), and delete these records as they cannot be used in the destination inference process.
5) Remove all the alighting counts from the dataset (the aim of the thesis is to test whether the algorithm can be used for PT systems with no alighting validation).
6) Isolate the first trip and the last trip of a particular Card Num ID, assign the destination of the first trip to be the origin of the last trip in a day. Table 2 illustrates Card Num ID's of two passengers with the Transaction Type column indicating boarding for all taps, as would be

expected in a flat fare system. For each passenger, their last transaction (boarding) of the day is found and assigned to be the alighting destination of the first trip of the day. The result of the process is illustrated in table 3.

Table 5-2: Sample data illustrating destination inference using Trip-chaining

CARD_NUM	ROUTE_NAME	STOP_NAME	TRANSACTION_TYPE	HOURS	MIN
18426375260815000000	D03 Mitchells Plain - Civic Centre	Marguerite	1st boarding	5	45
18426375260815000000	9999 (COCT)	Civic Centre	*1st boarding*	15	58
18258482901415500000	D03 Mitchells Plain - Civic Centre	Sesame	1st boarding	6	53
18258482901415500000	9999 (COCT)	Civic Centre	*1st boarding*	16	58

Table 5-3: Sample data illustrating destination inference using Trip-chaining (1)

CARD_NUM	ROUTE_NAME	STOP_NAME	TRANSACTION_TYPE	HOURS	MIN
18426375260815000000	D03 Mitchells Plain - Civic Centre	Marguerite	1st boarding	5	45
18426375260815000000	9999 (COCT)	Civic Centre	*Alighting*	15	58
18258482901415500000	D03 Mitchells Plain - Civic Centre	Sesame	1st boarding	6	53
18258482901415500000	9999 (COCT)	Civic Centre	*Alighting*	16	58

Refer to figure 4-3 for workflow diagram illustrating excel queries for full destination inference.

5.3.2 Destination reassignment

Different scenarios are present when inferring the destination of a trip, the next trip can be on the same route just in the opposite direction or on a different route (Cui, 2006). In such cases, it is required to reassign the destination

If the next trip is on a different bus route, the stop on the route interest closest to the boarding stop of the next trip is assigned as the alighting stop. If the next trip is a trip on a different bus route-direction, it can either be a walking distance away from the current route (Example A in Figure 5-2), be parallel to the current route (Example B in Figure 5-2), or intersect the current bus route (Example C in Figure 5-2).

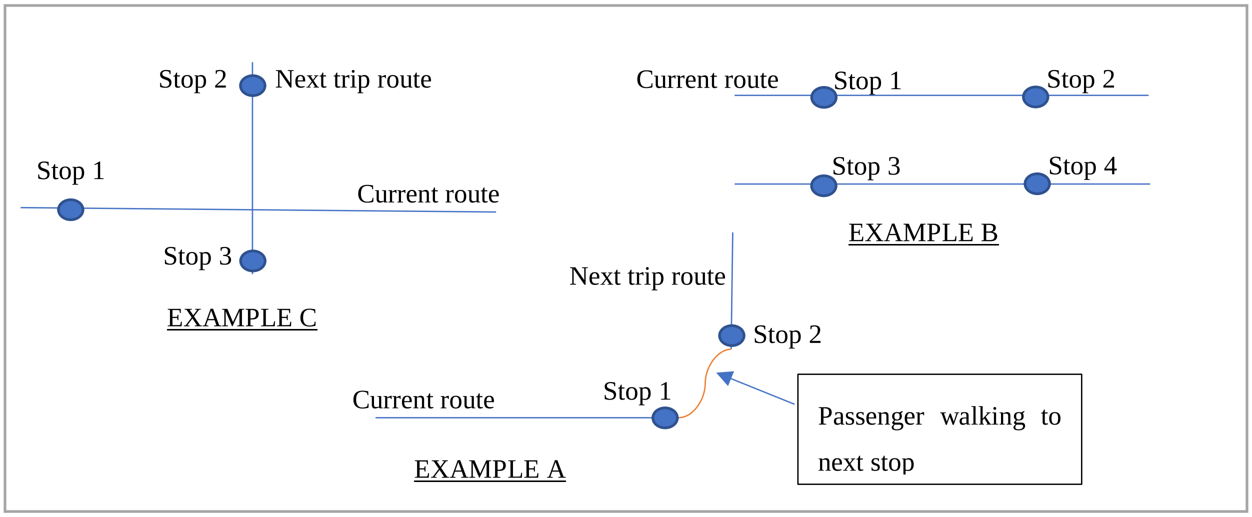


Figure 5-2: Destination reassignment scenarios

In Example A, if stop 2 is within a walkable distance (assumed to be 800m) the stop closest to the next route, stop 1, is inferred to be the alighting stop (barry, 2012). Considering Example B, whether the boarding stops of the next trip is stop 3 or stop 4 is relevant, since the closest stop (considering stop 1 and stop 2) to the boarding location will be inferred to be the alighting stop. For Example C, knowing the routes is the critical information, whether the boarding stop is at stop 2 or stop 3 does not matter stop 1 is clearly the closest stop and thus would be inferred as the destination stop.

At this point, there is a complete set of boarding and alighting data, which can be aggregated to form OD flows and consequently forming the seed matrix.

5.3.3 Seed matrix formation

The MyCiti AFC data consisted of 164 000 taps on the 9th of May 2017, consisting of 72 500 trips. The Trip-chaining method managed to infer 90% of the trips successfully according to Gordon (2012). Approximately 10% of the trips only had a single trip in the day, thus were removed from the data as it would be impossible to infer their destination using the algorithm. Cui (2006) states that the success rate of destination reassignment is 79%. At the end, the seed matrix represents approximate 71% (90% × 79%) of all MyCiti trips. Table 5-4 illustrates the seed matrix for Route DO1 Kuyasa - Civic Center, the is a total of 1706 OD flows occurring on the route. Thus, this seed matrix represents approximately 67% of trips for the route-direction combination. A lower sampling rate can be attributed to routes that have a low number of regular passenger, who use MyCiti for more than one trip a day.

Table 5-4: Seed matrix for Route DO1 Kuyasa- Civic Center

O/D	**Kuyasa Rail Station**	**Lindela**	**Dibana**	**Tutu**	**D Nyembe**	**Steve Biko**	**Charles Mokoena**	**Vuyani Taxi Rank**	**Civic Centre**	**Boardings**
Kuyasa Rail Station	1	0	0	0	0	1	0	2	368	372
Lindela		2	0	1	1	1	0	0	103	108
Dibana			0	0	0	0	0	0	116	116
Tutu				0	0	0	0	0	209	209
D Nyembe					0	0	0	0	13	13
Steve Biko						0	0	0	106	106
Charles Mokoena							0	0	113	113
Vuyani Taxi Rank								0	96	96
Civic Centre									0	0
Alighting	1	2	0	1	1	2	0	2	1124	1133

5.4 Boarding and Alighting Counts

The Automatic Passenger Counts (APC) boarding and alighting counts need to be adjusted so that the total boarding and total alighting's over the entire length of a one-way bus trip are equal. Thus, no passengers are on the bus before the trip start and no passengers are left on the bus at the end of the trip.

The boarding and alighting counts need to be aggregated at stop level and if required adjust the counts to ensure their totals match, by increasing the lower number in the seed matrix. Therefore, the adjustments should be proportional to either the boarding count or the alighting count. Cui (2006) state that this is a simple way to correct errors as APC equipment more likely to undercount. In addition, this ensures consistency as every boarding must correspond to an alighting.

The boarding and alighting counts can be obtained directly from the APC, by isolating each stop occurring on the route-direction combination of interest and aggregating the number of boardings and alightings. Table 5 illustrates the APC boarding and alighting counts for route DO1 Kuyasa-Civic Centre.

Table 5-5: boarding and alighting counts

Stops	**Kuyasa Rail Station**	**Lindela**	**Dibana**	**Tutu**	**D Nyembe**	**Steve Biko**	**Charles Mokoena**	**Vuyani Taxi Rank**	**Civic Centre**
Total Boardings	543	184	156	322	27	166	164	146	0
Total Alightings	4	0	0	2	0	3	2	5	1693

Chapter 6

MyCiti Single Bus Route OD Matrix Estimation

In this Chapter, the Demonstration Corridor is presented in Section 6.1 and actual (validation) OD matrix is presented in Section 6.2. Subsequently, the result of the Iterative Proportional Fitting (IPF) method for single route OD matrix estimation is presented in Section 6.3. The results from the estimated OD matrices are then validated against the Validation OD matrices in Section 6.4.

6.1 Demonstration Corridor

To demonstrate the methodology of the network level OD matrix estimation, ten single direction bus routes are used, most connected by Civic Center as the transfer point at shown in Figure 6-1.

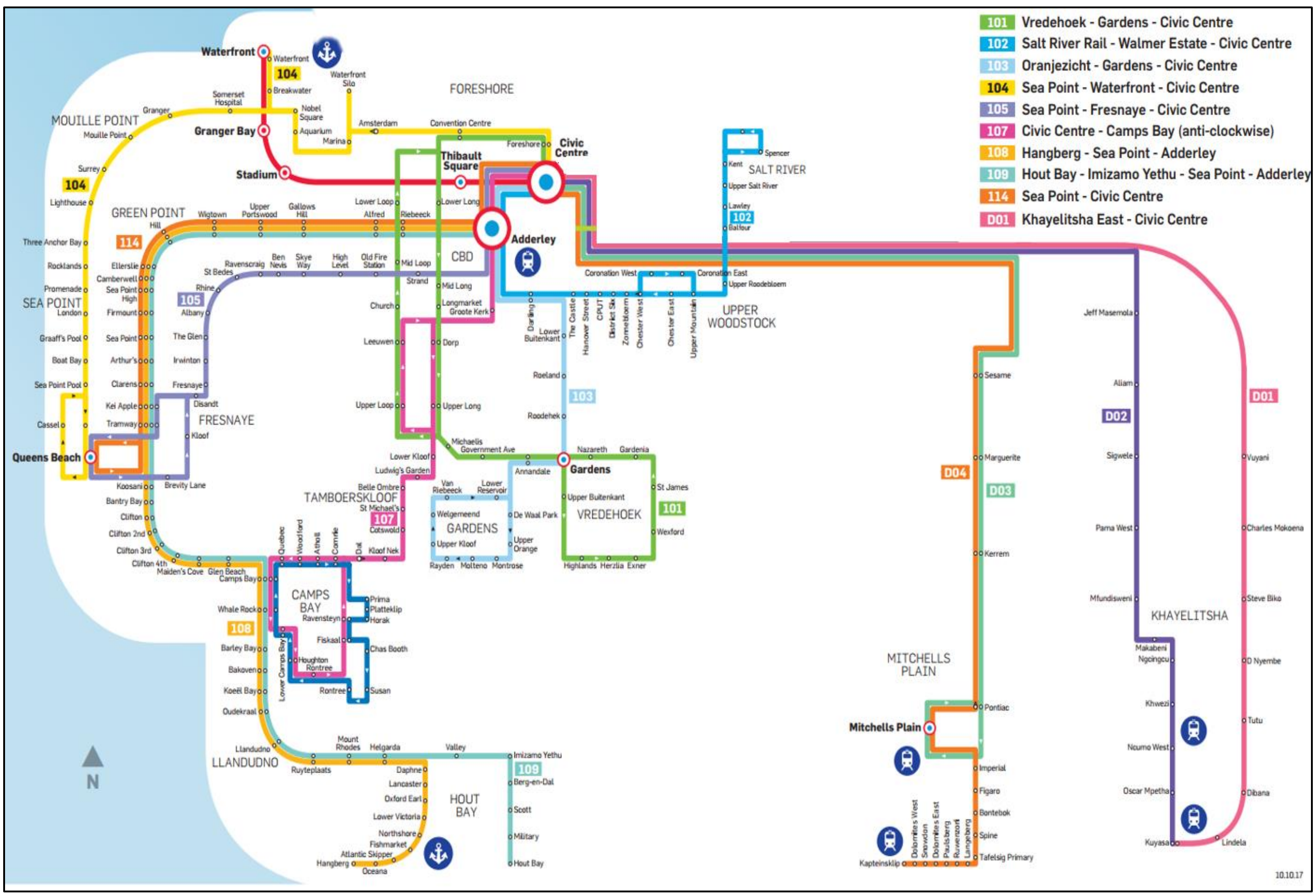


Figure 6-1: Demonstration Corridor containing partial MyCiti Network

Analysis of the demonstration corridor will mainly focus on single route OD matrices and only transfer flows from route DO1 to the other nine routes will be demonstrated (in Chapter 7).

6.2 Validation OD matrix

As described in section 5.1 the validation OD matrices are generated directly from the MyCiti data, without any statistical techniques used to scale the data. Thus, the validation OD Matrix represents the actual OD flows at stop level and can be used to validate the accuracy of the proposed procedure of OD matrix estimation presented in this thesis. Table 6-1, illustrates the actual single route OD matrix for route DO1 Kuyasa- Civic Center.

Table 6-1: Actual OD Matrix for Route DO1 Kuyasa to Civic Centre

	Kuyasa Rail Station	***Lindela***	***Dibana***	***Tutu***	***D Nyembe***	***Steve Biko***	***Charles Mokoena***	***Vuyani Taxi Rank***	***Civic Centre***	***Boardings***
Kuyasa Rail Station	*4*	*0*	*0*	*0*	*0*	*1*	*1*	*3*	*534*	*543*
Lindela		*0*	*0*	*2*	*0*	*2*	*0*	*1*	*179*	*184*
Dibana			*0*	*0*	*0*	*0*	*1*	*0*	*155*	*156*
Tutu				*0*	*0*	*0*	*0*	*1*	*322*	*323*
D Nyembe					*0*	*0*	*0*	*0*	*27*	*27*
Steve Biko						*0*	*0*	*0*	*166*	*166*
Charles Mokoena							*0*	*0*	*164*	*164*
Vuyani Taxi Rank								*0*	*146*	*146*
Civic Centre									*0*	*0*
Alighting	*4*	*0*	*0*	*2*	*0*	*3*	*2*	*5*	*1693*	*1709*

6.3 Iterative Proportional Fitting

The application of IPF was completed using Microsoft Excel, in which the column and row factors were obtained. All the column factors and row factors were greater than 1. Since the seed matrix is based on a sample of total passenger trips the sum of all cells of a seed matrix over a row (or column) should always be less than the total boarding count (or alighting count) for that row (or column). Cases were the row factor or column factor was less than 1 indicated inconsistency between the total boarding or alighting counts and the seed matrix. In such a case, Cui (2006) suggests adding constraints to deal with the inconsistencies such that $a_i = \prod_k a_i^k \geq 1$, thus in situation where the factor is less than 1 it is set to equal 1 instead.

In Table 6-2, the resulting column and row factors for Route D01 Kuyasa - Civic Center are shown. The OD flows in the seed matrix (table 6-1) are scaled by these factors resulting in the estimated single bus route OD matrix.

Table 6-2: IPF OD Matrix for DO1 Kuyasa to Civic Centre

O/D	**Kuyasa Rail Station**	**Lindela**	**Dibana**	**Tutu**	**D Nyembe**	**Steve Biko**	**Charles Mokoena**	**Vuyani Taxi Rank**	**Civic Centre**	**Boardings**	**row factor**
Kuyasa Rail Station	4	0	0	0	0	1.34	0	5	532.6	543	1.45
Lindela		0	0	2	0	1.64	0	0	180.4	184	1.75
Dibana			0	0	0	0	0	0	156	156	1.35
Tutu				0	0	0	0	0	323	323	1.55
D Nyembe					0	0	0	0	27	27	2.08
Steve Biko						0	0	0	166	166	1.57
Charles Mokoena							0	0	164	164	1.45
Vuyani Taxi Rank								0	146	146	1.52
Civic Centre									0	0	1
Alighting	4	0	0	2	0	3	0	5	1695	**1709**	
Column factor	2.76	1.00	1.00	1.14	0.00	0.94	0.00	1.72	1.00		

The factors converged at the 7th iteration and it can be observed that most of the column and row factors have values greater than 1, and those stops with relatively low boarding or alighting counts have values equal to 1 (due to the additional constraints).

Majority of destinations were inferred to be at Civic Center since it is the first stop from the residential areas to the CBD area. In addition, Civic Center is the transfer point for several other routes. The OD flows for other areas are relatively low indicating that the travel within the residential areas is limited. The result is to be expected as there is not much economic activity within townships, thus people must leave the area to reach the CBD or higher income suburban areas for work or better schools. The OD matrices of other route combinations can be obtained; however, they are not shown as the purpose of this thesis is to demonstrate the purposed procedure.

With the complete AFC system of MyCiti, it is possible to validate the estimated OD matrices with the actual OD matrices generated directly from MyCiti data. Thus, table 6-3 compares the results of the estimated OD matrix (in table 6-2) to the actual OD matrix (in table 6-3). To better represent the comparison between these two results, the OD passenger flows as percentages of total passenger flows for route DO1 Kuyasa- Civic Center is shown in table 6-3. The results shown in the brackets are the OD flows obtained from the validation OD matrix.

6.4 Validation of the Estimation Results

Table 6-3 below shows the relative weighting of the estimated OD flow pair to the total counts versus the Validation OD flow pair to the total counts.

Table 6-3: Relative comparison of the estimated OD flows against actual OD flows

O/D	Kuyasa Rail Station	Lindela	Dibana	Tutu	D Nyembe	Steve Biko	Charles Mokoena	Vuyani Taxi Rank	Civic Centre	Boardings
Kuyasa Rail Station	0.23% (0.23%)	0% (0%)	0% (0%)	0% (0%)	0% (0%)	0.08% (0.06%)	0% (0.06%)	0.29% (0.18%)	31.17% (31.25%)	31.77% (31.77%)
Lindela		0% (0%)	0% (0%)	0.12% 0.12%	0% (0%)	0.10% 0.12%	0% (0%)	0% (0.06%)	10.55% (10.47%)	10.77% (10.77%)
Dibana			0% (0%)	0% (0%)	0% (0%)	0% (0%)	0% (0.06%)	0% (0%)	9.13% (9.07%)	9.13% (9.13%)
Tutu				0% (0%)	0% (0%)	0% (0%)	0% (0%)	0% (0%)	18.9% (18.84)	18.9% (18.9%)
D Nyembe					0% (0%)	0% (0%)	0% (0%)	0% (0%)	1.58% (1.58%)	1.58% (1.58%)
Steve Biko						0% (0%)	0% (0%)	0% (0%)	9.71% (9.71%)	9.71% (9.71%)
Charles Mokoena							0% (0%)	0% (0%)	9.6% (9.59%)	9.6% (9.59%)
Vuyani Taxi Rank								0% (0%)	8.54% (8.54%)	8.54% (8.54%)
Civic Centre									0% (0%)	0% (0%)
Alighting	0.23% (0.23%)	0% (0%)	0% (0%)	0.12% 0.12%	0% (0%)	0.18% 0%	0% 0.12%	0.29% (0.30%)	99.18% (99.06%)	100% (100%)

Note: brackets are results from actual OD matrix. Rounding error may be present.

As can be seen on table 6-3 the OD flow percentages are similar with no significant difference to note. The OD flow with the biggest difference is 0.11% originating from Kuyasa Rail Station to Vuyani Taxi Rank, which is potentially the result of the Trip-chaining method over-estimating the OD flow. However, stop Charles Mokoena Trip-chaining underestimated the OD flows for the stop. A possible reason for the inconsistencies is because OD flows of regular passenger are more likely to be inferred than once of passengers.

To check the standard deviation of the residuals for the route of interest, a Root Mean Square Error (RMSE) calculation was completed for all route-directions in the demonstration corridor. RMSE is a frequently used measure of the difference between values predicted by a model and the values actually observed from the population that is being modelled (statisticshowto, 2017). The RMSE formula used is as illustrated in equation for each OD flow pair.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}} \tag{4}$$

The RMSE for route DO1 Kuyasa to Civic Center is 50.69 trips for the seed matrix and for the estimated OD matrix 0.796 trips, which can be rounded up to 1 trips. Thus, residuals for the route are found to have an average spread of 1 trip from the validation data, illustrating that the procedure is a relatively good predicter of OD flows of the population. The spread of the seed matrix is high at 50.69 trips, however, that is expected as the number of trips were not inferred for and the expansion factors have not been applied yet.

The accuracy of the estimated matrix is potentially due to the high quality of control totals at each stop and high sampling rate provided by the Trip-chaining method. Traditionally the control totals for each stop are usually estimated values, found from on-board surveys which are scaled up according to the particular situation. However, with the use of the AFC data, it is possible to obtain high-quality boarding and alighting counts for each stop and a good seed matrix. The accuracy of the MyCiti Automatic Data Collection systems is unique to MyCiti and is yet to be installed for PT systems such as GABS. These less advanced PT systems may require the more traditional approach of calculating control totals at each stop. Thus, using IPF can result in accurate information on the relative weight of OD flows, but the absolute value of the OD flows will depend on the accuracy of the total estimated boarding and alighting counts.

Chapter 7

Network Level OD Matrix Estimation

The estimation of a partial network level OD matrix for MyCiti bus system closely follows the general procedures proposed in Section 4.4. This Chapter will discuss the implementation and the results using the procedures previously described, thus Section 7.1 will discuss the process of obtaining transfer flow total for each route-direction combination. Section 7.2 and 7.3 will describe the distribution of the origin and destinations of the transfer flow totals obtained by using both the modified IPF and proportional distribution methods. Subsequently, Section 7.4 will validate the estimated OD matrices of each method against the actual (validation) OD matrix. Finally, Section 7.5 will briefly discuss transfer flow OD matrices.

7.1 Transfer Flow Totals

The basic goal of transfer flow assignment is to find passengers that have more than one trip leg for a single trip. After identifying these passengers, the transfers are assigned to a particular transfer point according to the location of the second trip leg. This information is obtained from the AFC dataset.

7.1.1 Transfer trips in AFC data

The transfer trips made using MyCiti cards are not specifically marked as transfer trips in the AFC data. Therefore, there was a need to have a definition of transfer trips. A transfer trip can be defined as a “series of passenger trips involving more than one unlinked trips, and all unlinked trips that are part of this series of transfer trips must start within two hours of each other” (Cui, 2006).

Using the AFC data, it is possible to find the Card_Num ID's of all the passengers using a specific route-direction combination, by filtering out all other routes. Subsequently, using the obtained list of Card_Num ID's, locations where the Card_Num ID's were used outside the route of interest are searched on the entire dataset and noted. These locations are cumulated according to route which they occur on. Thus, it is possible to find the transfer point along the route, for the demonstration corridor presented in this thesis Civic Center is the transfer point for majority of the routes. In addition, for this analysis, transfer trips outside the demonstration corridor were eliminated.

7.2 Proportional Distribution

The next logical step is to determine the distribution of origins and destinations for the transfer flow totals found in the previous section. This method as discussed in section 4.5.3, Proportional distribution method distributes the origins or destinations proportionally based on the total flows for the OD pairs involving the transfer stops.

The implementation of proportional distribution is demonstrated by taking the transfer flows from Route D01 Kuyasa-Civic and allocating them the nine possible transfer routes, as illustrated in the demonstration corridor. Only those with ten or greater passengers are considered, consequently one route did not meet the requirement, which left eight possible transfer routes. The total OD flows ending at the transfer point for Route DO1 Kuyasa- Civic Center is shown in Table 7-1 and the total OD flows originating from the transfer point is illustrated in Table 7-2. The OD flows with destinations at the transfer point for Route DO1 Kuyasa – Civic Center are illustrated in Table 7-3, and the OD flows with origins at the transfer point for the eight possible routes are shown in Table 7-4.

Table 7-1: Total OD flows Ending at the Transfer Point 12

		To
		Civic Centre
From	Kuyasa Rail Station	534
	Lindela	179
	Dibana	155
	Tutu	322
	D Nyembe	27
	Steve Biko	166
	Charles Mokoena	164
	Vuyani Taxi Rank	146

Table 7-2: Total OD Flows Originating from the Transfer Point

From	To: Destination at Route Level							
Transfer Point	Route 101 (Civic Centre - Gardens - Vredehoek)	Route 102 (Civic Centre - Walmer Est - Salt Riv)	Route 103 (103 Civic Centre - Oranjezicht)	Route 104 (Civic Centre - Waterfront - Sea Poin)	Route 105 (Civic Centre - Fresnaye - Sea Point)	Route 107 (Civic Centre - Camps Bay)	Route 108 (Civic Centre - Hangberg)	Route 114 (Civic Centre - Sea Point)
Civic Center	501	2315	989	384	815	2113	2077	1215

There are 603 daily transfer flows for Route DO1 Kuyasa- Civic Centre distributed to the 8 routes, which is distributed according to the above OD flows. The results for transfer flows are shown in Table 7-3 and 7-4.

Table 7-3: Transfer-only Flows with Origins Distributed

		To
		Civic Centre
From	Kuyasa Rail Station	190
	Lindela	64
	Dibana	55
	Tutu	115
	D Nyembe	10
	Steve Biko	59
	Charles Mokoena	58
	Vuyani Taxi Rank	52

Table 7-4: Transfer-only Flows with Destinations Distributed

From	**To: Destination at Route Level**							
Transfer Point	Route 101 (Civic Centre - Gardens - Vredehoek)	Route 102 (Civic Centre - Walmer Est - Salt Riv)	Route 103 (103 Civic Centre - Oranjezicht)	Route 104 (Civic Centre - Waterfront - Sea Poin)	Route (105 Civic Centre - Fresnaye - Sea Point)	Route (107 Civic Centre - Camps Bay)	Route 108 (Civic Centre - Hangberg)	Route 114 (Civic Centre - Sea Point)
Civic Center	29	134	57	22	47	123	120	71

7.3 Modified IPF

The modified IPF involves modifying the seed matrix and the boarding and alighting totals, by adding virtual stops for each destination beyond the transfer point.

7.3.1 Modifying the seed matrix

The initial step in implementing the modified IPF method is to modify the seed matrix used for a single-bus route OD matrix estimation. For Route DO1 Kuyasa- Civic Centre, eight virtual-columns and eight virtual-rows are added representing transfer to eight different routes as shown in Table 7-5. Additionally, the seed matrices for all the route-direction combinations in the demonstration corridor need to be modified by adding the virtual-columns.

Table 7-5: Modified Route DO1 Kuyasa-Civic Center with Virtual-columns and Virtual-rows representing transfers to other routes

	Kuyasa Rail Station	**Lindela**	**Dibana**	**Tutu**	**D Nyembe**	**Steve Biko**	**Charles Mokoena**	**Vuyani Taxi Rank**	**Civic Centre**	**Route 101 (Civic Centre - Gardens - Vredehoek)**	**Route 102 (Civic Centre - Walmer Est - Salt Riv)**	**Route 103 (103 Civic Centre - Oranjezicht)**	**Route 104 (Civic Centre - Waterfront - Sea Poin)**	**Route (105 Civic Centre - Fresnaye - Sea Point)**	**Route (107 Civic Centre - Camps Bay)**	**Route 108 (Civic Centre - Hangberg)**	**Route 114 (Civic Centre - Sea Point)**	**Boardings**
Kuyasa Rail Stat.	1	0	0	0	0	1	0	2	313	4	12	5	6	7	9	8	16	384
Lindela		2	0	1	1	1	0	0	91	0	2	4	2	0	1	2	4	111
Dibana			2	0	0	0	0	0	93	2	1	6	1	2	4	1	7	119
Tutu				0	0	0	0	0	178	2	6	6	4	3	7	3	6	215
D’Nyembe					0	0	0	0	12	0	0	0	0	0	0	0	1	13
Steve Biko						0	0	0	79	5	3	4	3	3	6	1	4	108
Charles Mokoena							0	0	94	1	3	3	1	2	3	5	3	115
Vuyani Taxi(rank)								0	78	2	1	3	2	3	4	2	2	97
Civic Centre									0	0	4	2	1	0	0	2	4	13
Alighting	1	2	2	1	1	2	0	2	938	16	32	33	20	20	34	24	47	1175

Comparing the seed matrix in table 7-5 to the seed matrix in table 5-4, the total boardings are similar. The main difference between the two seed matrices is the distribution of the of transfer flows from the transfer stop (Civic Centre).

7.3.2 Modifying boarding and alighting totals

The boarding and alighting totals are modified to include the total transfers in the virtual-columns and virtual-rows, and the boarding and alighting totals for transfer stops are modified to exclude the total transfers to become the total non-transfer boarding and alighting counts as shown in Table 7-6.

Table 7-6: Boarding and alighting Counts with Virtual-entries

	Kuyasa Rail Station	**Lindela**	**Dibana**	**Tutu**	**D Nyembe**	**Steve Biko**	**Charles Mokoena**	**Vuyani Taxi Rank**	**Civic Centre**	**Route 101**	**Route 102**	**Route 103**	**Route 104**	**Route 105**	**Route 107**	**Route 108**	**Route 114**
Boarding	534	179	155	322	27	166	164	146	15								
Alighting	4	0	0	2	0	3	2	5	1105	30	113	71	13	62	140	74	100

7.3.3 Modified IPF results

With the control totals and seed matrix value set up, it is possible to apply the IPF procedure to obtain the modified single-bus route OD matrices. The application of IPF was complete using Microsoft Excel, in which the column and row factors were obtained. In order to decrease the presence of zero OD flows, the transfer trips were generated at route level. The modified IPF OD matrix converged at the 13th iteration, all the row, and column factors were greater or equal to one. The resulting OD matrix from IPF method is shown in table 7-7.

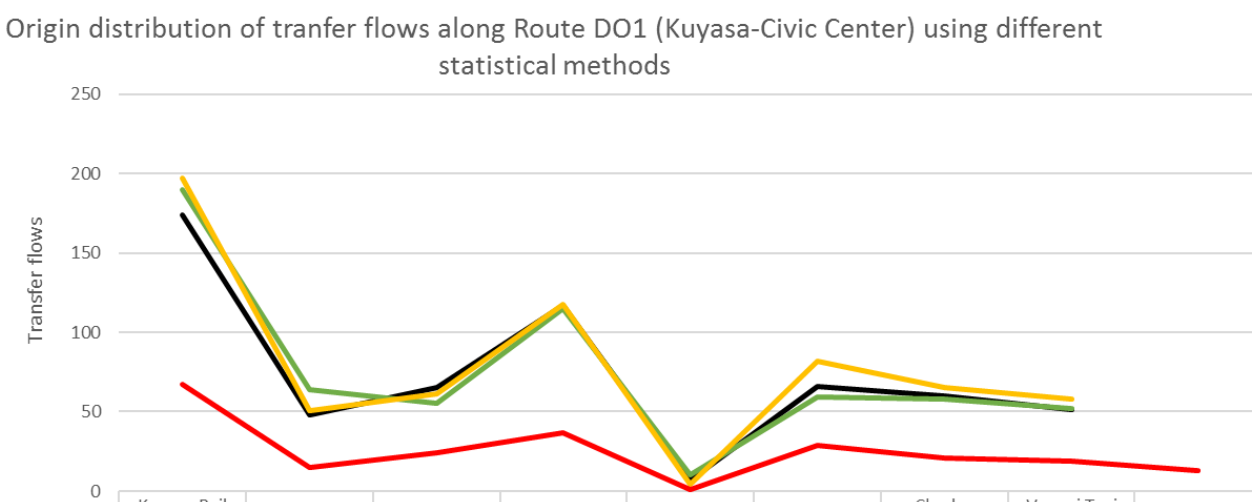
Table 7-7: Modified IPF Method with Virtual Columns and Rows

	Kuyasa Rail Station	Lindela	Dibana	Tutu	D Nyembe	Steve Biko	Charles Mokoena	Vuyani Taxi Rank	Civic Centre	Route 101 (Civic Centre - Gardens - Vredehoek)	Route 102 (Civic Centre - Walmer Est - Salt Riv)	Route 103 (103 Civic Centre - Oranjezicht)	Route 104 (Civic Centre - Waterfront - Sea Poin)	Route (105 Civic Centre - Fresnaye - Sea Point)	Route (107 Civic Centre - Camps Bay)	Route 108 (Civic Centre - Hangberg)	Route 114 (Civic Centre - Sea Point)	Boardings	Row Factor
Kuyasa Rail Station	4	0	0	0	0	1	0	5	332	8	52	11	4	22	37	27	36	539	1.32
Lindela		0	0	1	0	1	0	0	127	0	11	12	2	0	5	9	12	180	1.73
Dibana			0	0	0	0	0	0	95	4	4	13	1	6	16	3	15	156	1.27
Tutu				0	0	0	0	0	207	4	28	15	3	10	31	11	15	325	1.44
D Nyembe					0	0	0	0	23	0	0	0	0	0	0	0	4	27	2.39
Steve Biko						0	0	0	86	10	13	9	2	10	25	3	9	168	1.35
Charles Mokoena							0	0	101	2	13	7	1	6	12	17	7	166	1.33
Vuyani Taxi Rank								0	90	4	5	7	1	10	18	7	5	147	1.43
Civic Centre									0	0	7	2	0	0	0	3	4	16	0.52
Alighting	4	0	0	1	0	2	0	5	1060	33	134	74	14	64	144	81	107	1724	
Column Factor	3.01	0	0	1.15	0	0.98	0	1.88	0.83	1.38	2.75	1.58	0.47	2.27	2.98	2.33	1.59		

7.4 Validation of Results

The partial network level validation matrix of Route DO1 Kuyasa to Civic Center can be found in Appendix A (Figure A1), which was generated directly from the MyCiti data. The validation matrix can be utilized to compare the accuracy of the OD estimation procedure while implementing different statistical techniques (namely modified IPF and proportional distribution method). The comparison will solely focus on the transfer trip distribution, as this is a critical part of the analysis of a network level OD estimation.

The main difference between the methods used is (1) the difference in the basic assumptions for each method and (2) the effect of the seed matrix. As already highlighted in section 4.5.4 Proportional Distribution method assumes the passenger OD behavior is the same for transfer passengers and non-transfer passengers. This assumption is not required for the modified IPF method, which consequently makes it a more comprehensive statistical method. A seed matrix including transfer route-direction combinations as both virtual-origins and virtual-destinations provides a sample of actual passenger OD behavior. Thus, using the modified IPF method, it is possible to utilize this information to obtain an OD matrix estimation better correlated to the actual passenger behavior than assuming the transfer flows follow the same pattern as the total flows. However, the modified IPF method is heavily reliant on good quality boarding and alighting counts. With a large sample size, it is required to avoid the potential biases in the modified IPF method due to non-structural zeros, potentially resulting in a non-unique OD matrix estimation (Cui, 2006). The OD flows with destinations at the transfer point for Route DO1 Kuyasa – Civic Center generated using the two different statistical techniques are compared to the validation results as shown in figure 7-1.

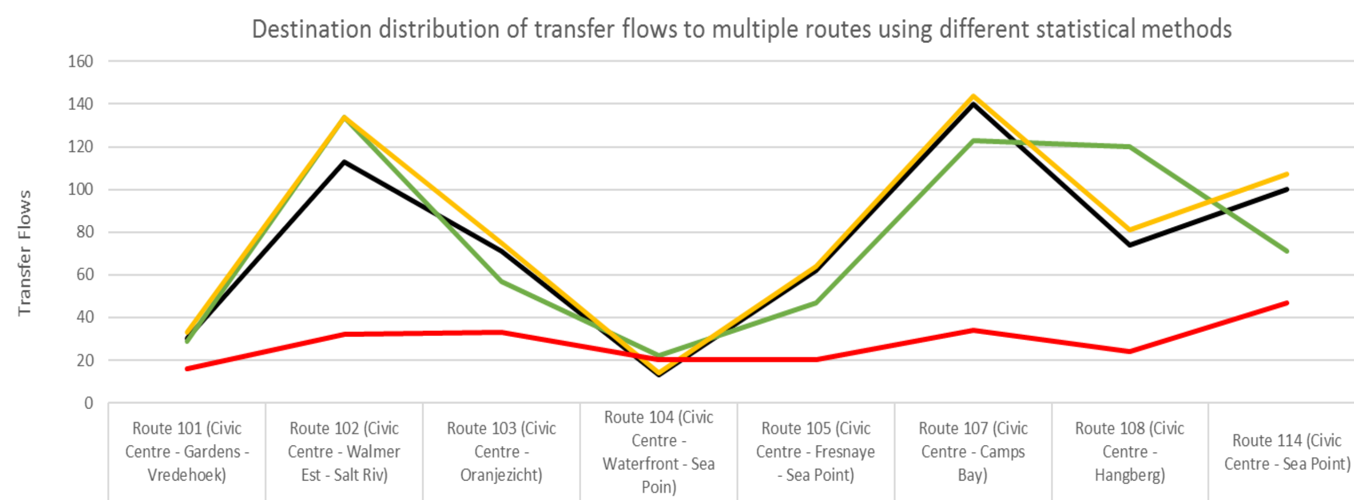


	Kuyasa Rail Station	Lindela	Dibana	Tutu	D Nyembe	Steve Biko	Charles Mokoena	Vuyani Taxi Rank	
Validation Matrix	174	48	65	117	7	66	60	51	
Proportional Distribution	190	64	55	115	10	59	58	52	
Modified IPF	197	51	61	118	4	82	65	58	
Seed (sample) matrix	67	15	24	37	1	29	21	19	13

Figure 7-1: comparison of origin distribution of transfer flows

From figure 7-1 above, there are similarities in the results of the two techniques, with each method following the trend of the validation results quite closely. However, between Lindela and Dibana, it was observed that proportional distribution had a negative gradient instead of the positive gradient followed by both the validation OD matrix results and Modified IPF method. Thus, this can be noted as a significant difference as relative fit is more critical than absolute fit, in situations where large samples cannot be obtained. Which, tends to happen frequently as collecting large sample data is a far too expensive exercise for some transport authorities. Additionally, relative fit provides information to transit authorities regards trends and patterns when certain events occur. In terms of absolute difference, the modified IPF method has the largest difference to the validation results of 23 transfer trips at the Kuyasa Rail Station. Further, figure 7-1 illustrates that the seed matrix follows a similar trend to the actual data before the seed matrix has been scaled up using the statistical techniques. This indicates that the seed matrix is a good representation of the population data and thus the Trip-chaining algorithm is a vital component in the estimation process. Additionally, the seed matrix can be effectively used by transit planners on a relative scale.

The OD flows with origins at the transfer point and destination at eight possible routes was also generated using the two statistical techniques and compared to the validation results as shown in figure 7-2.



	Route 101 (Civic Centre - Gardens - Vredehoek)	Route 102 (Civic Centre - Walmer Est - Salt Riv)	Route 103 (Civic Centre - Oranjezicht)	Route 104 (Civic Centre - Waterfront - Sea Poin)	Route 105 (Civic Centre - Fresnaye - Sea Point)	Route 107 (Civic Centre - Camps Bay)	Route 108 (Civic Centre - Hangberg)	Route 114 (Civic Centre - Sea Point)
Validation Matrix	30	113	71	13	62	140	74	100
Proportional Distribution	29	134	57	22	47	123	120	71
Modified IPF	33	134	75	14	64	144	81	107
Seed (sample) matrix	16	32	33	20	20	34	24	47

Figure 7-2: Comparison of destination distribution of transfer flows

From figure 7-2, the differences between the modified IPF method and Proportional distribution method is more apparent when considering the destination distribution of transfer flows. The proportional method has the highest absolute difference to the validation results with 46 transfer flows. In addition, between "virtual stops" route 108 and route 114, the gradient is negative for proportional fitting while the other two methods are showing a positive gradient. The modified IPF method follows the relative trend of the validation results, which is indicator of good control totals that ADC systems can provide transit authorities. The seed matrix in the case of destination distribution of transfer flows does not follow the trend of the actual OD counts as accurately as in figure 7-1. The probability of assigning a transfer destination accurately decreases as the number of transfer routes increases and some routes may have more regular passengers allowing their destinations to be inferred at a higher rate.

The RMSE for the network level OD matrix is 13.427 trips, which can be rounded up to 14 trips. Thus, residuals for the route are found to have an average spread of 14 trips from the validation data. This is higher than the RSME value found for the single route OD matrix, which is to be expected as network level OD matrices have a high number of destination stops and thus increased probability of OD flows make it difficult for the model to predict. However, an RSME value of 14 trips per OD flow is acceptable, considering the total number of passengers using the system.

With the implementation of both proportional distribution and the modified IPF method, the provides results which are similar to the validation results. This methodology can be further expanded to other parts of the network to form a complete matrix. In terms of the two statistical methods used to fully expand the partial network OD matrix, the modified IPF method provides a more sophisticated result. However, the modified IPF method requires a large sample size and high-quality control totals, while the proportional distribution is less stringent in requirements and computationally easier to apply. Thus, the use of these two methods largely depends on the type of dataset available and the accuracy of the result required.

7.5 Transfer flows Matrices

The final step in network level OD estimation is to generate the transfer OD Matrix, which is made from splitting the route level transfers to stop level transfers. However, the high number of possible transfer routes results in an even higher number of transfer stops, consequently resulting in a large number of non-structural zero flows. Therefore, the benefit of generating a transfer flow matrix at stop level is outweighed by the analytical and computational process of constructing it (Cui, 2006). For transit planning, route level assessment of transfer flows is adequate to provide transport authorities with the information they need to plan and build a general network matrix as was illustrated in figure 4-5.

Chapter 8

Conclusions and Recommendations

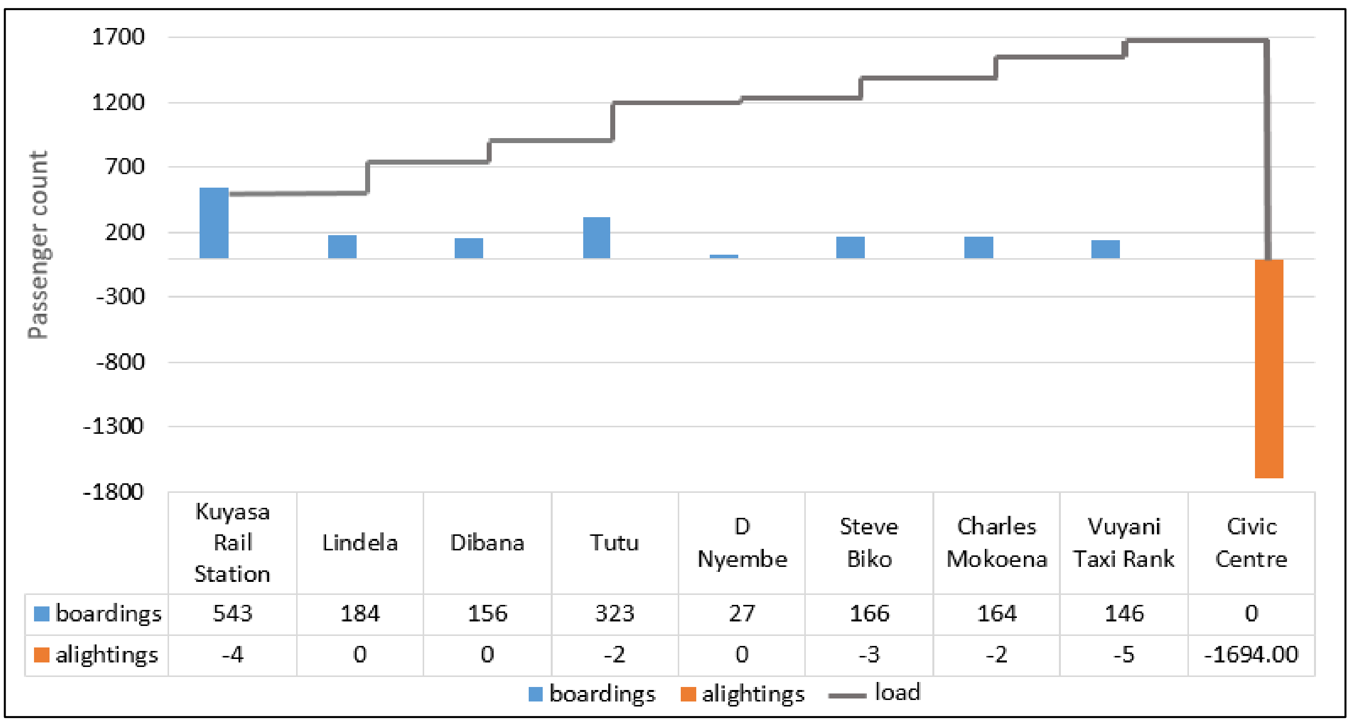
This Chapter begins with a discussion of the benefits of a network level OD matrix and the potential bias in the study in Section 8.1, followed by the potential application of the procedure to GABS in Section 8.2. Subsequently, recommendations are presented in Section 8.3 and finally topics of further research are proposed in Section 8.4.

8.1 Supplementary Discussion of Main Results

8.1.1 Applications of single route OD matrix estimation

Applications of estimating a single route OD matrix include and are not limited to the following:

1. Transit authorities can recognize the demand at particular stops or the lack of demand when deciding an upgrade of facilities within the network (Gordon, 2012).
2. The OD flows can be studied to find the trip purpose (Dumas, 2015).
3. The load capacity of a route can provide useful information for transit planners, such as accessing the effect of overcrowding on dwelling time, and serviceability reliability.



	Kuyasa Rail Station	Lindela	Dibana	Tutu	D Nyembe	Steve Biko	Charles Mokoena	Vuyani Taxi Rank	Civic Centre
boardings	543	184	156	323	27	166	164	146	0
alightings	-4	0	0	-2	0	-3	-2	-5	-1694.00

Figure 8-1: Load along a single bus route

8.1.2 Application of a network-level OD matrix estimation

The ability to estimate single route OD matrices using AFC data is a substantial improvement over conventional practice, however, the ability to estimate a network-level OD matrix provides significant additional value to the transit authorities. The benefits of estimating a network-level OD matrix include and are not limited to:

1. A better understanding of transfer flows provides information for better operations control strategies. For example, a decision to delay a bus (for whatever plausible reason) should consider the possibility that on-board passengers may need to make timed transfers downstream (Cui, 2006).
2. The transit authorities can identify critical transfer points and critical transfer directions where improvements of passenger transfer experiences, such as coordinated transfers. For example, when bus 1 arrives at the transfer stop and transfer bus 2 leaves a few minutes later, to ensure transferring passengers aren't left behind or waiting too long for bus 2.
3. Identifying transfer stops to other routes can help improve stop locations to minimize transfer walking distance.
4. The network level OD matrix provides a rational basis for network structure redesign or improvement

8.1.3 Discussion of potential Bias

The estimation procedure presented in this thesis does not always preserve randomness of samples, thus causing potential for bias. Traditionally this would require an extensive OD survey, which would cost transit authorities a large amount of money and time to validate the results (Gordon 2012). That would not be practical, considering that the aim of this thesis was to provide transit authorities alternative means to using the expensive and time-consuming surveys.

Thus, to deal with the bias and test the accuracy of the proposed procedure, this thesis provided a validation design which generated validation OD matrices using the sophisticated MyCiti ADC system. The estimated OD matrices were analyzed and compared to the validation OD matrices. The procedure was found to have an acceptable level of bias, as it was noted that regular passengers of the MyCiti system where more likely to have their destinations inferred compared to one-time users.

In conclusion, the bus OD matrix estimation process proposed in this thesis is unlikely to produce perfect results, however, it is superior to the traditional survey methods for bus OD matrix estimation. In addition, due to the relatively large sample sizes used, the proposed methodology is less likely to have bias in the resulting OD matrix.

8.2 Application of Procedure to GABS

The use of the procedure can be expanded to other PT systems which are flat fare based and have a AFC system as was tested during the period of this thesis. GABS is one such system, in 2015 GABS identified the handling of money as a payment method on buses as a high-risk factor in the company's business model. In order to mitigate the operational risk, GABS took a decision to invest in an AFC system at a cost of approximately R85 million. After a stringent procurement process, the tender for the company-wide roll-out of an AFC system was awarded to internationally acclaimed urban mobility company Parkeon (Interchange, 2016).

The new AFC system will consist of 1 250 Wayfarer200 driver console, which is an open platform industrial mobile computer configured with an Electronic Ticket Machine as illustrated in figure 1-1. In addition, the system will also feature dispatcher, point-of-sale units plus wireless LAN communications will remove the need for Driver Data Module, and data exchange between bus and the office system will become an automated process, removing the potential loss of data through Data Module failure (Interchange, 2016).

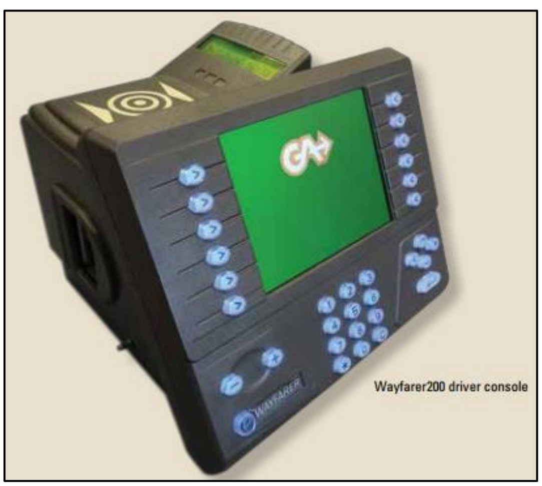


Figure 8-2: AFC system (wayfare200 driver console)

According to interchange (2016), the system will incorporate international best practices, as route based travel packages will be loaded onto smart cards and rides will be deducted when tapping on (when boarding the bus). Thus, GABS system will be different to MyCiti, as Golden Arrow's routes form an open system in which passengers are not expected to tap out (interchange, 2016); consequently, leaving no record of the alighting location. Without the alighting information, it is not a trivial task to generate an OD matrix, which is vital in transit planning, hence the procedure in this thesis has the potential to be calibrated for application to GABS AFC data.

8.3 Recommendations

8.3.1 Use of IPF method

The use of the IPF method is recommended when there is a way to accurately obtain the control totals (total boarding and alighting), thus the transit system being analyzed would be required to have an APC data system. In situations where there is no APC data, IPF method can still be used if the objective is to obtain the relative weightings of different OD flows. The high sampling rate of the seed matrix sufficiently captures the pattern of OD behavior. The relative weighting can provide transit authorities with the adequate information required for planning.

8.3.2 Use of modified IPF method or proportional distribution for transfer flows

Proportional distribution is computationally simpler and thus can be used as a quick analysis by transit authorities. Modified IPF method is more complicated than proportional distribution, however, provides a more accurate result, as the validation analysis conducted (in section 7.5) illustrated. In situations where there is low transfer OD flows, modified IPF at stop level is not recommended unless the non-structural zeros can be largely eliminated by utilizing more days or by analyzing at transfers at route level. In such situations, modified IPF method will be superiors the actual passenger transfer are taken in consideration in the formation of a transfer seed matrix.

8.4 Potential for Future Research

8.4.1 Use of MLE technique

Due to the computational difficulty of implementing MLE and limited time for the thesis, the MLE technique was not utilized. With the use of this method, instead of using sequential estimation approach, the network level OD matric could be estimated directly based on available data. Resulting in an OD matrix that could theoretically best explain all the observed data

8.4.2 Applying the methodology to other transit agencies

Transit agencies have different forms of available data, the OD estimation procedure described in this thesis needs to be adapted to utilize other agencies available data. Depending on the situation, there may be different challenges in applying the algorithm process described, and thus further studies should be conducted to apply the methodology to other transit agencies such GABS. The type of dataset produced by the GABS AFC system is required to conclusively say whether the procedure can be fully applied successfully to GABS. However, as a general method of using AFC data to estimate the OD matrix of a flat fare transport system the methodology produced in this thesis is adequate.

References

Alsger, A., Assemi, B., Mesbah, M. and Ferreira, L., 2016. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C: Emerging Technologies*, *68*, pp.490-506.

Barry, J., Newhouser, R., Rahbee, A. and Sayeda, S., 2002. Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board*, *(1817)*, pp.183-187.

Barry, J. J., Freimer, R., & Slavin, H. (2009). Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City. *Transportation Research Record: Journal of the Transportation Research Board*, *2112*(-1), 53–61. doi:10.3141/2112-07

Bagchi, M. and White, P., 2004. What role for smart-card data from bus systems. *Municipal Engineer*, *157*(1), pp.39-46.

Ben-Akiva, M.E., 1987. Methods to combine different data sources and estimate origin-destination matrices. *Transportation and traffic theory*.

Ben Akiva, M., P. P. Macke, and P. S. Hsu. Alternative Methods to Estimate Route-Level Trip Tables and Expand On-Board Surveys. In *Transportation Research Record 1037*, TRB, National Research Council, Washington, D. C., 1985, pp. 1-11.

CoCT. n.d. *TDI: transport analysis zones (TAZs)*. Available: http://www.tct.gov.za/docs/widgets/255/TDI3.png [2017, November 02].

Cui, A. *Bus* Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems. MS thesis. Massachusetts Institute of Technology, Cambridge, 2006.

Devillaine, F., Munizaga, M., & Trépanier, M. (2012). Detection of Activities of Public Transport Users by Analyzing Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*, *2276*(3), 48–55. doi:10.3141/2276-06

Dumas, R. A., 2012 Analysing Transit Equity Using Automatically Collected Data. MS thesis. Massachusetts Institute of Technology.

Farzin, J. M. Constructing an Automated Bus Origin-Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2072*, Transportation Research Board of the National Academies, Washington, D. C., 2008, pp. 30-37.

Gordon, J. B. Intermodal Passenger Flows on London's Public Transport Network: Automated Inference of Full Passenger Journeys Using Fare-Transaction and Vehicle-Location Data. MS thesis. Massachusetts Institute of Technology, Cambridge, 2012.

Interchange (2017). Quartly News Letter For Friends of Golden Arrow Bus Services. Available at: file:///C:/Users/QMBYAM001/Downloads/INTERCHANGE%20JULY%2016.pdf [2017, October 13]

Montero, L., Codina, E., & Barceló, J. (2015). Dynamic OD transit matrix estimation: formulation and model-building environment. In H. Selvaraj, D. Zydek, & G. Chmaj (Eds.), *Progress in Systems Engineering SE - 51*, Advances in Intelligent Systems and Computing (Vol. 1089, pp. 347–353). Springer International Publishing. doi:10.1007/978-3-319-08422-0_51

Munizaga, M. a., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, *24*, 9–18. doi:10.1016/j.trc.2012.01.007

Myciti. (2017). Complete MyCiTi Business Plan. Available at: https://myciti.org.za/docs/772/Complete_MyCiTi_Business_Plan_2012.pdf [2017, October 13]

Navick, D.S. and Furth, P.G., 1994. Distance-based model for estimating a bus route origin-destination matrix. *Transportation research record*, pp.16-16.

Navick, D. S., and P. G. Furth. Estimating Passenger Miles, Origin-Destination Patterns, and Loads with Location-Stamped Farebox Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1799*, Transportation Research Board of the National Academies, Washington, D. C., 2002, pp. 107-113.

Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557–568. doi:10.1016/j.trc.2010.12.003

Simon, J. and Furth, P.G., 1985. Generating a bus route OD matrix from on-off data. *Journal of Transportation Engineering*, *111*(6), pp.583-593.

Shelfer, K.M. and Procaccino, J.D., 2002. Smart card evolution. *Communications of the ACM*, *45*(7), pp.83-88.

Trépanier, M., Tranchant, N. and Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, *11*(1), pp.1-14.

Zhao, J., Rahbee, A. and Wilson, N.H., 2007. Estimating a Rail Passenger Trip Origin‐ Destination Matrix Using Automatic Data Collection Systems. *Computer‐ Aided Civil and Infrastructure Engineering*, *22*(5), pp.376-387.

Appendix A

(Additional OD Matrices Generated)

	Kuyasa Rail Station	Lindela	Dibana	Tutu	D Nyembe	Steve Biko	Charles Mokoena	Vuyani Taxi Rank	Civic Centre	Route 101 (Civic Centre - Gardens - Vredehoek)	Route 102 (Civic Centre - Walmer Est - Salt Riv)	Route 103 (103 Civic Centre - Oranjezicht)	Route 104 (Civic Centre - Waterfront - Sea Poin)	Route (105 Civic Centre - Fresnaye - Sea Point)	Route (107 Civic Centre - Camps Bay)	Route 108 (Civic Centre - Hangberg)	Route 114 (Civic Centre - Sea Point)	Boardings
Kuyasa Rail Station	4	0	0	0	0	1	1	3	360	4	42	16	1	20	36	20	35	534
Lindela		0	0	2	0	2	0	1	131	1	13	2	2	5	12	4	9	179
Dibana			0	0	0	0	1	0	90	5	10	13	1	7	15	7	7	155
Tutu				0	0	0	0	1	205	10	15	17	3	10	31	12	19	322
D Nyembe					0	0	0	0	20	0	2	1	0	2	2	0	0	27
Steve Biko						0	0	0	100	5	9	6	3	4	16	14	9	166
Charles Mokoena							0	0	104	1	13	10	0	10	8	7	11	164
Vuyani Taxi Rank								0	95	4	6	4	3	3	17	7	7	146
Civic Centre									0	0	3	2	0	1	3	3	3	15
Alighting	4	0	0	2	0	3	2	5	1105	30	113	71	13	62	140	74	100	1708

Figure A1: Validation OD matrix for Route DO1 Kuyasa- Civic Centre

ORIGIN/DETINATION	Tramway	Queens Beach	Brevity Lane	Kloof	Disandt	Fresnaye	Irwinton	The glen	Albany	Rhine	St Bedes	Ravenscraig	Ben Nevis	Skye Way	High Level	Old Fire Station	Strand	Adderley	Civic Centre	Total Boardings
Tramway		0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	4
Queens Beach			0	0	0	0	0	0	0	0	0	1	0	1	0	4	4	2	4	16
Brevity Lane				0	0	0	0	0	0	1	0	0	0	1	0	1	1	6	10	20
Kloof					0	0	0	0	0	1	0	0	0	0	0	0	3	12	12	28
Disandt						0	0	0	0	0	0	0	0	0	0	0	5	23	13	41
Fresnaye							0	0	0	4	0	0	0	0	0	1	6	18	15	44
Irwinton								0	0	0	1	0	0	2	0	1	11	7	12	34
The glen									0	0	0	0	0	0	0	3	5	6	3	17
Albany										3	0	0	0	0	0	2	6	9	12	32
Rhine											0	0	0	1	0	0	4	7	4	16
St Bedes												0	1	0	0	1	5	6	5	18
Ravenscraig													0	0	0	0	2	7	3	12
Ben Nevis														0	0	2	3	3	6	14
Skye Way															0	1	4	11	4	20
High Level																0	2	2	6	10
Old Fire Station																	0	0	5	5
Strand																		0	7	7
Adderley																			6	6
Civic Centre																			0	0
Total alighting	0	0	0	1	0	0	0	0	0	9	1	1	1	6	0	16	62	119	128	344

Figure A2: Single Route Estimation of Route 105 Tramway- Civic Centre

	Kapteinsklip	Snowdon	Paulsberg	Langeberg	Spine	Imperial	Mitchells Plain (Town Centre)	Pontiac	Kerrem	Marguerite	Sesame	Civic Centre	TOTAL BOARDINGS
Kapteinsklip	1	0	1	1	1	0	3	0	1	3	1	139	151
Snowdon		0	0	0	1	1	3	1	1	1	0	60	68
Paulsberg			0	0	0	0	10	1	1	0	1	87	100
Langeberg				0	0	0	13	1	2	1	1	88	106
Spine					0	0	7	1	0	0	0	66	74
Imperial						0	0	0	0	0	0	26	26
Mitchells Plain (Town Centre)							3	0	2	3	3	17	28
Pontiac								0	0	1	0	28	29
Kerrem									0	0	1	50	51
Marguerite										0	1	100	101
Sesame											0	88	88
Civic Centre												0	0
TOTAL ALIGHTING	1	0	1	1	2	1	39	4	7	9	8	749	822

Figure A3: Single Route OD estimation of Route D04 Kapteinsklip - Civic Centre

	Mitchells Plain (Town Centre)	Pontiac	Kerrem	Marguerite	Sesame	Civic Centre	Boardings
Mitchells Plain (Town Centre)	8	0	5	3	0	47	63
Pontiac		0	0	2	0	20	22
Kerrem			0	1	2	27	30
Marguerite				1	1	96	98
Sesame					0	67	67
Civic Centre						32	32
Alighting	8	0	5	7	3	289	312

Figure 25:Single Route OD matrix estimation of Route D03 Mithcells Plain to Civic Centre

	Civic Centre	Jeff Masemola	Aliam	Sigwele	Pama West	Mfundisweni	Makabeni	Ngcingcu	Kwezi Khayelitsha	Ncumo West	Oscar Mpheta	Kuyasa Rail Station	boardings
Civic Centre	181	28	31	26	13	18	56	20	41	30	27	16	487
Jeff Masemola		0	0	0	0	0	0	0	0	0	0	0	0
Aliam			0	0	0	0	0	0	1	0	0	0	1
Sigwele				0	0	0	0	0	0	0	0	0	0
Pama West					0	0	0	0	0	0	0	0	0
Mfundisweni						0	0	0	0	0	0	0	0
Makabeni							0	0	0	0	1	3	4
Ngcingcu								0	0	0	0	2	2
Kwezi Khayelitsha									0	0	1	1	2
Ncumo West										0	0	0	0
Oscar Mpheta											0	0	0
Kuyasa Rail Station												0	0
Alighting	181	28	31	26	13	18	56	20	42	30	29	22	496

Figure A4: Single Route OD matrox estimation for Route D02 Civic Center to Kuyasa

It can be observed that the OD flow pair Civic Centre to Civic Centre has quite a high count compared to the other pairs, this indicates transfer flows from other routes.

Appendix B

(Ethics Approval)

Application for Approval of Ethics in Research (EiR) Projects
Faculty of Engineering and the Built Environment, University of Cape Town

APPLICATION FORM

Please Note:
Any person planning to undertake research in the Faculty of Engineering and the Built Environment (EBE) at the University of Cape Town is required to complete this form **before** collecting or analysing data.The objective of submitting this application *prior* to embarking on research is to ensure that the highest ethical standards in research, conducted under the auspices of the EBE Faculty, are met. Please ensure that you have read, and understood the **EBE Ethics in Research Handbook**(available from the UCT EBE, Research Ethics website) prior to completing this application form: http://www.ebe.uct.ac.za/usr/ebe/research/ethics.pdf

APPLICANT'S DETAILS		
Name of principal researcher, student or external applicant		Ziphozihle Luvuno
Department		Civil Engineering
Preferred email address of applicant:		Lvnzip001@myuct.ac.za
If a Student	Your Degree: e.g., MSc, PhD, etc.,	BSc Civil Engineering
	Name of Supervisor (if supervised):	Mark Zuidgeest
If this is a researchcontract, indicate the source of funding/sponsorship		N/A
Project Title		Estimation and Validation of O-D Matrix Algorithm for PT

I hereby undertake to carry out my research in such a way that:

- there is no apparent legal objection to the nature or the method of research; and
- the research will not compromise staff or students or the other responsibilities of the University;
- the stated objective will be achieved, and the findings will have a high degree of validity;
- limitations and alternative interpretations will be considered;
- the findings could be subject to peer review and publicly available; and
- I will comply with the conventions of copyright and avoid any practice that would constitute plagiarism.

SIGNED BY	Full name	Signature	Date
Principal Researcher/ Student/External applicant	Ziphozihle Luvuno		**30 Aug 2017**

APPLICATION APPROVED BY	Full name	Signature	Date
Supervisor (where applicable)	Mark Zuidgeest		**30 Aug 2017**
HOD (or delegated nominee) Final authority for all applicants who have answered NO to all questions in Section1; and for all Undergraduateresearch (Including Honours).	M Vandersch[illegible]		30 Aug 2017
Chair : Faculty EIR Committee For applicants other than undergraduate students who have answered YES to any of the above questions.	Click here to enter text.		Click here to enter a date.

RE: [UCT Ethics in Research] Estimation and Validation of O-D Matrix Algorithm for PT

isabel.ncube=uct.ac.za@email.submittable.com
on behalf of
UCT Ethics in Research <isabel.ncube@uct.ac.za>

Reply all|

Thu 08-31, 12:55 PM

Ziphozihle Luvuno

To help protect your privacy, some content in this message has been blocked. To re-enable the blocked features, click here.

To always show content from this sender, click here.

Dear Student,

Thank you for your Ethics Application. Your application has been reviewed and approved by Ms Marianne Vanderschuren.

Kind regards, Isabel Ncube

You can go here to view the submission:
https://universityofcapetown.submittable.com/user/submissions/8420668

Appendix C

(Excel Procedures Completed)

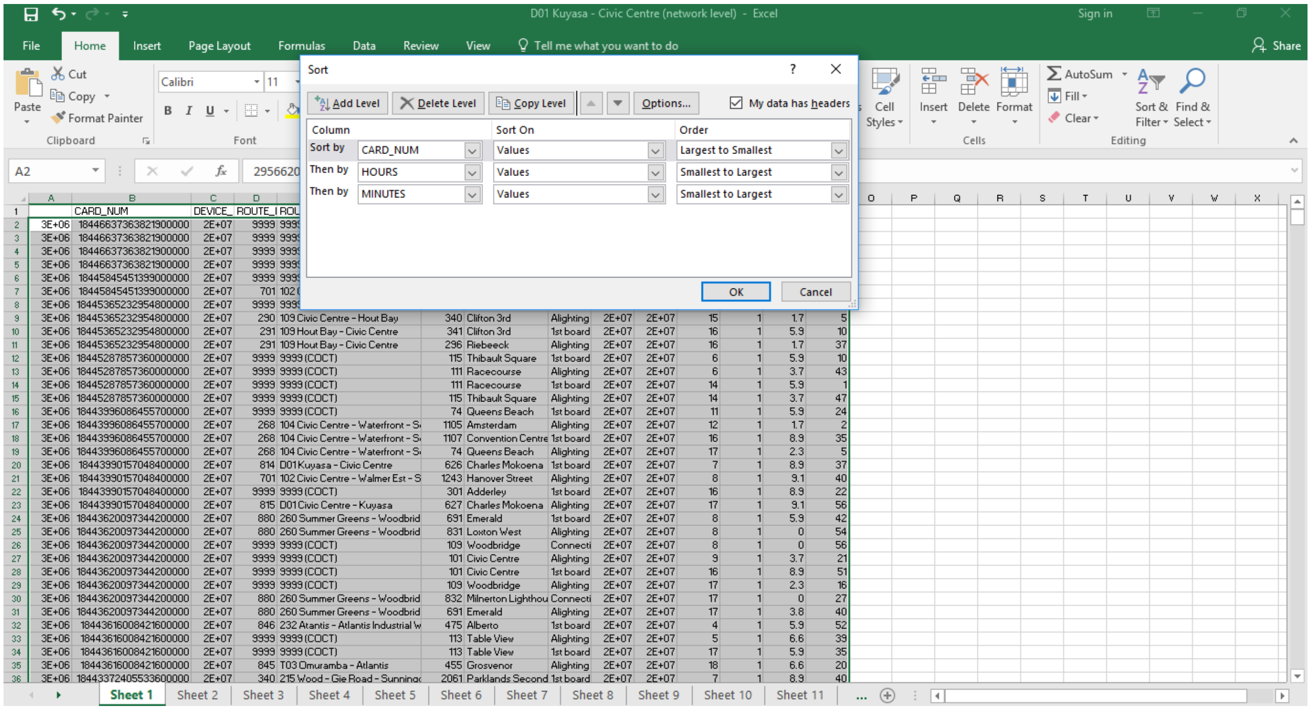


Figure C1

Load AFC data. Sort according to Card_Num ID, Hours and Minutes. Sorting provides an easier interpretation of the data, as it is possible to view a passenger morning and/or afternoon trips.Filter out all other routes to isolate D01 Kuyasa- Civic Centre and copy to sheet 2.

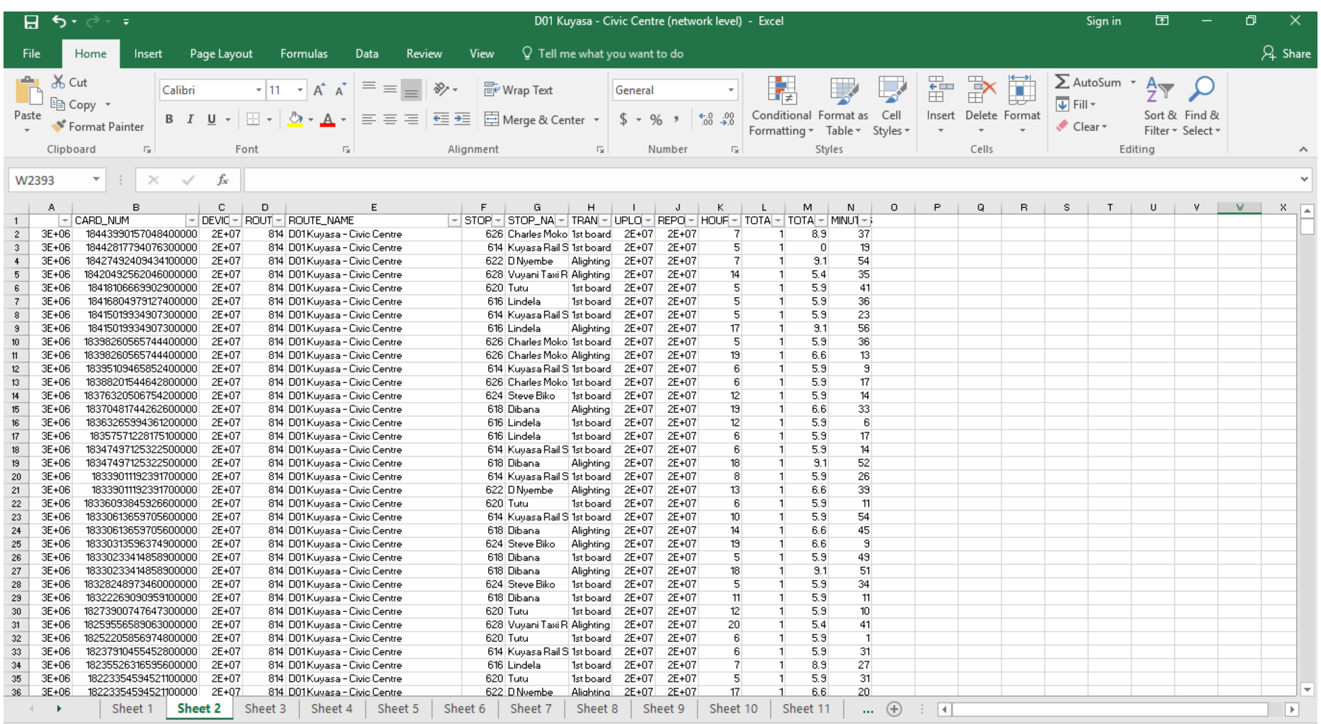


Figure C2

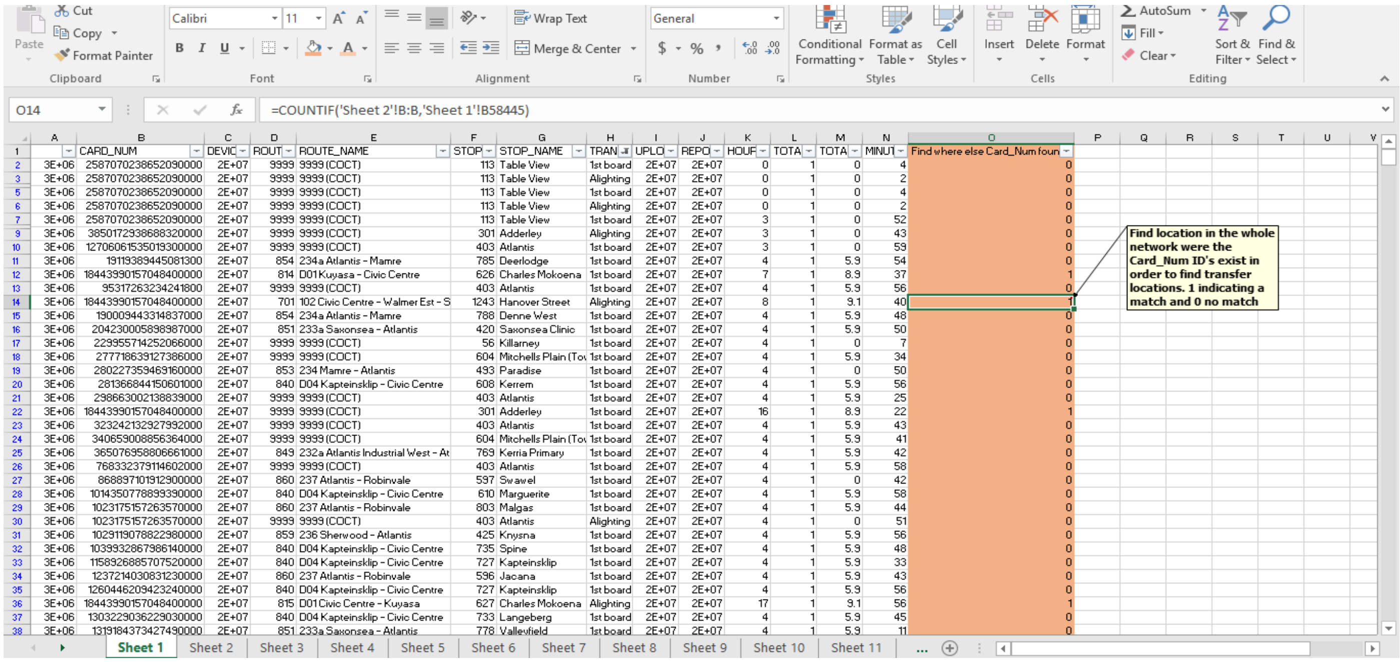


Figure C3

Use Countif formula to search locations elsewhere on the network were the Card_NUM ID's found on sheet 2 occur, 1 indicates a match and 0 indicates no match. Afterwhich the no matches are filtered out and matches are copied to sheet 3, the result shown in the figure 4 below. This step is crucial for transfer point inference, as the locations outside the route of interest are virtual stops.

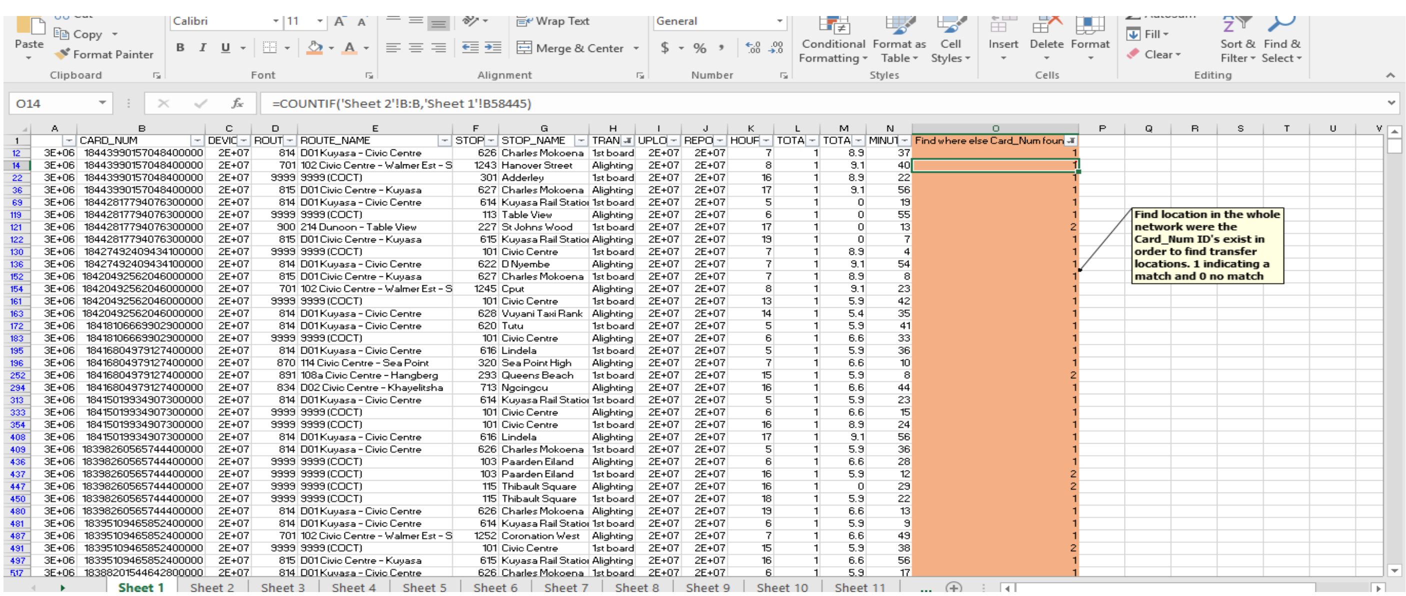


Figure C4

Countif statement is used to find the Card_Num which were not used along the route of interest.

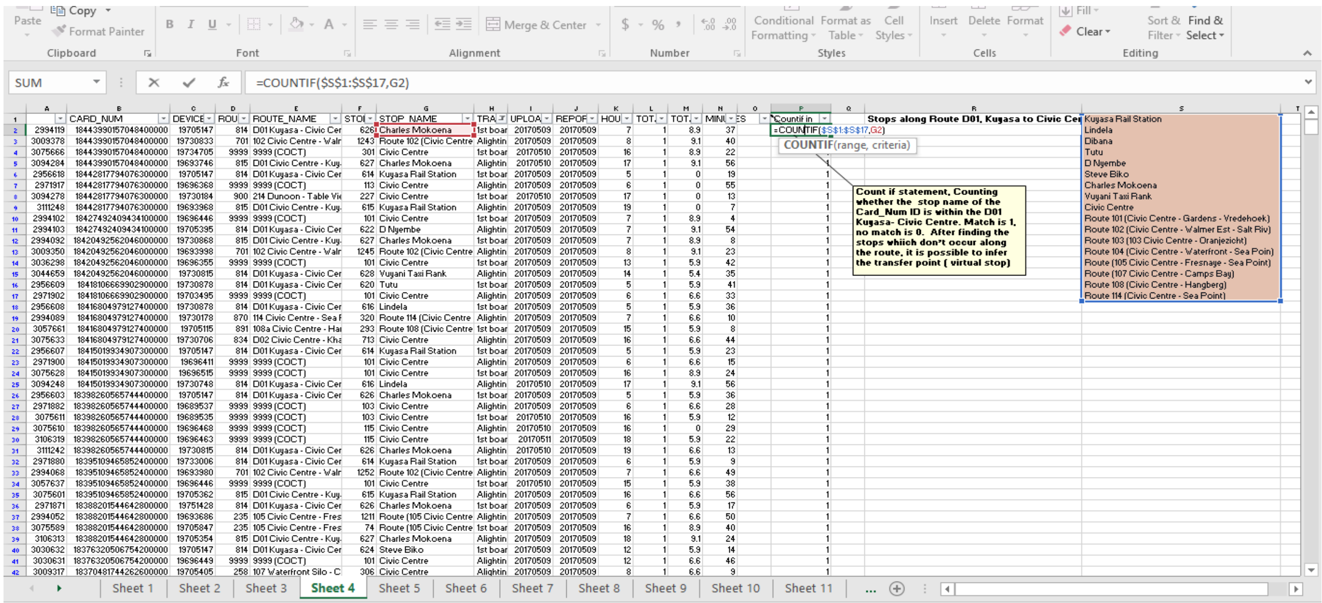


Figure C5

Refering to the figure 5 above , the function searches through the list of stops highlighted and when there is is a match it is indicated as 1 and no match 0. The 1's are filtered out in order to isolate the values which are outside the route of interest. Depending on which route the transfer occurs a virtual stop is inferred. Copy the result to sheet 5 as shown in the figure 6 below.

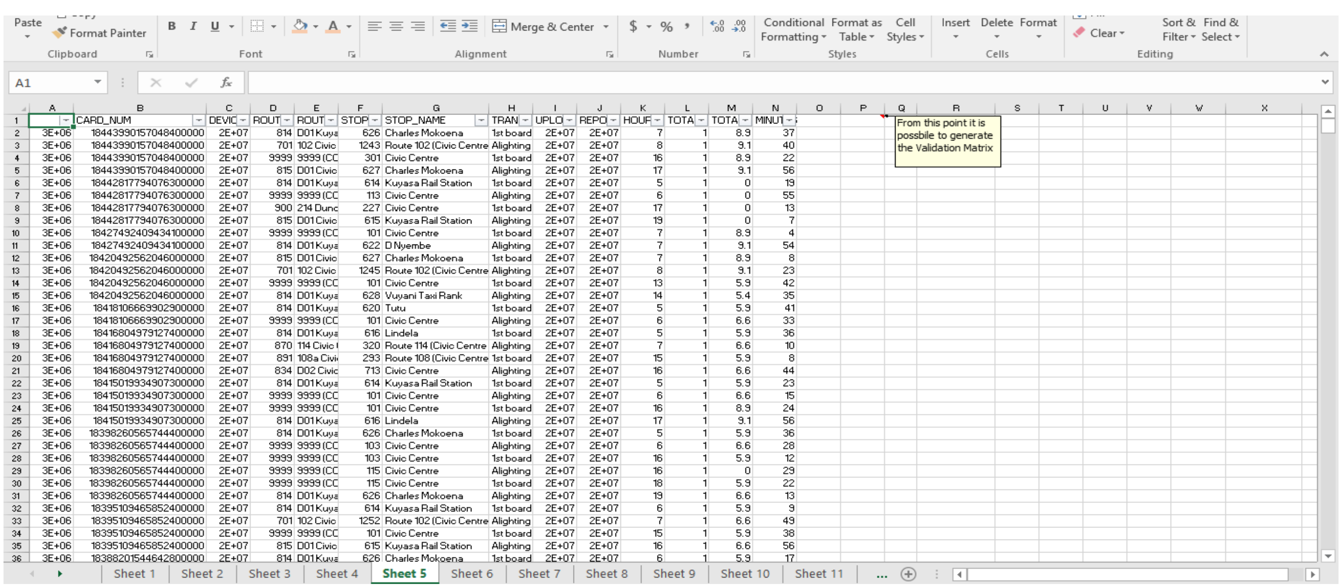


Figure C6

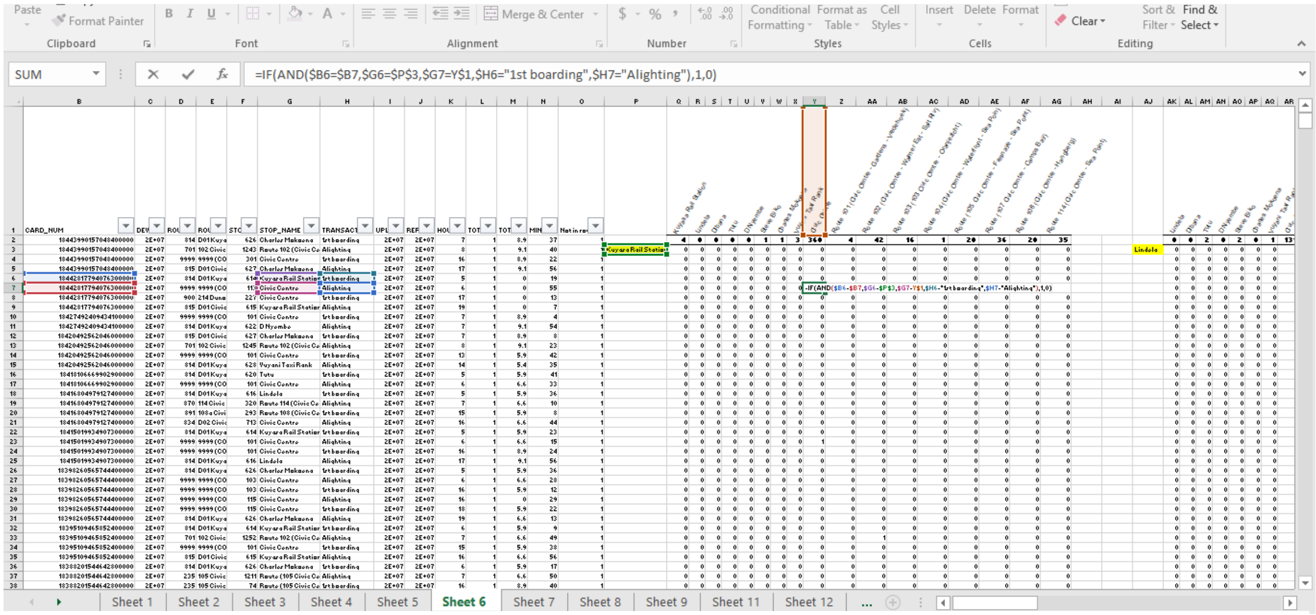


Figure C7

Copy the results from sheet 5 to sheet 6. Using logic tests (if statements) it is possible to count OD flow pair as illustrated in the figure 8 above. The method, counts when certain conditions are met , such as for the example when two consecutive Card_Num ID's are highlighted indicating one trip. Next condition involves the probability of moving from one station to the next station, finally the boarding and alighting taps must meet the direction requirements. When all conditions have been meet , the binary variable 1 is chosen and summed up to give the total of each OD flow pair. The process provides a quick count of possible OD flows, however the previous mentioned steps are required to prepare the data for theA 120 possible stop combinations were checked for 1709 trips.

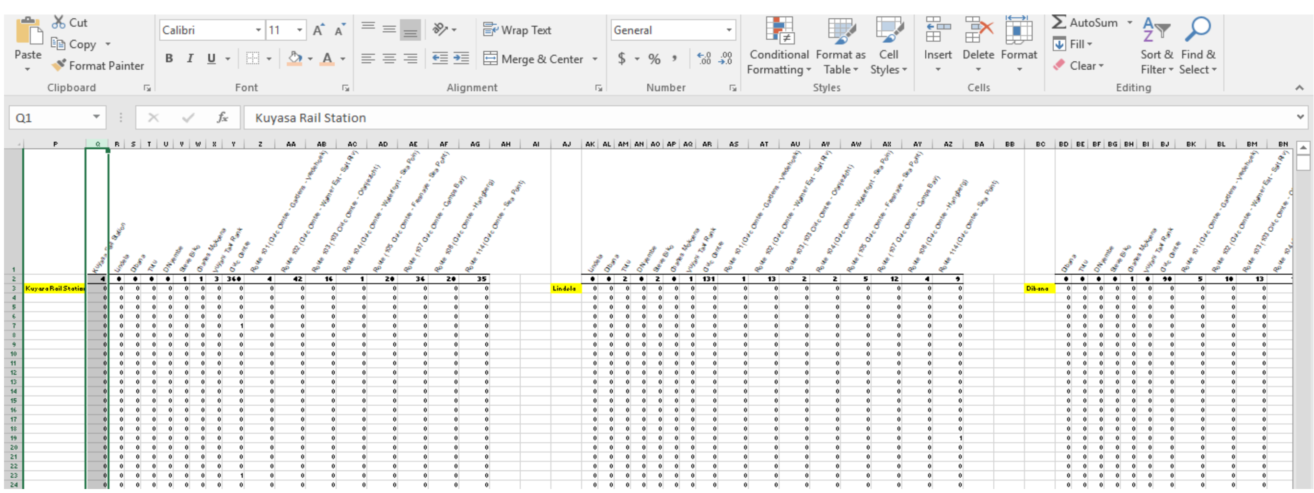
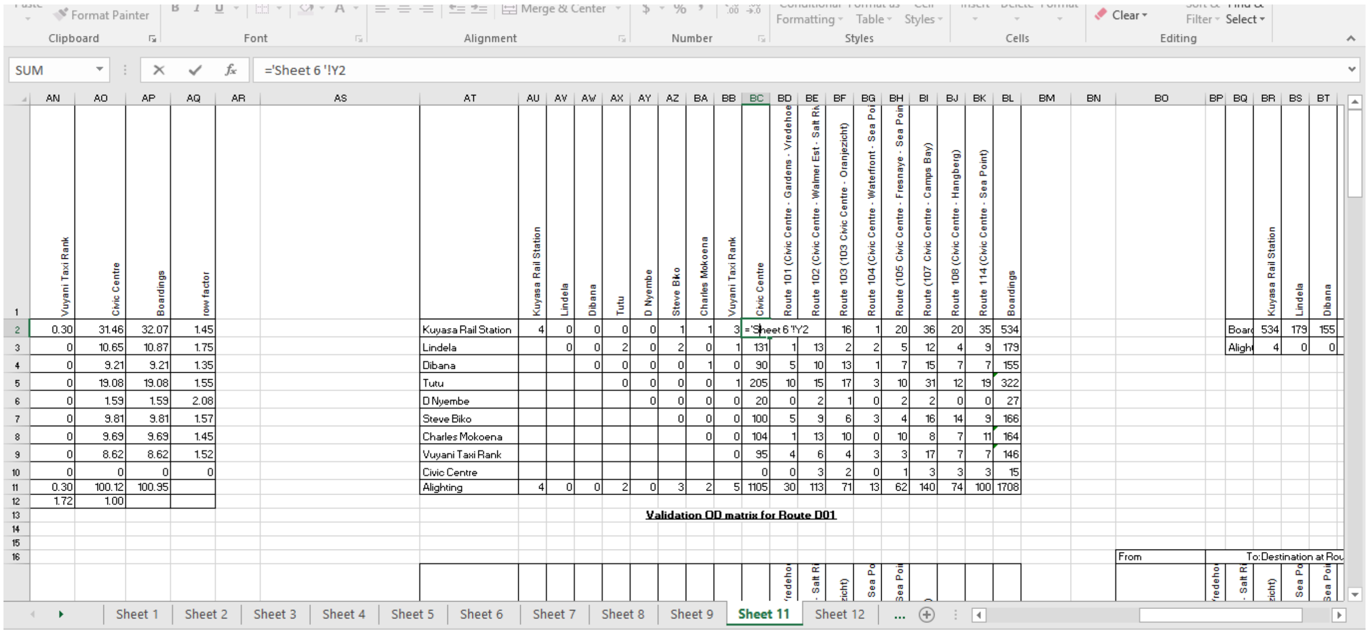


Figure C8

After adding up the total flows the Validation OD matrix can be constructed by linking the totals to an OD matrix table as shown in the following figure C9.



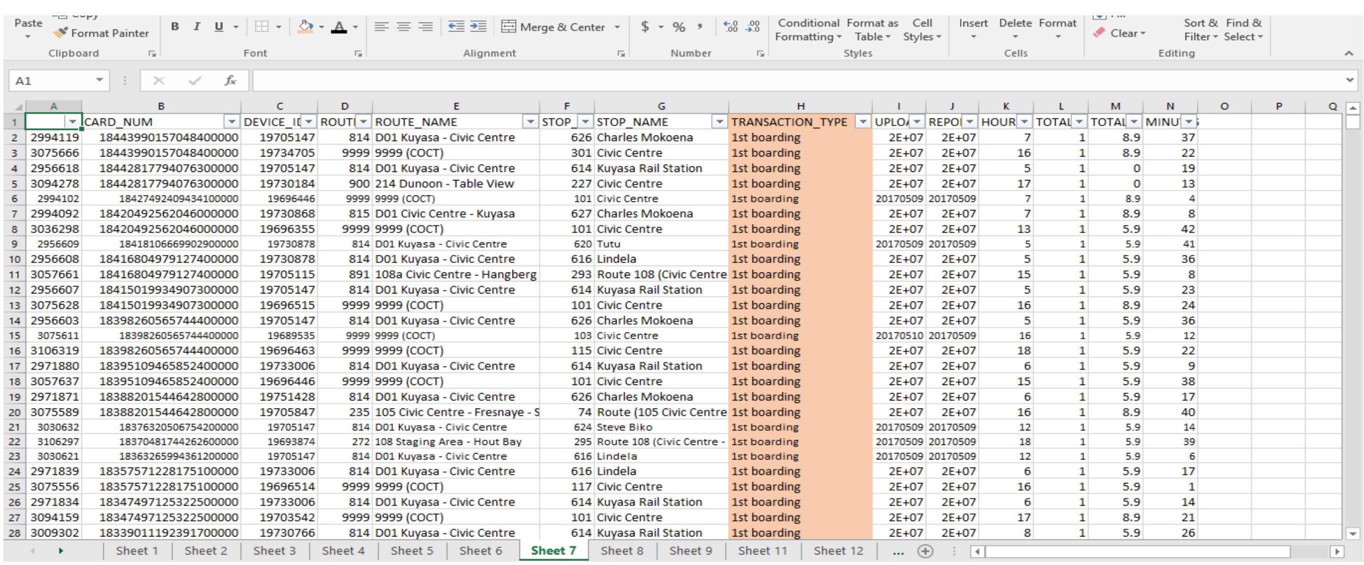
	Vuyani Taxi Rank	Civic Centre	Boardings	row factor
2	0.30	31.46	32.07	1.45
3	0	10.65	10.87	1.75
4	0	9.21	9.21	1.35
5	0	19.08	19.08	1.55
6	0	1.59	1.59	2.08
7	0	9.81	9.81	1.57
8	0	9.69	9.69	1.45
9	0	8.62	8.62	1.52
10	0	0	0	0
11	0.30	100.12	100.95	
12	1.72	1.00		

	Kuyasa Rail Station	Lindela	Dibana	Tutu	D Nyembe	Steve Biko	Charles Mokoena	Vuyani Taxi Rank	Civic Centre	Route 101 (Civic Centre - Gardens - Vredehoe	Route 102 (Civic Centre - Walmer Est - Salt Ri	Route 103 (103 Civic Centre - Oranjezicht)	Route 104 (Civic Centre - Waterfront - Sea Poi	Route (105 Civic Centre - Fresnaye - Sea Poi	Route (107 Civic Centre - Camps Bay)	Route 108 (Civic Centre - Hangberg)	Route 114 (Civic Centre - Sea Point)	Boardings
Kuyasa Rail Station	4	0	0	0	0	1	1	3	='Sheet 6 '!Y2		16	1	20	36	20	35	534	
Lindela		0	0	2	0	2	0	1	131	1	13	2	2	5	12	4	9	179
Dibana			0	0	0	0	1	0	90	5	10	13	1	7	15	7	7	155
Tutu				0	0	0	0	1	205	10	15	17	3	10	31	12	19	322
D Nyembe					0	0	0	0	20	0	2	1	0	2	2	0	0	27
Steve Biko						0	0	0	100	5	9	6	3	4	16	14	9	166
Charles Mokoena							0	0	104	1	13	10	0	10	8	7	11	164
Vuyani Taxi Rank								0	95	4	6	4	3	3	17	7	7	146
Civic Centre									0	0	3	2	0	1	3	3	3	15
Alighting	4	0	0	2	0	3	2	5	1105	30	113	71	13	62	140	74	100	1708

Validation OD matrix for Route D01

Figure C9

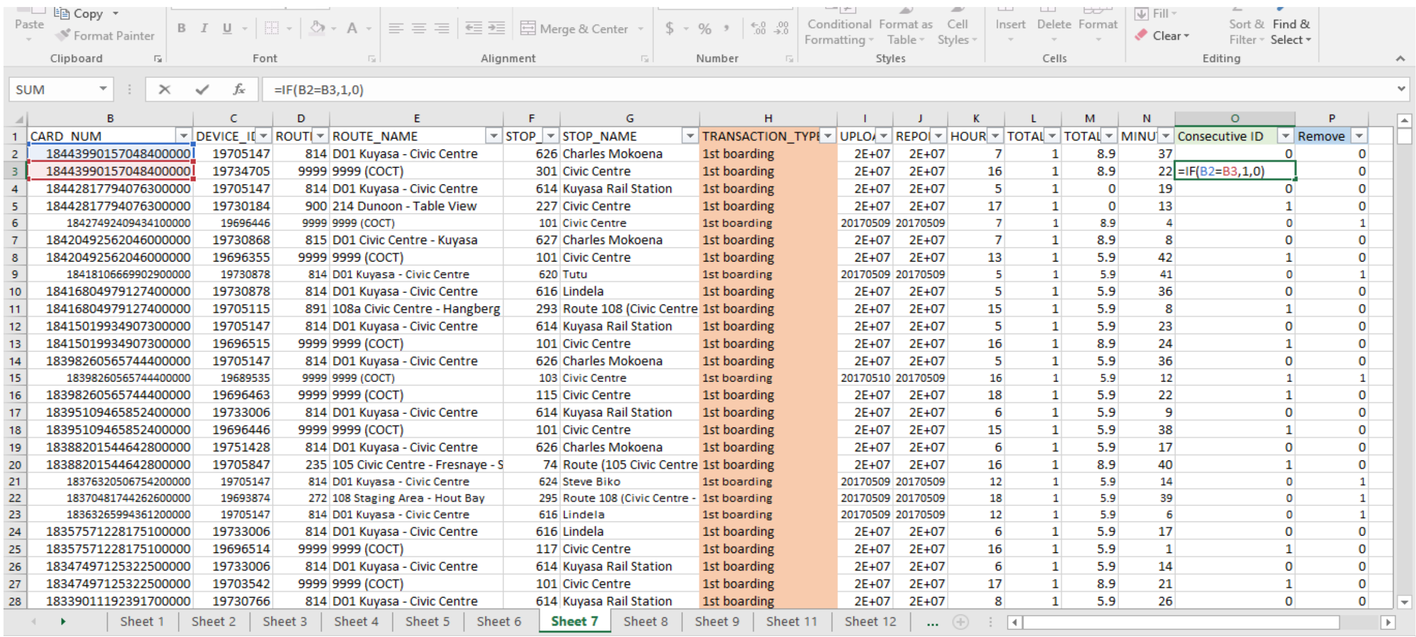
After generating the Validation OD matrix, and obtaining the alighting and boarding counts, the next step is to generate a seed matrix. The seed matrix is generated using the Trip-chaining algorithm. Thus return to sheet 5, and filter out all the alighting transactions leaaving on the boarding transactions (to simulate a Flat Fare system). The result is illustrated in figure 10 below.



	CARD_NUM	DEVICE_ID	ROUTE	ROUTE_NAME	STOP_	STOP_NAME	TRANSACTION_TYPE	UPLOA	REPOI	HOUR	TOTAL	TOTAL	MINU
2994119	18443990157048400000	19705147	814	D01 Kuyasa - Civic Centre	626	Charles Mokoena	1st boarding	2E+07	2E+07	7	1	8.9	37
3075666	18443990157048400000	19734705	9999	9999 (COCT)	301	Civic Centre	1st boarding	2E+07	2E+07	16	1	8.9	22
2956618	18442817794076300000	19705147	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	5	1	0	19
3094278	18442817794076300000	19730184	900	214 Dunoon - Table View	227	Civic Centre	1st boarding	2E+07	2E+07	17	1	0	13
2994102	18427492409434100000	19696446	9999	9999 (COCT)	101	Civic Centre	1st boarding	20170509	20170509	7	1	8.9	4
2994092	18420492562046000000	19730868	815	D01 Civic Centre - Kuyasa	627	Charles Mokoena	1st boarding	2E+07	2E+07	7	1	8.9	8
3036298	18420492562046000000	19696355	9999	9999 (COCT)	101	Civic Centre	1st boarding	2E+07	2E+07	13	1	5.9	42
2956609	18418106669902900000	19730878	814	D01 Kuyasa - Civic Centre	620	Tutu	1st boarding	20170509	20170509	5	1	5.9	41
2956608	18416804979127400000	19730878	814	D01 Kuyasa - Civic Centre	616	Lindela	1st boarding	2E+07	2E+07	5	1	5.9	36
3057661	18416804979127400000	19705115	891	108a Civic Centre - Hangberg	293	Route 108 (Civic Centre	1st boarding	2E+07	2E+07	15	1	5.9	8
2956607	18415019934907300000	19705147	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	5	1	5.9	23
3075628	18415019934907300000	19696515	9999	9999 (COCT)	101	Civic Centre	1st boarding	2E+07	2E+07	16	1	8.9	24
2956603	18398260565744400000	19705147	814	D01 Kuyasa - Civic Centre	626	Charles Mokoena	1st boarding	2E+07	2E+07	5	1	5.9	36
3075611	18398260565744400000	19689535	9999	9999 (COCT)	103	Civic Centre	1st boarding	20170510	20170509	16	1	5.9	12
3106319	18398260565744400000	19696463	9999	9999 (COCT)	115	Civic Centre	1st boarding	2E+07	2E+07	18	1	5.9	22
2971880	18395109465852400000	19733006	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	6	1	5.9	9
3057637	18395109465852400000	19696446	9999	9999 (COCT)	101	Civic Centre	1st boarding	2E+07	2E+07	15	1	5.9	38
2971871	18388201544642800000	19751428	814	D01 Kuyasa - Civic Centre	626	Charles Mokoena	1st boarding	2E+07	2E+07	6	1	5.9	17
3075589	18388201544642800000	19705847	235	105 Civic Centre - Fresnaye - S	74	Route (105 Civic Centre	1st boarding	2E+07	2E+07	16	1	8.9	40
3030632	18376320506754200000	19705147	814	D01 Kuyasa - Civic Centre	624	Steve Biko	1st boarding	20170509	20170509	12	1	5.9	14
3106297	18370481744262600000	19693874	272	108 Staging Area - Hout Bay	295	Route 108 (Civic Centre -	1st boarding	20170509	20170509	18	1	5.9	39
3030621	18363265994361200000	19705147	814	D01 Kuyasa - Civic Centre	616	Lindela	1st boarding	20170509	20170509	12	1	5.9	6
2971839	18357571228175100000	19733006	814	D01 Kuyasa - Civic Centre	616	Lindela	1st boarding	2E+07	2E+07	6	1	5.9	17
3075556	18357571228175100000	19696514	9999	9999 (COCT)	117	Civic Centre	1st boarding	2E+07	2E+07	16	1	5.9	1
2971834	18347497125322500000	19733006	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	6	1	5.9	14
3094159	18347497125322500000	19703542	9999	9999 (COCT)	101	Civic Centre	1st boarding	2E+07	2E+07	17	1	8.9	21
3009302	18339011192391700000	19730766	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	8	1	5.9	26

Figure C10

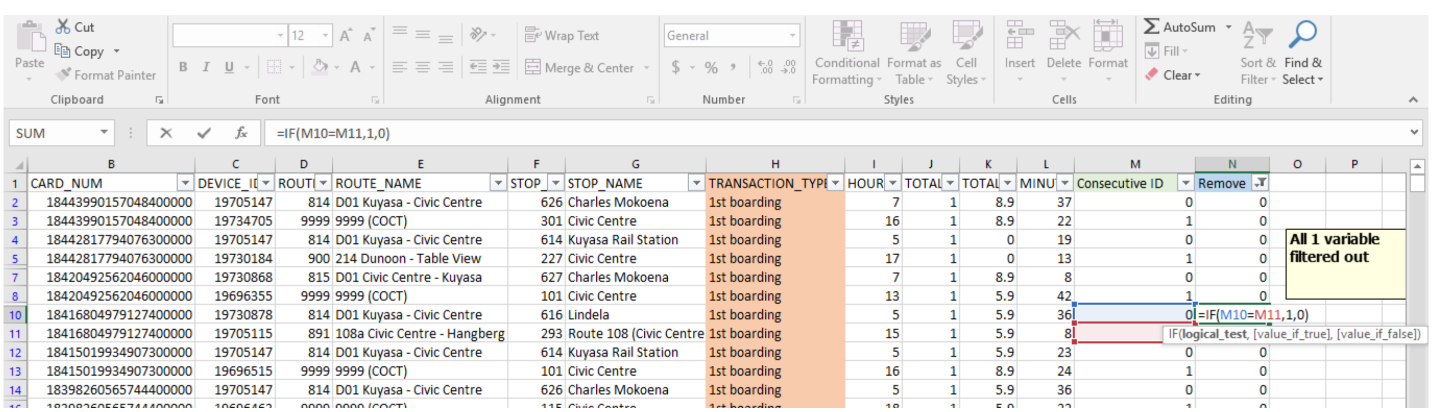
In order to infer destinations using the trip chain algorithm, the passenger must have at least two trips in a day, all the passengers that don't meet the requirement are omitted from the data.



	CARD_NUM	DEVICE_ID	ROUTI	ROUTE_NAME	STOP_	STOP_NAME	TRANSACTION_TYPE	UPLOA	REPOI	HOUR	TOTAL	TOTAL	MINUT	Consecutive ID	Remove
2	18443990157048400000	19705147	814	D01 Kuyasa - Civic Centre	626	Charles Mokoena	1st boarding	2E+07	2E+07	7	1	8.9	37	0	0
3	18443990157048400000	19734705	9999	9999 (COCT)	301	Civic Centre	1st boarding	2E+07	2E+07	16	1	8.9	22	=IF(B2=B3,1,0)	0
4	18442817794076300000	19705147	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	5	1	0	19	0	0
5	18442817794076300000	19730184	900	214 Dunoon - Table View	227	Civic Centre	1st boarding	2E+07	2E+07	17	1	0	13	1	0
6	18427492409434100000	19696446	9999	9999 (COCT)	101	Civic Centre	1st boarding	20170509	20170509	7	1	8.9	4	0	1
7	18420492562046000000	19730868	815	D01 Civic Centre - Kuyasa	627	Charles Mokoena	1st boarding	2E+07	2E+07	7	1	8.9	8	0	0
8	18420492562046000000	19696355	9999	9999 (COCT)	101	Civic Centre	1st boarding	2E+07	2E+07	13	1	5.9	42	1	0
9	18418106669902900000	19730878	814	D01 Kuyasa - Civic Centre	620	Tutu	1st boarding	20170509	20170509	5	1	5.9	41	0	1
10	18416804979127400000	19730878	814	D01 Kuyasa - Civic Centre	616	Lindela	1st boarding	2E+07	2E+07	5	1	5.9	36	0	0
11	18416804979127400000	19705115	891	108a Civic Centre - Hangberg	293	Route 108 (Civic Centre	1st boarding	2E+07	2E+07	15	1	5.9	8	1	0
12	18415019934907300000	19705147	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	5	1	5.9	23	0	0
13	18415019934907300000	19696515	9999	9999 (COCT)	101	Civic Centre	1st boarding	2E+07	2E+07	16	1	8.9	24	1	0
14	18398260565744400000	19705147	814	D01 Kuyasa - Civic Centre	626	Charles Mokoena	1st boarding	2E+07	2E+07	5	1	5.9	36	0	0
15	18398260565744400000	19689535	9999	9999 (COCT)	103	Civic Centre	1st boarding	20170510	20170509	16	1	5.9	12	1	1
16	18398260565744400000	19696463	9999	9999 (COCT)	115	Civic Centre	1st boarding	2E+07	2E+07	18	1	5.9	22	1	0
17	18395109465852400000	19733006	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	6	1	5.9	9	0	0
18	18395109465852400000	19696446	9999	9999 (COCT)	101	Civic Centre	1st boarding	2E+07	2E+07	15	1	5.9	38	1	0
19	18388201544642800000	19751428	814	D01 Kuyasa - Civic Centre	626	Charles Mokoena	1st boarding	2E+07	2E+07	6	1	5.9	17	0	0
20	18388201544642800000	19705847	235	105 Civic Centre - Fresnaye - S	74	Route (105 Civic Centre	1st boarding	2E+07	2E+07	16	1	8.9	40	1	0
21	18376320506754200000	19705147	814	D01 Kuyasa - Civic Centre	624	Steve Biko	1st boarding	20170509	20170509	12	1	5.9	14	0	1
22	18370481744262600000	19693874	272	108 Staging Area - Hout Bay	295	Route 108 (Civic Centre -	1st boarding	20170509	20170509	18	1	5.9	39	0	1
23	18363265994361200000	19705147	814	D01 Kuyasa - Civic Centre	616	Lindela	1st boarding	20170509	20170509	12	1	5.9	6	0	1
24	18357571228175100000	19733006	814	D01 Kuyasa - Civic Centre	616	Lindela	1st boarding	2E+07	2E+07	6	1	5.9	17	0	0
25	18357571228175100000	19696514	9999	9999 (COCT)	117	Civic Centre	1st boarding	2E+07	2E+07	16	1	5.9	1	1	0
26	18347497125322500000	19733006	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	6	1	5.9	14	0	0
27	18347497125322500000	19703542	9999	9999 (COCT)	101	Civic Centre	1st boarding	2E+07	2E+07	17	1	8.9	21	1	0
28	18339011192391700000	19730766	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	2E+07	2E+07	8	1	5.9	26	0	0

Figure C11

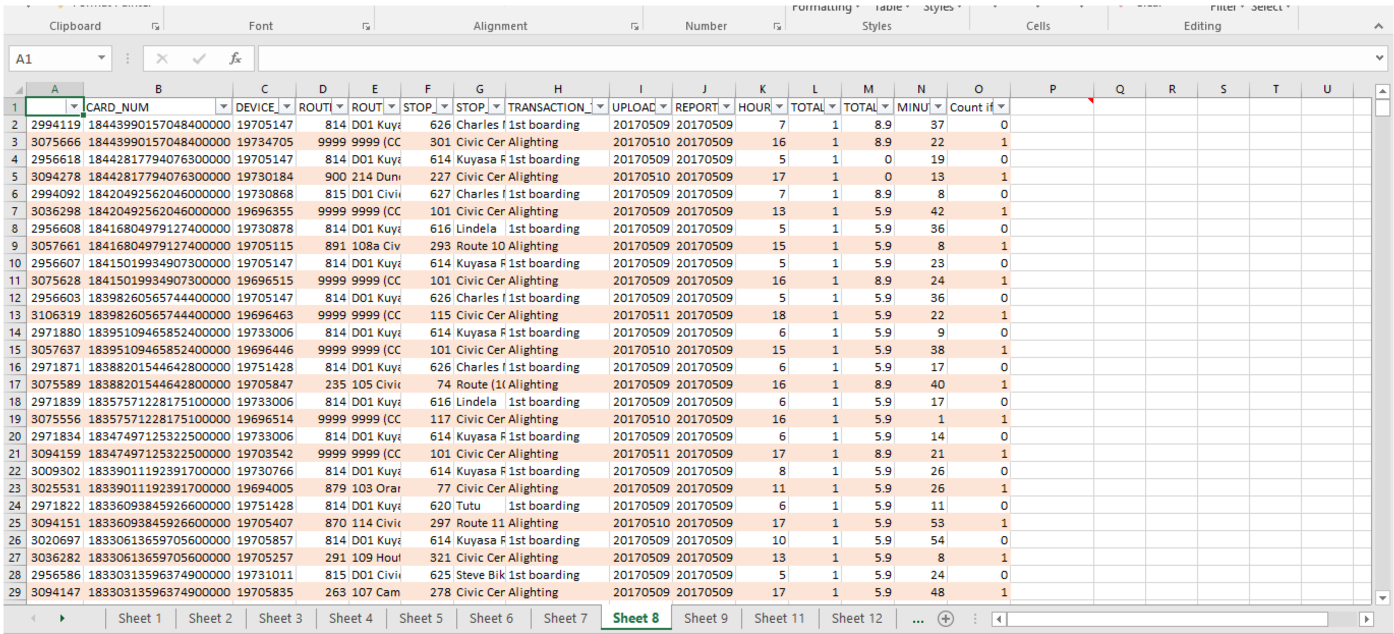
The checking procedure involves the use of binary variable, to find Card_Num ID's that appear consecutivly, the next step involves finding the Card_Num which don't break the 0,1,0,1 sequence. These are found under the remove column, with consecutive zeros indicating the sequence of 0,1,0,1 is being maintained and 1 indicating that the sequence has been broken e.g. 0,0,1,0 thus the Card entry of the second zero must be removed. This is simply done by filtering out all the 1 variables in the remove columns the result being passengers with two trips in a day. The result is shown in the figure C12 below.



	CARD_NUM	DEVICE_ID	ROUTI	ROUTE_NAME	STOP_	STOP_NAME	TRANSACTION_TYPE	HOUR	TOTAL	TOTAL	MINUT	Consecutive ID	Remove
2	18443990157048400000	19705147	814	D01 Kuyasa - Civic Centre	626	Charles Mokoena	1st boarding	7	1	8.9	37	0	0
3	18443990157048400000	19734705	9999	9999 (COCT)	301	Civic Centre	1st boarding	16	1	8.9	22	1	0
4	18442817794076300000	19705147	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	5	1	0	19	0	0
5	18442817794076300000	19730184	900	214 Dunoon - Table View	227	Civic Centre	1st boarding	17	1	0	13	1	0
7	18420492562046000000	19730868	815	D01 Civic Centre - Kuyasa	627	Charles Mokoena	1st boarding	7	1	8.9	8	0	0
8	18420492562046000000	19696355	9999	9999 (COCT)	101	Civic Centre	1st boarding	13	1	5.9	42	1	0
10	18416804979127400000	19730878	814	D01 Kuyasa - Civic Centre	616	Lindela	1st boarding	5	1	5.9	36	0	=IF(M10=M11,1,0)
11	18416804979127400000	19705115	891	108a Civic Centre - Hangberg	293	Route 108 (Civic Centre	1st boarding	15	1	5.9	8		
12	18415019934907300000	19705147	814	D01 Kuyasa - Civic Centre	614	Kuyasa Rail Station	1st boarding	5	1	5.9	23	0	0
13	18415019934907300000	19696515	9999	9999 (COCT)	101	Civic Centre	1st boarding	16	1	8.9	24	1	0
14	18398260565744400000	19705147	814	D01 Kuyasa - Civic Centre	626	Charles Mokoena	1st boarding	5	1	5.9	36	0	0

Figure C12

Copy the result to sheet 8. From this point it is possible to infer the alighting transaction, using time of tap. The first boarding activity in terms of time is left as boarding, the second boarding activity occurs later in the day, for the second trip. Therefore, according to Trip-chainging algorithm the boarding of the second trip is the alighting point of the first boarding activity.Thus the second boarding activity in terms of time is infered to be the alighting point and the result is shown in the figure C13 below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		CARD_NUM	DEVICE_	ROUTI	ROUT	STOP_	STOP_	TRANSACTION_	UPLOAD	REPORT	HOUR	TOTAL	TOTAL	MINU	Count if
2	2994119	18443990157048400000	19705147	814	D01 Kuya	626	Charles I	1st boarding	20170509	20170509	7	1	8.9	37	0
3	3075666	18443990157048400000	19734705	9999	9999 (CC	301	Civic Cer	Alighting	20170510	20170509	16	1	8.9	22	1
4	2956618	18442817794076300000	19705147	814	D01 Kuya	614	Kuyasa F	1st boarding	20170509	20170509	5	1	0	19	0
5	3094278	18442817794076300000	19730184	900	214 Dun	227	Civic Cer	Alighting	20170510	20170509	17	1	0	13	1
6	2994092	18420492562046000000	19730868	815	D01 Civi	627	Charles I	1st boarding	20170509	20170509	7	1	8.9	8	0
7	3036298	18420492562046000000	19696355	9999	9999 (CC	101	Civic Cer	Alighting	20170509	20170509	13	1	5.9	42	1
8	2956608	18416804979127400000	19730878	814	D01 Kuya	616	Lindela	1st boarding	20170509	20170509	5	1	5.9	36	0
9	3057661	18416804979127400000	19705115	891	108a Civ	293	Route 10	Alighting	20170509	20170509	15	1	5.9	8	1
10	2956607	18415019934907300000	19705147	814	D01 Kuya	614	Kuyasa F	1st boarding	20170509	20170509	5	1	5.9	23	0
11	3075628	18415019934907300000	19696515	9999	9999 (CC	101	Civic Cer	Alighting	20170509	20170509	16	1	8.9	24	1
12	2956603	18398260565744400000	19705147	814	D01 Kuya	626	Charles I	1st boarding	20170509	20170509	5	1	5.9	36	0
13	3106319	18398260565744400000	19696463	9999	9999 (CC	115	Civic Cer	Alighting	20170511	20170509	18	1	5.9	22	1
14	2971880	18395109465852400000	19733006	814	D01 Kuya	614	Kuyasa F	1st boarding	20170509	20170509	6	1	5.9	9	0
15	3057637	18395109465852400000	19696446	9999	9999 (CC	101	Civic Cer	Alighting	20170510	20170509	15	1	5.9	38	1
16	2971871	18388201544642800000	19751428	814	D01 Kuya	626	Charles I	1st boarding	20170509	20170509	6	1	5.9	17	0
17	3075589	18388201544642800000	19705847	235	105 Civic	74	Route (1(	Alighting	20170509	20170509	16	1	8.9	40	1
18	2971839	18357571228175100000	19733006	814	D01 Kuya	616	Lindela	1st boarding	20170509	20170509	6	1	5.9	17	0
19	3075556	18357571228175100000	19696514	9999	9999 (CC	117	Civic Cer	Alighting	20170510	20170509	16	1	5.9	1	1
20	2971834	18347497125322500000	19733006	814	D01 Kuya	614	Kuyasa F	1st boarding	20170509	20170509	6	1	5.9	14	0
21	3094159	18347497125322500000	19703542	9999	9999 (CC	101	Civic Cer	Alighting	20170511	20170509	17	1	8.9	21	1
22	3009302	18339011192391700000	19730766	814	D01 Kuya	614	Kuyasa F	1st boarding	20170509	20170509	8	1	5.9	26	0
23	3025531	18339011192391700000	19694005	879	103 Orar	77	Civic Cer	Alighting	20170509	20170509	11	1	5.9	26	1
24	2971822	18336093845926600000	19751428	814	D01 Kuya	620	Tutu	1st boarding	20170509	20170509	6	1	5.9	11	0
25	3094151	18336093845926600000	19705407	870	114 Civic	297	Route 11	Alighting	20170510	20170509	17	1	5.9	53	1
26	3020697	18330613659705600000	19705857	814	D01 Kuya	614	Kuyasa F	1st boarding	20170509	20170509	10	1	5.9	54	0
27	3036282	18330613659705600000	19705257	291	109 Hout	321	Civic Cer	Alighting	20170509	20170509	13	1	5.9	8	1
28	2956586	18330313596374900000	19731011	815	D01 Civi	625	Steve Bik	1st boarding	20170509	20170509	5	1	5.9	24	0
29	3094147	18330313596374900000	19705835	263	107 Cam	278	Civic Cer	Alighting	20170509	20170509	17	1	5.9	48	1

Figure C13

From figure C13 above it can be observed that the alightings infered are all at a later time during the day. Thus indicating that the alighitng inference is correct as this was originally the boarding location from the second trip, which is basis for trip reassignment for Trip-chaining method. If the second trip boarding activity which has now been infered to be the alighting point occurs outside the route-direction of interest, the transfer point is used as the stop location as was previously mentiioned. Thus it can be observed that all the stops present in the above picture all occur with D01 Kuyasa to Civic Centre. Hence, all the trip-chaining assumptions and tranfer flow assignment have been applied it is possible to count the OD flow pairs. Copy the information on sheet 8 to sheet 9 and proceed to count the OD flows using the if statement logic test setup.

Thus from the data in sheet 8 it is possible to start the OD flow pair counts in order to generate the seed matrix. The process is similar as for the validation matrix, in which all the possible stop combinations are considered, the results are illustrated in figure C8. It must be noted that the logic test is set in such a way to test a particular route-direction combination.

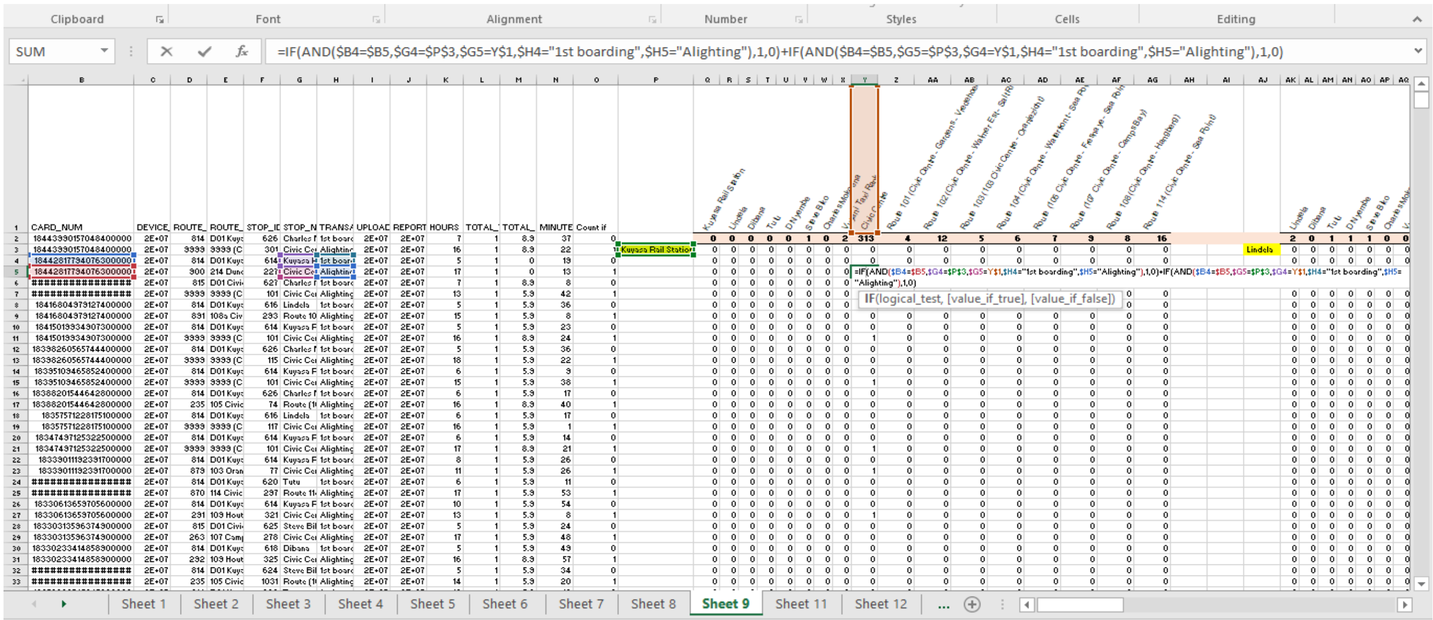
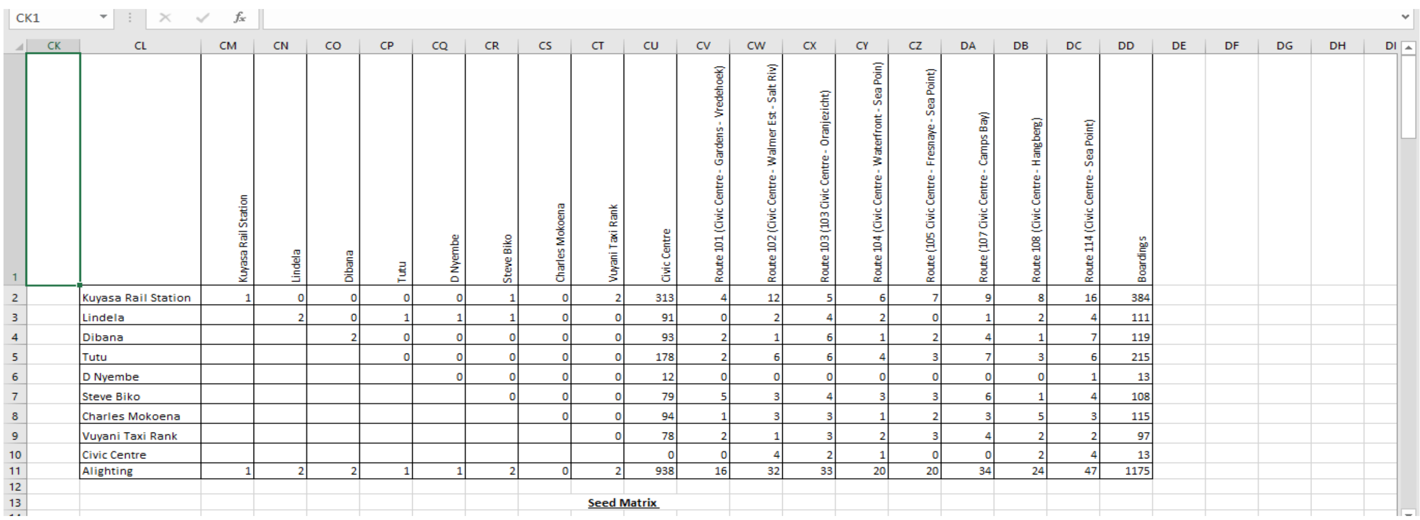


Figure C14

After running the process a total of 120 possible stop combinations were checked for 1175 trips. The total of the OD flow counts are then put into the matrix table as illustrated on figure C15.



	Kuyasa Rail Station	Lindela	Dibana	Tutu	D Nyembe	Steve Biko	Charles Mokoena	Vuyani Taxi Rank	Civic Centre	Route 101 (Civic Centre - Gardens - Vredehoek)	Route 102 (Civic Centre - Walmer Est - Salt Riv)	Route 103 (103 Civic Centre - Oranjezicht)	Route 104 (Civic Centre - Waterfront - Sea Poin)	Route (105 Civic Centre - Fresnaye - Sea Point)	Route (107 Civic Centre - Camps Bay)	Route 108 (Civic Centre - Hangberg)	Route 114 (Civic Centre - Sea Point)	Boardings
Kuyasa Rail Station	1	0	0	0	0	1	0	2	313	4	12	5	6	7	9	8	16	384
Lindela		2	0	1	1	1	0	0	91	0	2	4	2	0	1	2	4	111
Dibana			2	0	0	0	0	0	93	2	1	6	1	2	4	1	7	119
Tutu				0	0	0	0	0	178	2	6	6	4	3	7	3	6	215
D Nyembe					0	0	0	0	12	0	0	0	0	0	0	0	1	13
Steve Biko						0	0	0	79	5	3	4	3	3	6	1	4	108
Charles Mokoena							0	0	94	1	3	3	1	2	3	5	3	115
Vuyani Taxi Rank								0	78	2	1	3	2	3	4	2	2	97
Civic Centre									0	0	4	2	1	0	0	2	4	13
Alighting	1	2	2	1	1	2	0	2	938	16	32	33	20	20	34	24	47	1175

Seed Matrix

Figure C15

At this stage the expansion techniques namely, proportional distribution and modified IPF techniques can be applied. The application of both techniques was sufficiently documented and thus refer to chapter 6 and chapter 7 of this thesis for further reading on applications.

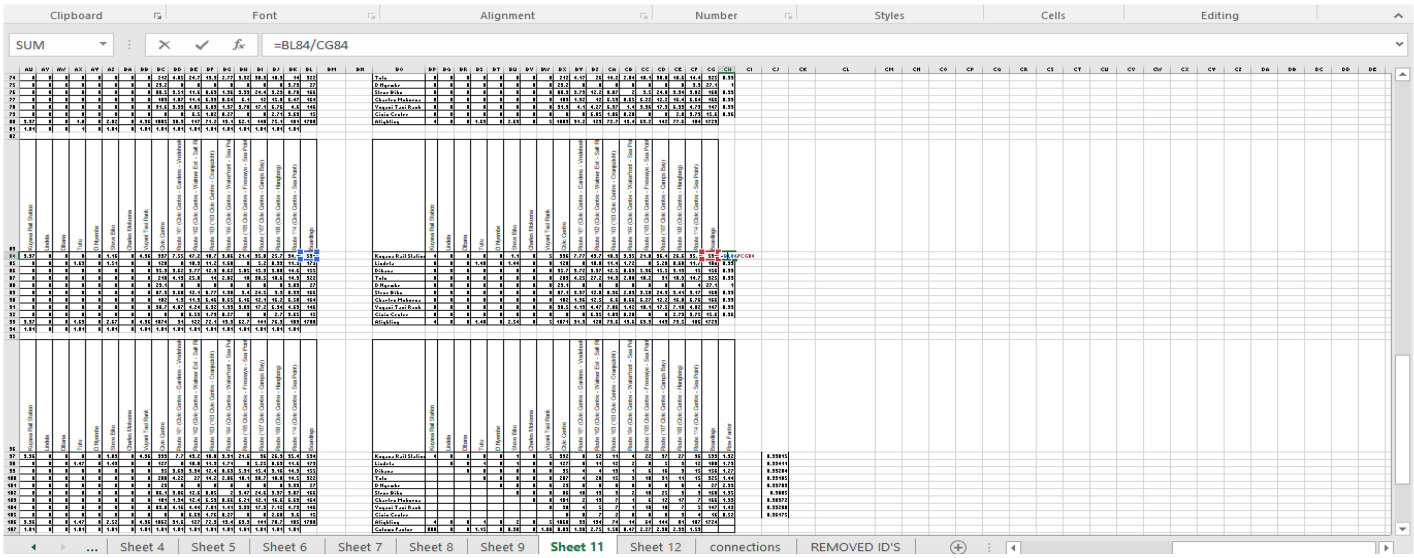
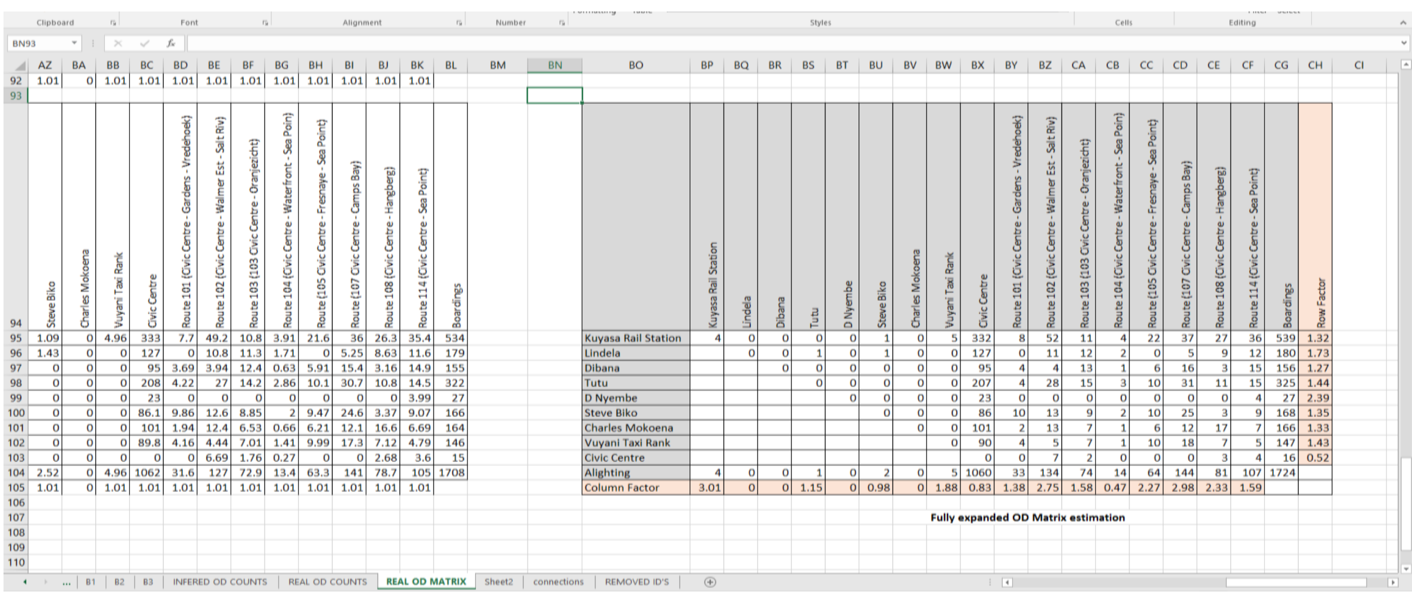


Figure C16

Figure C16 above illustrates the process of iterations of the column and row factors, the IPF expansion resulted in 13 iterations before convergence of the row and column factors, the result is illustrated in figure C17 below.



Steve Biko	Charles Mokoena	Vuyani Taxi Rank	Civic Centre	Route 101 {Civic Centre - Gardens - Vredehoek}	Route 102 {Civic Centre - Walmer Est - Salt Riv}	Route 103 {103 Civic Centre - Oranjezicht}	Route 104 {Civic Centre - Waterfront - Sea Poin}	Route {105 Civic Centre - Fresnaye - Sea Point}	Route {107 Civic Centre - Camps Bay}	Route 108 {Civic Centre - Hangberg}	Route 114 {Civic Centre - Sea Point}	Boardings
1.09	0	4.96	333	7.7	49.2	10.8	3.91	21.6	36	26.3	35.4	534
1.43	0	0	127	0	10.8	11.3	1.71	0	5.25	8.63	11.6	179
0	0	0	95	3.69	3.94	12.4	0.63	5.91	15.4	3.16	14.9	155
0	0	0	208	4.22	27	14.2	2.86	10.1	30.7	10.8	14.5	322
0	0	0	23	0	0	0	0	0	0	0	3.99	27
0	0	0	86.1	9.86	12.6	8.85	2	9.47	24.6	3.37	9.07	166
0	0	0	101	1.94	12.4	6.53	0.66	6.21	12.1	16.6	6.69	164
0	0	0	89.8	4.16	4.44	7.01	1.41	9.99	17.3	7.12	4.79	146
0	0	0	0	0	6.69	1.76	0.27	0	0	2.68	3.6	15
2.52	0	4.96	1062	31.6	127	72.9	13.4	63.3	141	78.7	105	1708
1.01	0	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	

	Kuyasa Rail Station	Lindela	Dibana	Tutu	D Nyembe	Steve Biko	Charles Mokoena	Vuyani Taxi Rank	Civic Centre	Route 101 {Civic Centre - Gardens - Vredehoek}	Route 102 {Civic Centre - Walmer Est - Salt Riv}	Route 103 {103 Civic Centre - Oranjezicht}	Route 104 {Civic Centre - Waterfront - Sea Poin}	Route {105 Civic Centre - Fresnaye - Sea Point}	Route {107 Civic Centre - Camps Bay}	Route 108 {Civic Centre - Hangberg}	Route 114 {Civic Centre - Sea Point}	Boardings	Row Factor
Kuyasa Rail Station	4	0	0	0	0	1	0	5	332	8	52	11	4	22	37	27	36	539	1.32
Lindela		0	0	1	0	1	0	0	127	0	11	12	2	0	5	9	12	180	1.73
Dibana			0	0	0	0	0	0	95	4	4	13	1	6	16	3	15	156	1.27
Tutu				0	0	0	0	0	207	4	28	15	3	10	31	11	15	325	1.44
D Nyembe					0	0	0	0	23	0	0	0	0	0	0	0	4	27	2.39
Steve Biko						0	0	0	86	10	13	9	2	10	25	3	9	168	1.35
Charles Mokoena							0	0	101	2	13	7	1	6	12	17	7	166	1.33
Vuyani Taxi Rank								0	90	4	5	7	1	10	18	7	5	147	1.43
Civic Centre									0	0	7	2	0	0	0	3	4	16	0.52
Alighting	4	0	0	1	0	2	0	5	1060	33	134	74	14	64	144	81	107	1724	
Column Factor	3.01	0	0	1.15	0	0.98	0	1.88	0.83	1.38	2.75	1.58	0.47	2.27	2.98	2.33	1.59		

Fully expanded OD Matrix estimation

Figure C17

Next proportional distribution technique was implemented as described in chapter 4 and 7 on sheet 11 of the spreedsheet.

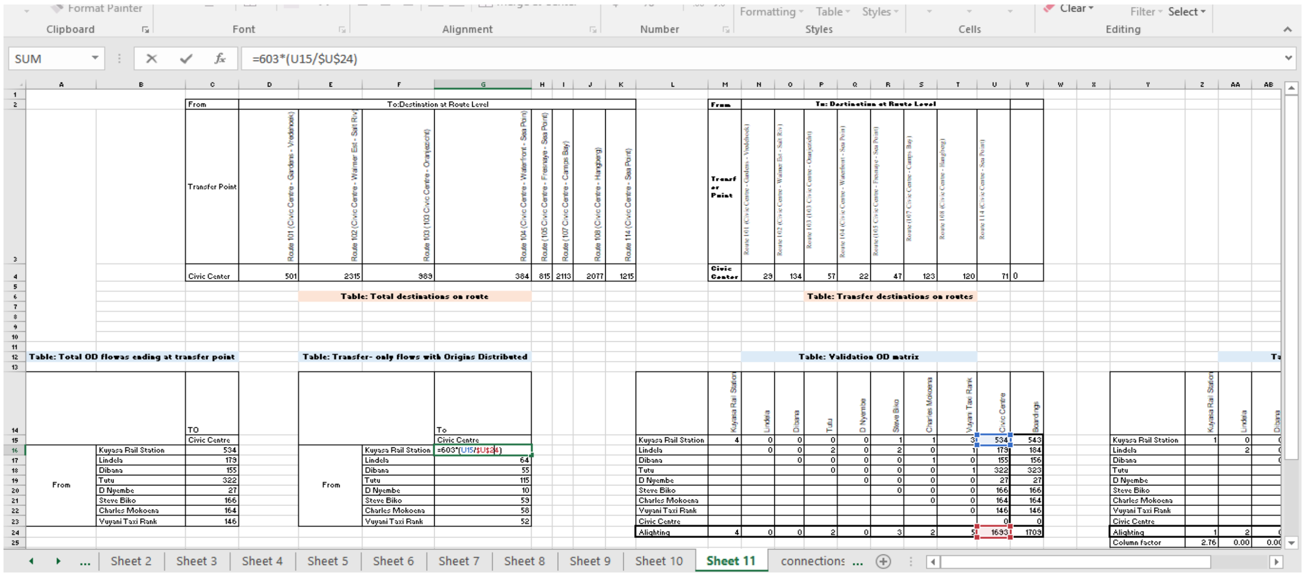


Figure C18

The transfer flows with origins distributed were found and transfer destinations were assigned to the routes. These follows were combined with the seed matrix to obtain a fully expanded network level OD matrices.

At this point it was possible to validate the transfers flows of the estimated matrices against the validation matrix. This was illustrated in the form of a graph as and the RMSE method was applied to each OD flow pair to obtain the total RMSE value.

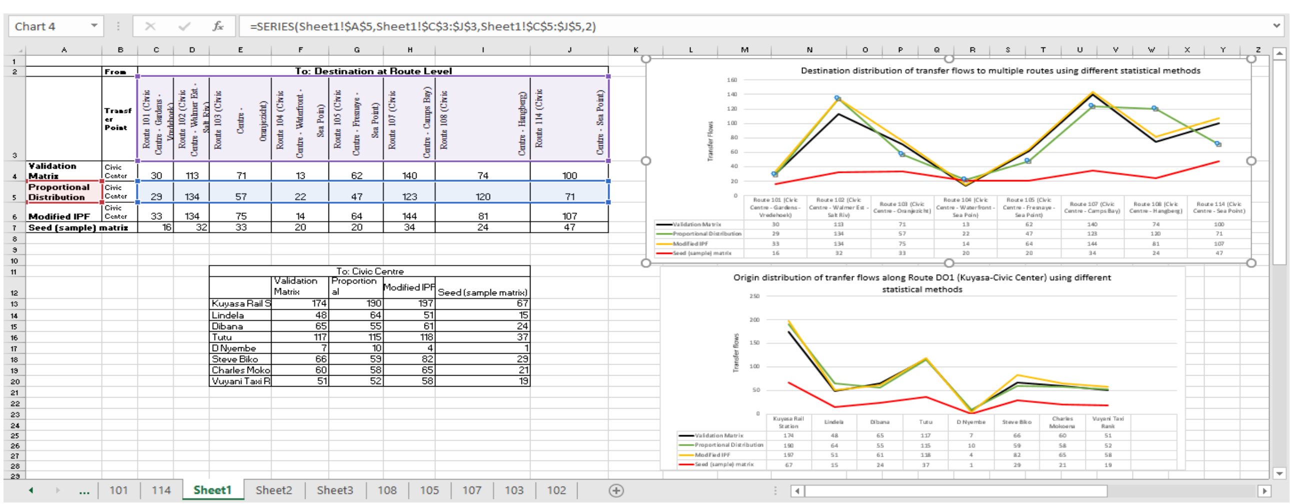


Figure C19

APPENDIX D

(Map of the Complete Myciti Network)

MyCiTi System Map

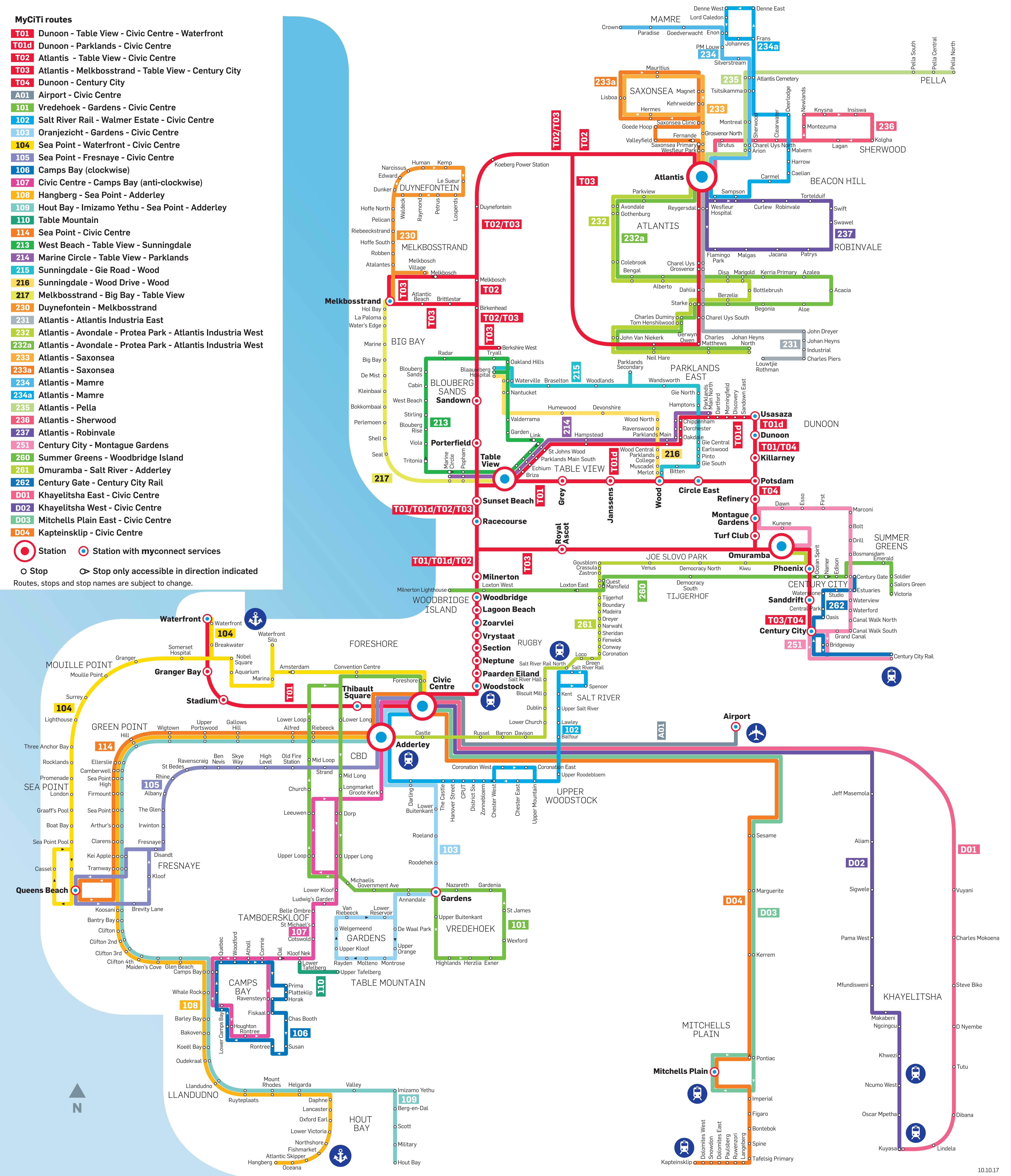
MyCiTi routes

- T01 Dunoon - Table View - Civic Centre - Waterfront
- T01d Dunoon - Parklands - Civic Centre
- T02 Atlantis - Table View - Civic Centre
- T03 Atlantis - Melkbosstrand - Table View - Century City
- T04 Dunoon - Century City
- A01 Airport - Civic Centre
- 101 Vredehoek - Gardens - Civic Centre
- 102 Salt River Rail - Walmer Estate - Civic Centre
- 103 Oranjezicht - Gardens - Civic Centre
- 104 Sea Point - Waterfront - Civic Centre
- 105 Sea Point - Fresnaye - Civic Centre
- 106 Camps Bay (clockwise)
- 107 Civic Centre - Camps Bay (anti-clockwise)
- 108 Hangberg - Sea Point - Adderley
- 109 Hout Bay - Imizamo Yethu - Sea Point - Adderley
- 110 Table Mountain
- 114 Sea Point - Civic Centre
- 213 West Beach - Table View - Sunningdale
- 214 Marine Circle - Table View - Parklands
- 215 Sunningdale - Gie Road - Wood
- 216 Sunningdale - Wood Drive - Wood
- 217 Melkbosstrand - Big Bay - Table View
- 230 Duynefontein - Melkbosstrand
- 231 Atlantis - Atlantis Industria East
- 232 Atlantis - Avondale - Protea Park - Atlantis Industria West
- 232a Atlantis - Avondale - Protea Park - Atlantis Industria West
- 233 Atlantis - Saxonsea
- 233a Atlantis - Saxonsea
- 234 Atlantis - Mamre
- 234a Atlantis - Mamre
- 235 Atlantis - Pella
- 236 Atlantis - Sherwood
- 237 Atlantis - Robinvale
- 251 Century City - Montague Gardens
- 260 Summer Greens - Woodbridge Island
- 261 Omuramba - Salt River - Adderley
- 262 Century Gate - Century City Rail
- D01 Khayelitsha East - Civic Centre
- D02 Khayelitsha West - Civic Centre
- D03 Mitchells Plain East - Civic Centre
- D04 Kapteinsklip - Civic Centre

Station — Station with myconnect services

Stop — Stop only accessible in direction indicated

Routes, stops and stop names are subject to change.



APPENDIX E

(ESCA outcomes)

ECSA Outcome 1: Problem Solving

Problem solving was required in order to analysis the AFC data, which consisted of over 164 000 commuter entries. Due to the size of the dataset certain analysis techniques had to be implemented. This included: determining how obtain OD flow counts for passengers originating from specific area and going to another area, this required a procedural approach in which a number of logical tests where's set up. In addition, the thesis problem itself posed a question of how to infer the alighting locations of passengers on a public transport system (flat fare based) that only has boarding validation.

ECSA outcome 2: Application of scientific and engineering knowledge

The research project required extensive engineering knowledge regarding transportation concepts such as trip generation, OD matrix estimation techniques, and the role of technological advances on transportation analysis. Furthermore, scientific knowledge was required in terms of producing a research framework that minimized bias and adhered to good experimental techniques.

ECSA outcome 4: Investigations, experiments and data analysis

A number of investigations were completed, in the form of a desktop study, interview with industry expert, literature review and evaluating a number of analysis platforms, such as Python, Stata and Excel. An investigation for creating a procedure for OD matrix estimation was completed in which the alighting locations were inferred. Majority of the work involved statistical analysis, certain travel scenarios had to be considered in setting up the inference procedure. After obtaining the OD estimations, validation analysis was completed, and it was concluded that the procedure provide acceptable results within an acceptable margin of bias.

ECSA 6 Professional and technical communication

The research document required a certain level of professional and technical communication in order to be of an acceptable standard. A technical format obtain through the years of study at the department of Civil Engineering was followed. A large number of figures and tables were used to summarise and communicate data and information. In conclusion this thesis was compiled for a technical audience and a poster is to be presented for a non-technical audience.

ECSA outcome 8: Individual, team and multi-disciplinary working

This research project was an individual assignment and therefore most of the work was performed on an individual basis, thus a level of competence had to be shown in working effectively individually, although guidance from my supervisor was provided. The project had a clear objective of obtaining a OD matrix estimation procedure for public transport and finding the potential uses of MyCiti AFC data. All tasks were completed on time and deadlines met.

ECSA outcome 9: Independent learning ability

Independent learning was achieved through a comprehensive review of literature regarding procedures for OD estimation using AFC data and a number of case studies were reviewed. Furthermore, independent learning was achieved through a desktop study of the applying the trip-chaining algorithm and obtaining transfer OD flows among other things.

ECSA outcome 10: Engineering professionalism

Engineering professionalism was demonstrated through the interviews that were had with the industry expert as well as the research project supervisor. Furthermore, communication with the relevant engineering professionals was conducted in a professional and ethical manner. All deadlines required by this research project were met.