Implementation of finite mixture models for route choice estimation in large metro networks

Tamás Nádudvari

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
Institute for Transport Studies

February 2020

"The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others."

Chapter 4 of the thesis includes work of the following conference paper, of which the author drafted the papers and the co-authors provided commentary and comments throughout:

> **Nadudvari, T.** Liu, R. and Balijepalli, C. 2016. The reasonable route choice set in large and complex metro networks; an implementation of the K-shortest path algorithm for the London Underground, paper presented at the *21st International Conference of Hong Kong Society for Transportation Studies*, Hong Kong, China, 10th-12th December 2016

Chapter 5 and of the thesis includes work of the following conference papers, of which the author drafted the papers and the co-authors provided commentary and comments throughout:

> **Nadudvari, T.** Liu, R. and Hess, S. 2015. Modelling passengers' route choice behaviour on the London Underground: application of two choice modelling approaches, paper presented at the *47th Annual Conference of Universities' Transport Study Group*, London, United Kingdom, 5th-7th January 2015.
>
> **Nádudvari, T.** Liu, R. Balijepalli, C. Fu Q. 2019. The superstation representation of metro networks for overcoming data availability issues of station-to-station origin destination pairs – an application on the London Underground, paper presented at the *24th International Conference of Hong Kong Society for Transportation Studies*, Hong Kong, China, 14th-16th December 2019

Acknowledgements

“That which has been is what will be,
That which is done is what will be done,
And there is nothing new under the sun.
Is there anything of which it may be said,
“See, this is new”?
It has already been in ancient times before us.”

Ecclesiastes 1:9-10

I thank God for the PhD experience. I believe that the process He made me to pass through is much worthier than the PhD title itself. I thank Him for being near me in every moment. I thank for the brothers and sisters in the UK, Italy and all over the world who prayed for me and gave me spiritual advice in the difficult moments. I can’t find words to thank my wife, Mengting for her love, patience, understanding and support in all aspects. I thank my daughters: Dóra and Evelin, my father, mother, brother and sister.

Of course, I thank my supervisors Ronghui Liu, Chandra Balijepalli and Stephane Hess for leading me throughout these years step by step from the commencement of the PhD until the submission. I am especially thankful to Qian Fu, Dániel Hörcher, Jan-Dirk Schmöcker, Yu Jiang and Nigel Wilson that we had the opportunity to meet and that I could get specific help on my research topic. Surely, it helped me a lot to go ahead. I also thank János Tóth, Domokos Esztergár-Kiss and Csilla Szigligeti that they provided an office at the University where I could work on my thesis writing up, while I was in Budapest. I thank all staff and students at ITS Leeds, whom I met during these years.

I am thankful to those in Italy who gave me an opportunity to present my research and provided valuable feedback in different aspects: Angelo Martino and Davide Fiorello at TRT; Maurizio Arnone, Tiziana Delmastro and Brunella Caroleo at Links Foundation as well as Maurizio Agrillo and Nadia Dovano at GTT. Many thanks to the Engineering and Physical Sciences Research Council as well as to the University of Leeds for funding my PhD study.

Abstract

This thesis contributes to the research area of route choice estimation with smart card data in large metro[1] networks by addressing the issues with finite mixture models.

The motivation for this research comes from the problem that public transport authorities need to know passengers' route choice for their key functions. Recently, many cities adopted smart cards, which produced a wealth of data for researchers. However they reveal only the entry/exit station, not the chosen route.

Within the scope of this research is to address the following research problems:

Firstly, to propose a model that generates automatically the route choice set for all types of OD pairs in a metro network by finding a set of shortest routes with the K shortest path algorithm, and narrowing down this set by applying the generalised cost proportion of routes as the attribute cut-off.

Secondly, to introduce the concept of superstations by grouping those stations from/to which passengers have similar route choice patterns; and to aggregate the Observed Journey Times (OJT) of station-to-station OD pairs, so that the finite mixture model can be applied on a larger dataset.

Thirdly, to investigate the question of fail-to-board delays in two aspects: considering that at different origin stations, the fail-to-board delays may be different; as well as updating the route choice estimates, with the information on the fail-to-board delays along different routes.

The methodologies are illustrated through the case studies on the London Underground (LU) network, using Oyster data.

This research could enable a broader implementation of route choice estimation in large metro networks, especially when researchers can only rely on open data.

[1] In different cities different terms are used for the metro mode: "underground" or "tube" in London, "subway" in the cities of the United States, "metro" in many other cities (e.g. Paris, Shanghai)

Table of Contents

List of Figures

List of Tables

List of Acronyms

AEI:	Access Egress Interchange
AFC:	Automatic Fare Collection
CCOJT:	Centroid-to-Centroid adjusted Observed Journey Time
DLR:	Docklands Light Railway
EM:	Expectation-Maximisation
LO:	London Overground
LU:	London Underground
MNL:	Multinomial Logit
NR:	National Rail
OD:	Origin Destination
OJT:	Observed Journey Time
PDF:	Probability Density Function
PTAM:	Passenger-to-Train Assignment Model
RODS:	Rolling Origin Destination Survey
SJT:	Scheduled Journey Time
TAM:	Transit Assignment Model
TfL:	Transport for London

Chapter 1
Introduction

1.1 Background

Metro networks can be illustrated as the arteries of the public transport network of a city and its metropolitan area. Firstly, because they provide a faster and more reliable connection for a larger volume of passengers than the surface transport (i.e. bus, tram). Furthermore, the metro map often serves as a guidance to visitors and tourists to orient themselves in the city (Guo, 2011). Therefore, to maintain the high standard of metro networks, it is crucial to have adequate information on passengers' route choice, so that this information can be used for the key planning and operational tasks.

Conducting manual surveys to observe passengers' route choice is expensive as it requires human workforce and can reach only a small sample at certain stations and time periods. In the past two decades, many public transport operators have adopted Automated Data Collection Systems, such as Automatic Fare Collection (e.g. smart card), Automatic Vehicle Location (e.g. train tracking data from the signalling system), Automatic Passenger Count (e.g. sensors at platforms) and mobile services (e.g. cell phone data, WiFi, Bluetooth), which provide a larger data sample for the whole network from a longer time period (Koutsopoulos et al., 2017).

Among these data sources, smart card data have been widely used by researchers to understand passengers' route choices in metro networks. More recently, initiatives have been made to use the data extracted from the connection request of the passengers' devices to the WiFi access points (Transport for London, 2017) for the same purpose. The undoubted advantage of WiFi data is that passengers can be traced throughout their entire journey, therefore route choice can be directly observed; unlike smart card data, which reveals only the entry and exit station of the passengers and requires an appropriate model for the route choice. Whilst the collection and analytics of WiFi data is still in experimentation/pilot stage, there is the need to continue to learn how route choice can be modelled from more established, widely available automated data sources, such as public transport smart card data.

The question of modelling passengers' route choice from smart card data has been addressed by researchers at different levels of detail. As one extremity, approaching this question at network level, the additional information on journey time observations can be used to calibrate the parameters of discrete choice models. The issue here is, that in large and complex metro networks, different origin destination (OD) pairs may have different decision rules; and using the smart card data of the whole network to calibrate one utility function could not account for these differences across OD pairs. Theoretically, it could be possible to use OD specific route choice models and to calibrate them accordingly, however the definition of these OD pair categories is not always straightforward. As the other extremity, at an individual level, each passenger can be assigned to a train, knowing his/her entry/exit time to/from the metro network from smart card data and the departure/arrival time of trains from train tracking data. The challenge here is, that due to the large amount of data and to the more detailed representation of the problem, these models may require exceedingly high computational times. This implies the necessity to explore those methods, which can estimate route choice between these two extremities at OD level: estimate the route choice of an OD pair from the smart card data of that same OD pair.

In order to estimate route choice at OD level, the Observed Journey Time (OJT) distribution from smart card data can be analysed with appropriate statistical methods, such as the Kolmogorov-Smirnov test or the finite mixture model. While the former approach is limited to specific OD problems (Tirachini et al., 2016), the latter can be used for any type of OD pairs. Therefore the focus in this thesis is on the application of finite mixture models. The key concept here is that the distribution of OJTs of an OD pair can be decomposed as a mixture of the journey time distributions of the corresponding routes and the proportion of each component can be associated with the aggregate route choice.

Applying finite mixture models in complex metro networks, the following issues arise.

The first issue is, that most of finite mixture models require the number of components as an input; and setting it incorrectly, the model may give unrealistic results. In the context of route choice, the number of mixture components corresponds to the number of reasonable routes, which could be understood from the route choice set of the OD pair. Determining that in a complex metro network is a challenging question as theoretically there might be many possible routes, but only a few of them are reasonable.

The second issue is, that although a massive amount of data is available for the whole metro network, this sample for station-to-station OD pairs is very few. This is especially crucial, when only open data is available for the researcher, which contains only a smaller sample of all cardholders for a shorter time period of observation. Applying the finite mixture model on a small and not well distributed OJT sample may not give reliable results.

The third issue is related to the fact that a longer OJT does not necessarily mean that the passenger has taken the longer route, but it can also correspond to fail-to-board event, which is especially crucial in peak times when trains and platforms are overcrowded.

1.2 Research scope and objectives

Based on the research background presented in the previous section, the overall aim of this research is to develop a model, which can give more reliable estimates of route choice from automated data sources, such as smart card data. This could serve as a powerful tool for public transport operators to gain a better understanding on passenger flows in metro networks as it can replace or complement existing manual surveys. Currently, smart card data is still processed off-line (Koutsopoulos et al., 2017), therefore the proposed model is still limited to off-line functions, such as timetable planning or behavioural change communications. However, it is expected that in a few years' time it will be possible to have the technology for real-time data processing, and hence the proposed model could be applied for providing real-time information to passengers on the crowding along the lines and at the stations.

The work in this thesis brings forward the existing research on the application of finite mixture models for route choice problems; by addressing the above described issues that arise, when they are applied in the context of large and complex metro networks. To address these issues the following objectives are set:

- Develop a route choice set generation model that can find automatically the set of reasonable routes for different types of OD pairs within a metro network;
- Establish rules to group OD pairs with similar properties, so that the corresponding OJTs can be aggregated, and hence a larger data sample can be obtained; and
- Refine the data aggregation and route choice estimation method so that it can also account for fail-to-board delays at origin and at interchange stations

1.3 Methodological framework and outline of the thesis

The methodological framework of the thesis is presented on **Figure 1-1**. The rest of this thesis is structured as follows:

Chapter 2 sets the context for the thesis by introducing the problem of route choice estimation in metro networks. After describing the classical approaches; it presents a literature review on recent studies on route choice estimation from smart card data. The purpose for this is to select the implemented method and to identify the gaps in the field, which thesis is to address.

Chapter 3 focuses on the selected approach for route choice estimation: the finite mixture model. After describing the formulation and solution method, its convergence and validation is discussed. Applying it on the case study OD pairs of the London Underground (LU) network, and based on the results three major issues are raised (marked with **blue** on **Figure 1-1**):

1) It requires the number of mixture components (i.e. reasonable routes) as an input
2) Few smart card data available for station-to-station OD pairs
3) Longer OJT can mean either longer route or fail-to-board delay

These three issues are addressed in the subsequent chapters of the thesis.

Chapter 4 addresses the first issue. Reviewing route choice set generation methods, the chosen approach is presented: K shortest path algorithm for pathfinding followed by the application of the attribute cut-off based on the generalised cost of routes.

Chapter 5 addresses the second issue. Reviewing existing approaches for station grouping; the concept of superstations is introduced, referring to groups of stations from/to which passengers have similar route choice patterns. Following this, a method is presented to adjust the OJTs to superstation centroids and to aggregate them spatially; this way obtaining a larger data sample. Finally, the previously presented finite mixture model is applied on this larger dataset of OJTs to evaluate the benefits of the superstation representation.

Chapter 6 and **Chapter 7** addresses the third issue. **Chapter 6** focuses on fail-to-board delays at origin stations. Following a literature review, the quasi-dynamic approach is selected for inferring fail-to-board delays. A further OJT adjustment method is proposed, to take into consideration the difference in fail-to-board delays at different origin stations.

Chapter 7 focuses on the fail-to-board delays at interchange stations. A Bayesian approach is formulated to update the route choice estimates of the finite mixture model with the additional condition on fail-to-board delays.

The results of this analysis are three sets of route choice results depending on the consideration of the OD pairs and of the fail-to-board delay (marked with **green** on **Figure 1-1**):

- Station-to-station OD pairs (**Chapter 3**)
- Superstation-to-superstation OD pairs (**Chapter 5**)
- Superstation-to-superstation OD pairs consideration also the fail-to-board delay (**Chapter 6** and **Chapter 7**)

Chapter 8 concludes the thesis and proposes questions for further research.

.

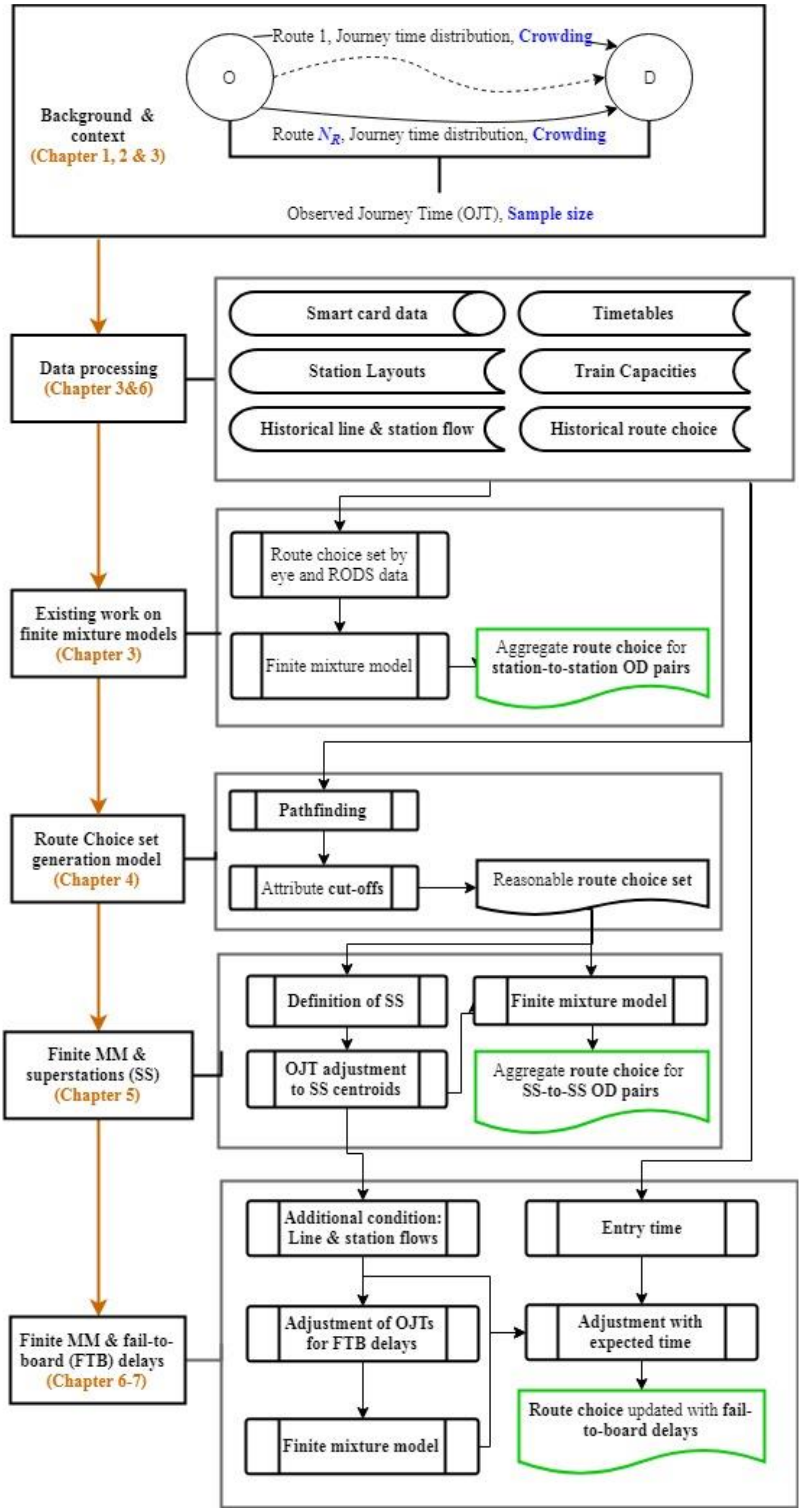

Route 1, Journey time distribution, Crowding
O
D
Background & context
(Chapter 1, 2 & 3)
Route N_R, Journey time distribution, Crowding
Observed Journey Time (OJT), Sample size
Data processing
(Chapter 3&6)
Smart card data
Timetables
Station Layouts
Train Capacities
Historical line & station flow
Historical route choice
Existing work on finite mixture models
(Chapter 3)
Route choice set by eye and RODS data
Finite mixture model
Aggregate route choice for station-to-station OD pairs
Route Choice set generation model
(Chapter 4)
Pathfinding
Attribute cut-offs
Reasonable route choice set
Finite MM & superstations (SS)
(Chapter 5)
Definition of SS
Finite mixture model
OJT adjustment to SS centroids
Aggregate route choice for SS-to-SS OD pairs
Finite MM & fail-to-board (FTB) delays
(Chapter 6-7)
Additional condition: Line & station flows
Entry time
Adjustment of OJTs for FTB delays
Adjustment with expected time
Route choice updated with fail-to-board delays
Finite mixture model

Figure 1-1 Structure of the thesis

1.4 Contributions

In this thesis two types of contributions can be distinguished. First and foremost, this thesis presents methodological contributions that makes a step towards the application of smart card data for estimating route choice in complex metro networks (**Section 1.4.1**). Additionally, there are notable technical contributions that are applied in the program code of the proposed algorithm for network representation and pathfinding (**Section 1.4.2**).

1.4.1 Methodological contributions

The work of this thesis

I. makes a step toward setting a general rule that can be applied to generate the reasonable route choice set for all types of OD pairs of a metro network;

II. introduces the concept of working with groups of stations for the purpose of overcoming data availability issues for station-to-station OD pairs; and

III. brings actual observations of journey time and crowding into passenger flow estimation models.

1.4.2 Technical contributions

The program code of the proposed algorithm

IV. creates the matrix of link times automatically from the input data of on-board times and headways of metro lines as well as of the access egress interchange times at metro stations; and

V. tailors the K shortest path algorithm for metro networks by creating the function to eliminate additional links in the network model, to avoid the generation of routes, which differ only in their access egress interchange movement.

Chapter 2
Route choice estimation in metro networks

2.1 Introduction

Metro networks of large metropolises (e.g. London, Shanghai, New York) are really complex, with many lines and stations serving the city and its metropolitan area. Metro services are usually high-frequency, high-capacity services that provide a faster connection than surface transport (i.e. bus and tram); as normally they run on dedicated tracks under or above the ground level, physically separated from other vehicle and pedestrian traffic, and the distance between its stations is longer than for other surface public transport modes, which allows a higher commercial speed. Therefore it is a convenient alternative for commuters; as well as for visitors and tourists, who often orient themselves by the metro map in the city (Guo, 2011).

This key role of metro networks within a metropolis and the associated high standards they should meet brings daily challenges to operators and transport authorities. One of the most crucial challenges is the problem of crowding, which occurs daily in the morning and afternoon peak as well as during special events (Parkes et al., 2016) or disruption (Freemark, 2013). This challenge determines the long and short term key functions; such as planning of new lines, vehicle and crew scheduling, behavioural change information to passengers, ticket pricing, revenue distribution and response to disruption. The final objective in performing these duties is, that metro networks could provide a reliable service, obtaining greater customer satisfaction and attracting private car users to choose more sustainable modes, this way relieving congestion from the roads. The main building block for these key functions and objectives is the adequate information on passenger flow on the metro lines and through the station passageways (Koutsopoulos et al., 2017).

Nowadays still many operators conduct manual surveys (e.g. questionnaires with passenger counts) to gain a better understanding on passenger flow (e.g. the Rolling Origin and Destination Survey (RODS) in the London Underground (LU), see **Section 3.6.3**). The advantage of these data sources is, that in addition to passenger flow, they can also collect information on the socio-demographic background of respondents as well as on trip purpose. However they are very expensive as they need to use human workforce for data collection and processing. Furthermore, they can reach only a small sample of the total population at limited number of stations and time periods.

In recent years, initiatives have been made to explore further, how passenger flow can be observed from automatic data collection systems, such as WiFi (Transport for London, 2017) or cell phone location data (Holleczek et al., 2015). Whilst these more advanced technologies can provide detailed tracking of individuals' movements in a metro network, they are still in experimentation/pilot stage. Therefore there is still the need to learn how passenger flow can be understood from more established, widely available automated data sources, such as public transport smart card data.

The main issue with smart card data is, that it records only the entry and exit location of passengers, but it is unknown, how they moved within the metro network. Therefore a model needs to be developed to estimate route choice and hence passenger flow based on the available information from smart card data.

The rest of this chapter is structured as follows. Firstly, in **Section 2.2**, the classical route choice modelling approach is summarised with the purpose to point out the need for innovative methods that can introduce automated data in the estimation process. Following this, in **Section 2.3**, a literature review is presented on recent methods that estimate route choice from smart card data. This chapter is concluded by a discussion in **Section 2.4**, where these methods are compared, the relevant issues in complex metro networks are highlighted and the implemented method is selected.

2.2 Discrete choice modelling methods

In general, modellers have relatively adequate information from the available data on the properties of the metro network and on OD demand. The main modelling challenge is to estimate passenger flows by assigning this OD demand to the metro network. This task is also mentioned as the fourth stage in the four-stage modelling (Ortúzar and Willumsen, 2011). In the context of metro networks, this stage is called the Transit Assignment Model (TAM).

A comprehensive explanation on the theory of TAMs can be found in Gentile and Noekel (2016)(pp. 287-481). In essence, TAMs include the following sub-models (**Figure 2-1**):

(i) generalised cost function of routes;

(ii) route choice set generation;

(iii) route choice estimation and

(iv) link loading.

The literature review of the different sub-models are organised according to the structure of this thesis. The considerations for generalised costs (sub-model (i)) are discussed together with the route choice set generation model (sub-model (ii)) in **Chapter 4**. The sub-model of link loading (iv), in relation to the question of capacity constraints is presented in **Chapter 6**. This section focuses on route choice estimation (sub-model (iii)) with discrete choice modelling methods.

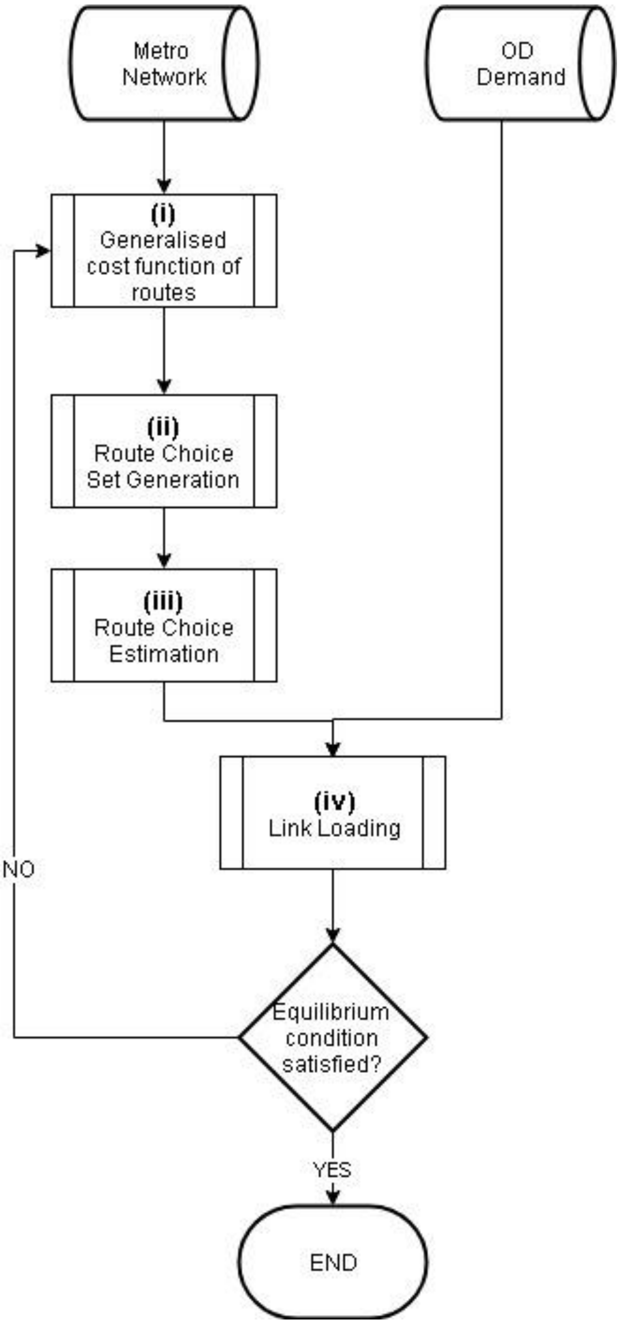

Metro Network
OD Demand
(i) Generalised cost function of routes
(ii) Route Choice Set Generation
(iii) Route Choice Estimation
(iv) Link Loading
NO
Equilibrium condition satisfied?
YES
END

Figure 2-1 The sub-models of transit assignment models

2.2.1 The random utility theory

Traditionally, discrete choice modelling methods are used for route choice estimation. The word “discrete”, refers to the fact that individuals can choose from a finite set of alternatives.

The theoretical framework in discrete choice modelling methods is the random utility theory (Domencich and McFadden, 1975), which postulates that the decision rules that individuals make are compensatory. This means, that each alternative has various attributes, and the good performance on one attribute of an alternative compensates the poor performance on its other attributes. The utility of an alternative is given by adding up these performances, and it is assumed that individuals choose the alternative with the highest utility.

However the utility of an alternative cannot be completely measured as human behaviour is not always deterministic and also the modeller does not possess complete information on the decision process that individuals make. Unobserved or incorrectly measured attributes, unobserved alternatives as well as the variation of preferences can bring in uncertainties in the system which cannot be captured by the modeller. The utility (u_k) of alternative k has two parts: a measurable part (v_k), which represents the attributes that the modeller can measure; and a random part (ϵ_k), which represents the above discussed uncertainties:

$$u_k = v_k + \epsilon_k \tag{2-1}$$

The random utility theory assumes that individuals choose the alternative with the highest utility. However due the random part introduced into the equation, modellers do not have the perfect information, which alternative has the highest utility. For this reason, they cannot approach the question of choice deterministically, stating that all individuals choose the alternative with the highest measurable utility. They only can approach this question stochastically, stating that the alternatives with higher measurable utility will be chosen with higher probability.

2.2.2 The Multinomial Logit (MNL) model

The basic model that uses the random utility theory is the Multinomial Logit (MNL) model (Domencich and McFadden, 1975). Assuming that the random part (ϵ_k) follows a Gumbel distribution, and that it is distributed independently and identically across alternatives and respondents; it can be mathematically proved that the formula for the choice probability contains only the measurable part of the utility, but not the random part. The probability (p_k) that alternative k is chosen is:

$$p_k = \frac{exp(v_k)}{\sum_{k' \in K} exp(v_{k'})} \qquad (2\text{-}2)$$

where k' represents all the alternatives in the choice set K.

2.2.3 Route choice models

Route choice modelling is a more complex task than other choice modelling tasks in transport (e.g. mode, car ownership), due to two main reasons. One reason is, that although there are many theoretically possible routes between an OD pair, only a few of them are reasonable (see **Chapter 4**). Another reason is the question of overlapping routes, which should not be considered as completely distinct alternatives, but their degree of overlapping – which is called correlation – should be modelled.

Prato (2009) gives a comprehensive review on route choice models, where two main approaches are presented to include correlation in route choice models. One approach is to introduce a correction term in the measurable part of the utility maintaining the MNL structure. Such models are the C-Logit (Cascetta et al., 1996) and the Path Size Logit (Ben-Akiva and Bierlaire, 1999). Another approach is to create a model specification, which includes correlation between the random parts of the utilities of the different alternatives. The Nested Logit (Williams, 1977) and its improvements, such as the Cross Nested Logit (Vovsha, 1997) and the Generalised Nested Logit (Wen and Koppelman, 2001) models are developed for modelling this correlation.

Furthermore, route choice in the context of public transport networks is even more a complex issue, where the key difficulty lies in the question of interchanges: A public transport route, in fact, can include two or more separate on-board trips with the corresponding interchanges in between, which have substantially different characteristics in the utility perception of passengers. While during the on-board trips the passenger is standing or sitting inside the vehicle and can utilise that time for a short activity (e.g. reading a book, checking e-mails), interchanges require walking between platforms often including stairs or escalators. This is not present in the road based context, where – even though there are junctions between the road segments – the route can be considered with homogeneous characteristics.

2.2.4 Parameter calibration of choice models

Once the utility function of the choice model is formulated, the task is to calibrate its parameters. The classical method for parameter calibration is the maximum likelihood method, which requires data on passengers' route choices. It sets the value of the questioned parameters, so that the route choice estimated with the choice model should reproduce the choices understood from the data: For each data observation, the choice probability of the chosen alternative is calculated in function of the parameters, and then the log-likelihood is obtained by summing the logarithms of these probabilities. The objective function is to find those parameters, which maximise the log-likelihood.

2.2.5 Discussions

The key point for parameter calibration is to have the appropriate understanding on the data source that is used. In the field of choice modelling, it is quite popular to use Stated Preference surveys. The main issue here is, that it does not give information on actual observations of passengers; but it comes from hypothetical scenarios. It is also shown, that passengers do not always respond the same in those scenarios as they would act in the real-life situations (Fifer et al., 2014).

In order to introduce actual observation on passengers' choices, Revealed Preference surveys are used (i.e. RODS data in the LU, cf. **Section 3.6.3**). Its main limitation – apart from the high cost of manual surveys – is, that it is collected over a long period, different years at different stations, usually only in certain times of the year; therefore it could not reflect route choice of a specific time period, it could only serve as an average value.

Recently, pilot studies have been conducted to explore how route choice can be observed from extrinsic mobility data, such as WiFi or cell phone location data (cf. **Section 2.1**), which could be also used as a source for parameter calibration. However, in this case, even though these data sources reveal the chosen route of passengers; their journey time – which is an important attribute in their utility function – cannot be fully understood, as only the connection time to the access point is recorded, not their exact entry/exit or boarding/alighting time.

In this research the focus is on exploring how route choice can be inferred from intrinsic mobility data, such as smart card data. As mentioned earlier (**Section 2.1**), smart card data reveals only the entry and exit time of passengers, not their chosen route; therefore an appropriate model is required for its estimation. As it follows, modelling approaches are reviewed that can infer route choice from smart card data (**Section 2.3**).

2.3 Estimating route choice in metro networks with smart card data

The original purpose of smart cards is to introduce a smarter way of fare collection, replacing cash or paper tickets and reducing fare evasion. For this reason they are also called Automatic Fare Collection (AFC) systems. Smart card data is generated as a by-product of AFC systems. Its great advantage with respect to manual surveys is; that even though the installation of ticket gates requires a capital cost; once they are in operation, they can collect data from the whole network for a continuous time period, obtaining a large sample of data at a low marginal cost (Chu, 2010; Pelletier et al., 2011; Koutsopoulos et al., 2017).

Since the beginning of the 2000s, many cities all over the world have adopted the AFC systems, which generated a wealth of smart card data available for further analysis. This data have been applied for various modelling tasks, among which notable research has been done regarding OD matrix estimation. This includes scaling up OD matrices in metro networks (Gordillo, 2006; Chan, 2007); inferring alighting stops in AFC systems, which record only the boarding stop (i.e. buses and in some metro networks) (Barry et al., 2002; Zhao et al., 2007; Cui, 2006; Wang et al., 2011; Trépanier et al., 2007; Munizaga and Palma, 2012); linking trips to obtain multimodal journey OD matrix (Seaborn et al., 2009; Munizaga and Palma, 2012; Gordon et al., 2013; Nassir et al., 2015a); as well as the application of OD matrices for bus route choice model calibration (Jánošíková et al., 2014) and for inferring mode choice patterns for zonal OD pairs (Viggiano et al., 2016).

Furthermore, smart card data have been applied also to measure service reliability (Chan, 2007; Zhao et al., 2013; Leahy et al., 2015; Silva, 2017; Ross, 2017), to identify trip purpose (Utsunomiya et al., 2006; Morency et al., 2007; Ortega-Tong, 2013; Kusakabe and Asakura, 2014) as well as to model wait time distribution of passengers (Wahaballa et al., 2017; Ingvardson et al., 2018) and their behaviour during service disruption (Freemark, 2013; Ross, 2017).

This section focuses on the application of smart card data for route choice estimation. The literature is classified into three main categories, according to the detail level of estimation (see **Table 2-1**). The least detailed level is the "network level" (see **Section 2.3.1**), where the route choice model still maintains the structure of random utility models and it uses smart card data for parameter calibration. The next detail level is called "OD level" (see **Section 2.3.2**), where the model moving away from the structure of random utility models, estimates route choice of an OD pair from the smart card data of that same OD pair. Finally, the most detailed level is the individual level (see **Section 2.3.3**), where disaggregate smart card data is applied together with train tracking data and each passenger is assigned to a train.

Table 2-1 Overview of methods that estimate route choice from smart card data

Section	**Level of estimation**	**Method**	**Reference**	**Case study network**
2.3.1	Network	Calibrating logit model parameters	Sun et al. (2015)	Singapore
			Xu et al. (2018)	Shanghai
2.3.2	OD	Finite mixture models	Sun and Xu (2012)	Beijing
			Fu (2014)	London
			Lee and Sohn (2015)	Seoul
		Kolmogorov-Smirnov statistics	Tirachini et al. (2016)	Singapore
2.3.3	Individual	Passenger-to-Train Assignment	Paul (2010)	London
			Hong et al. (2015)	Seoul
			Hörcher et al. (2017)	Hong Kong
			Zhu et al. (2017)	Hong Kong

.

2.3.1 Network level

In metro networks, smart card data does not reveal explicitly the chosen route of individuals, only their entry and exit time. Therefore the parameters of random utility models cannot be calibrated with the maximum likelihood method (cf. **Section 2.2.4**); but it requires more advanced techniques, such as the Bayesian framework, where the Observed Journey Times (OJT) from smart card data are used as an input.

Such approach was used in Sun et al. (2015) for the Singapore metro network, where the parameters of on-board and interchange time were calibrated in a MNL model. They included also reliability in the model, stating that trains may not be running on time. Therefore not only the parameters, but also the journey time attributes were left as an unknown, and it was estimated in an integrated Bayesian approach. They used OJTs from smart card data to update the priors of the parameters, where the priors came from their earlier study (Sun et al., 2012).

Xu et al. (2018) in addition to travel time took into consideration also crowding attributes in the MNL in terms of standing and fail-to-board proportions. To calibrate its parameters, they used also historical train loading data, operator's information on train properties (number of seats, maximum capacity) and timetables as observations apart from smart card data. They applied the case study for the Shanghai metro network.

Applying one utility function for all the OD pairs of a metro network, one would assume, that the order magnitude of their attributes (i.e. journey time components) are similar. However, in reality, this is not always the case. This can be illustrated through the example of two OD pairs in the LU (see **Figure 2-2**). The **Victoria** – **Holborn** OD pair represents a trip within Central London. There, the distances and hence the on board times are very short (5-10 minutes); but the interchange stations (**Oxford Circus** and **Green Park** stations) are very complex, therefore interchange times are relatively long (3-4 minutes). On the contrary, the **Stanmore** – **Bond Street** OD pair represents a trip from Outer London to Central London, where on-board times are longer (30-40 minutes); but interchange times are not an issue at all, because the interchange happens between adjacent platforms (at **Wembley Park** and **Finchley Road** stations). From this it can be learned that the relationship between interchange time and utility is not always linear, but would require a more detailed function specification, so that it could be applicable for all types of OD pairs in a metro network. This function specification, apart from the interchange time, would include also other attributes, such as the case of adjacent platforms or the presence of escalators (Raveau et al., 2014). This implies the necessity to go towards those methods that can estimate route choice of an OD pair from the smart card data of that same OD pair (see **Section 2.3.2** and **Section 2.3.3**).

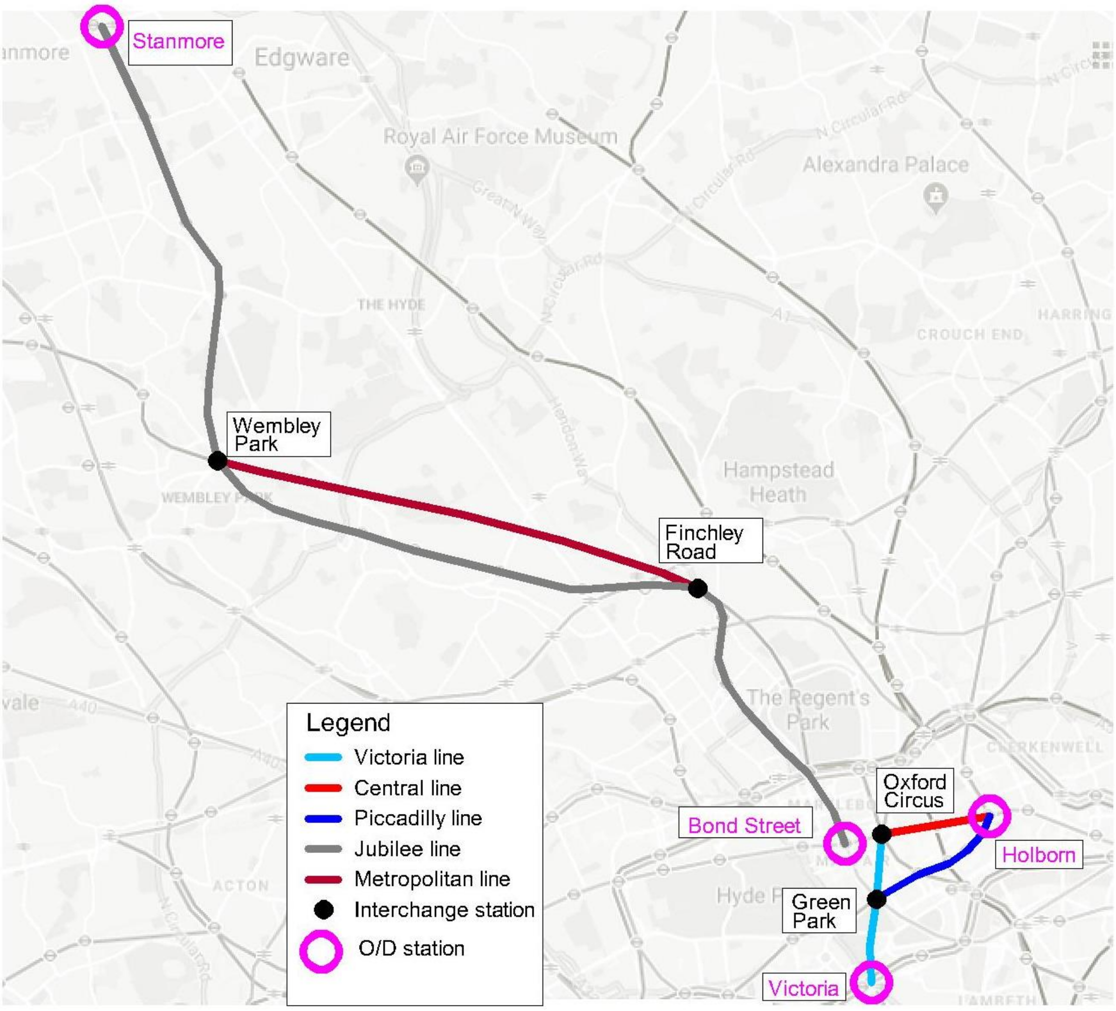


Figure 2-2 Route choice patterns of different OD pairs in the London Underground, presented on a geographical map

2.3.2 OD level

For methods that estimate route choice at OD level, the input data is the empirical distribution of OJTs, known from smart card data. The OJT distribution is, in fact, the mixture of the OJTs of passengers travelling on different routes. This formulates the problem to establish the connection between the OJT of a passenger and his/her chosen route. One possible method to solve this problem is to apply finite mixture models. The key concept of finite mixture models is, that the empirical OJT distribution is estimated as a mixture of component distributions. In this setting, the connection can be established between the mixture components and the actual routes of the given OD pair in the following way:

1) Number of mixture components corresponds to the number of reasonable routes

2) The statistical distribution of mixture components (e.g. Gaussian, lognormal) corresponds to the journey time distribution of routes
3) The proportion of the mixture components corresponds to the aggregate route choice probabilities

Sun and Xu (2012) applied the finite mixture model for an OD pair in the Beijing metro. There they assumed the number of routes to be known from the map. Furthermore, they used timetables and manual surveys to calculate the journey time distribution of routes. In their model formulation only the component proportions remained unknown, which could be solved analytically.

Fu (2014), applying the finite mixture model in the LU, also assumed the number of routes to be known from map and RODS data; however in his model both the journey time distribution and the choice probability of the routes were estimated with the finite mixture model. Having both the parameters of the mixture components and their proportion unknown, the problem could not be solved analytically, but the Expectation-Maximisation algorithm (Dempster et al., 1977) was applied for the numerical estimation of the results.

In the model specification of Lee and Sohn (2015), in addition to the journey time distribution and choice probability of routes, also the number of routes were treated as an unknown. To estimate all parameters in the finite mixture model, a more advanced solution algorithm, a reversible-jump Markov chain Monte Carlo simulation is applied, following the concept in Richardson and Green (1997).

These three examples on finite mixture models are compared in **Table 2-2**.

Tirachini et al. (2016) focused on the specific route choice problem of travelling forward or backward for getting a seat. There they could simplify the problem by including the additional constrain of travel time difference, as all alternatives are on the same line. With this simplification, journey time distribution and choice probabilities of routes could be obtained even without the need of applying finite mixture models. They used the Kolmogorov-Smirnov statistics instead.

In reality, a longer OJT may not be necessarily attributed to a longer route; but it can be caused by various other reasons, such as failure to board, carrying a heavy luggage or microscopic station features (longer path within the station, further ticket gate, alighting from the other end of the train). Relying purely on the OJTs, the actual reason for the longer journey time cannot be fully understood. This implies the necessity to review those methods that can estimate route choice at a more detailed, individual level (see **Section 2.3.3**).

Table 2-2 Comparison of finite mixture models; input and unknown parameters; solution methods

Reference	**Number of routes**	**Journey time**	**Route choice**	**Solution method**
Sun and Xu (2012)	Input	Input	Unknown	Explicit
Fu (2014)	Input	Unknown	Unknown	Expectation-Maximisation
Lee and Sohn (2015)	Unknown	Unknown	Unknown	Markov Chain Monte Carlo

2.3.3 Individual level

Estimating route choice at individual level means that from smart card records not only the OJT distributions of OD pairs are extracted, but also the entry/exit timestamps of individual passengers. Using this more detailed data together with the additional information on the departure/arrival time of trains, each individual is assigned to a train. Therefore this method is called Passenger-to-Train Assignment Model (PTAM).

Earlier PTAMs (Kusakabe et al., 2010; Xu and Zhou, 2012) used timetable information for the departure and arrival time of trains, assuming that trains run on time, which might not be valid in many cases. More recent models (Paul, 2010; Hong et al., 2015; Hörcher et al., 2017; Zhu, 2017) – moving away from scheduled times and taking into consideration delay of trains – worked with actual departure and arrival times known from the train tracking data, which is generated by the signalling system.

As the first step of the PTAM, a set of feasible itineraries are identified. An itinerary is considered feasible, if its first train departs after the entry time and its last train arrives before the exit time of the passenger. Following this, passengers are assigned to these feasible itineraries. The assignment process can be either deterministic (Kusakabe et al., 2010; Xu and Zhou, 2012; Paul, 2010) or probabilistic (Hong et al., 2015; Hörcher et al., 2017; Zhu, 2017).

The key problem in PTAM is that – in case of OD pairs with multiple reasonable routes or interchanges – there are too many feasible itineraries, which makes the computational process more complicated. Therefore, first this set needs to be narrowed down, and then passengers can be assigned to the remaining itineraries.

Kusakabe et al. (2010) – focusing on the urban rail network of Osaka, Japan – proposed to narrow down the set of feasible itineraries by using the criteria of minimising the total of wait time at the entry station and the lost time at the exit station as well as of excluding itineraries with unreasonable interchanges.

Xu and Zhou (2012) – taking the Beijing metro network as an example – introduced the concept of matching degree, which is calculated based on the time between the arrival of the passenger to the platform and the departure of the train. They assigned passengers to itineraries with the highest matching degree. Another novelty in their method is that the algorithm calculates backwards, starting from the exit station and going towards the entry station, as they highlighted that passengers are less likely to experience delays at the exit station. This approach has been followed in many subsequent studies.

Paul (2010) – applying the case study on the LU network – worked with the distribution of AEI times, which she obtained from smart card and train tracking data. In this process, she made the assumption, that at a given station the ratio of the access, egress and interchange time distributions is identical with the ratio of the corresponding times from the AEI survey of TfL. She further considered, that passengers walk with the same speed throughout all their journey. Using these AEI distributions, she excluded those itineraries, for which the expected access time is greater than the time between the entry of the passenger and the departure of the train; as well as those, which the expected transfer time is greater than the time between the arrival of the first train and the departure of the second train.

Hörcher et al. (2017) – applying a PTAM for the entire Hong Kong metro network – followed the approach of working with AEI time distributions. The novelty in their method is, that they used only automated data sources: smart card and train tracking data. They made the assumption that the egress time distribution is identical for all trips and hence they inferred the delayed access time and interchange time distributions. Following this, they assigned passenger to itineraries in a probabilistic setting according to the likelihood of the corresponding access, egress or interchange time.

Zhu et al. (2017), looking at OD pairs without route choice or interchange, estimated the access and egress time distributions of trips including those with one and more feasible itineraries. Using this, they inferred left behind probabilities (Zhu et al., 2018). This the work was extended for OD pairs with route choice and interchange (Zhu, 2017).

PTAMs have been used for operational tasks in metro networks, such as system performance measures, capacity utilisation of trains, crowding assessment at stations (Zhu, 2017) as well as for crowding cost estimation (Hörcher et al., 2017). Making one step ahead, Koutsopoulos et al. (2017) elucidated the possibility of running real time PTAMs to give short-term prediction on the loads on arriving trains, the expected spaces for newly boarding passengers, expected number of passengers at platforms and the expected number of passengers left behind.

For the accurate estimation of passenger flows, it would be necessary to run the PTAM on all OD pairs of the entire metro network, as crowding on one link can come from the demand of all OD pairs. However, this could lead to exceedingly high computational times, especially in complex metro networks. This was identified in Hörcher et al. (2017), where the PTAM on the Hong Kong metro with 1 day's smart card data required a run time of 2 days.

Despite the potential of PTAMs to analyse route choice at a more detailed level, there are still some microscopic station features (multiple paths between platforms and ticket gates, multiple station entrances, alighting through different doors of long trains), which cannot be captured in the model, only appropriate assumptions can be made.

2.4 Discussions

In this chapter the classical approaches for route choice estimation (i.e. discrete choice modelling methods) were presented. Within this framework it was established that the aim of this research is to move away from those methods and to explore how actual observations from intrinsic mobility data, such as smart card can be utilised to infer route choice and hence passenger flow.

For this purpose the literature was reviewed on methods that can infer route choice from smart card data; and they were classified according to the detail level of estimation: network, OD and individual level. Considering the size and complexity of the LU network, it was established, that – as different OD pairs may have substantially different decision rules – there is a necessity to estimate route choice at a more detailed level, rather than just calibrating the parameters of random utility models (cf. **Section 2.3.1**). On the other hand, estimating route choice at the level of individual passengers and trains is out of the scope of this thesis, as it would require a more complex network model and hence exceedingly high computational times (cf. **Section 2.3.3**). Based on this, the core objective of this thesis is set to explore, at what extent route choice can be understood at the OD level, using only the information of OJT distribution from smart card data (cf. **Section 2.3.2**).

As the case studies in the LU are not limited the problem of travelling forward or backward on the same line; the Kolmogorov-Smirnov statistics with the additional constraints of travel time difference (Tirachini et al., 2016) would not be applicable, but finite mixture models are required. Approaching this question from the prospective of reliability, it is not appropriate to use the scheduled journey time of routes as an input in the finite mixture model (Sun and Xu, 2012), but it is preferable to use methods which estimate both the journey times and the route choice with the finite mixture model itself. Among those methods the simpler Expectation-Maximisation method applied in Fu (2014) is already considered to be sufficient for the objectives set in this thesis. Estimating the number of reasonable routes also from smart card data (Lee and Sohn, 2015) is out of the scope of this thesis. This question is discussed in **Chapter 4**, where the focus will be on route choice set generation models based on network properties.

The formulation and solution method for the implemented finite mixture model is described in **Chapter 3**, where also case studies are presented. This gives a solid ground to understand the issues that are raised when these models are applied to estimate route choice in complex metro networks, such as the LU.

Chapter 3
Finite mixture models for route choice estimation and its application on the London Underground

3.1 Introduction

In **Chapter 2** – following a comprehensive literature review – it was established that this thesis focuses on modelling route choice from smart card data at the detail level of origin destination (OD) pairs. This means that the route choice of an OD pair is modelled from the Observed Journey Time (OJT) distribution of that same OD pair. More specifically, finite mixture models are applied for route choice estimation. Among the recent applications of finite mixture models Fu (2014) was chosen, where both the journey time distribution and choice probability of the routes are unknown and estimated by the model. In that setting only the route choice set (i.e. number of reasonable routes, which corresponds the number of mixture components) is supposed to be known and used as an input for the model.

The rest of this chapter is structured as follows: **Section 3.2** describes the formulation and solution method of implemented finite mixture model. **Section 3.3** discusses the parameters that influence the convergence of the model. **Section 3.4** presents how the results of the mixture model are matched with the actual routes on the London Underground (LU) network; and highlights the difference between the settings applied in Fu (2014) and in this thesis. The software implementation of the algorithms applied in the finite mixture model is resumed in **Section 3.5**.

Following this the finite mixture model is applied on the case OD pairs of the London Underground (LU). **Section 3.6** describes the data sources, and the case studies are presented in **Section 3.7**.

The purpose for these case studies is to point out the issues that arise, when the finite mixture model is applied in complex metro networks (**Section 3.8**). These issues are addressed in the following chapters of the thesis.

3.2 Formulation and solution method of finite mixture models for route choice estimation

In this section the implemented finite mixture model (Fu, 2014) is formulated and the solution method is presented. In this chapter notation is used as follows. As the methodology focuses on one OD pair, the index of OD pairs is omitted for all variables.

Variable identifiers

r	Index of a mixture component
q	Individual passenger
k	Index of an actual LU route
l, k	l-th journey leg[2] on route k
s, k	s -th interchange station on route k

Sets

R	Route choice set for an OD pair
Q	Statistical population of passengers travelling between origin and destination
Q_r	A subpopulation of passengers in Q travelling on route r

Variables

N_R	Number of routes in route choice set R
N_Q	Number of passengers in statistical population Q

[2] In the context of public transport networks a route may consist of two or more separate trips on different public transport lines with the corresponding interchanges in between. These trips are called "journey legs" (cf. **Section 2.2.3**).

$choice_{qr}$	Elementary event that passenger q have chosen route r
OJT	Initial dataset of Observed Journey Times (minutes)
OJT^0	Valid dataset Observed Journey Times (minutes)
δ_q^{OJT}	Journey time observation (OJT) of passenger q (minutes)
Δ_q	Elementary event that the OJT of passenger q is δ_q^{OJT}
δ_r	Random variable of journey time on route r (minutes)
$c_r(\delta)$	Probability density function of the journey time distribution δ_r, corresponding to component distributions
δ	Random variable of journey time for the OD pair (minutes)
$m(\delta)$	Probability density function of the journey time distribution of the OD pair (δ), corresponding to the mixture distribution
ω_r	Proportion of component distribution $c_r(\delta)$ in the mixture $m(\delta)$
μ_r	Mean journey time on route r (minutes)
σ_r	Standard deviation of journey time on route r (minutes)
n^{OJT}	Number of records in the OJT^0 dataset
$r^{[q]}$	Categorical variable, expressing the route r chosen by passenger q
OJT_r^{KMS}	Subset r of OJT^0 produced by the K-means clustering algorithm (minutes)
μ_r^{KMS}	Mean of sub-dataset OJT_r^{KMS} (minutes)
σ_r^{KMS}	Standard deviation of sub-dataset OJT_r^{KMS} (minutes)
ω_r^{KMS}	Proportion of sub-dataset OJT_r^{KMS} in dataset OJT^0
μ_r^{MIX}	Mean journey time for mixture component r (minutes)

σ_r^{MIX}	Standard deviation of journey time for mixture component r (minutes)
ω_r^{MIX}	Proportion for mixture component r
t_k^{SJT}	Scheduled Journey Time[3] of actual LU route k (minutes)
$t_{1,k}^{acc}$	Access time for the first journey leg of route k (minutes)
$t_{l,k}^{wait}$	Wait time for the first coming train on the l-th journey leg of route k (minutes)
$t_{l,k}^{ob}$	On-board time on the l-th journey of route k (minutes)
$t_{s,k}^{ic}$	Interchange time at the s-th interchange station of route k (minutes)
$t_{L,k}^{egr}$	Egress time from the last journey leg of route k (minutes)
$N_{L,k}$	Total number of journey legs on route k
$N_{S,k}$	Total number of interchange stations on route k
T^{entry}	Entry timestamp at origin station (minutes after midnight)
T^{exit}	Exit timestamp at destination station (minutes after midnight)
$f_{l.k}$	Frequency of trains on the l-th journey leg of route k (trains/hour)
ω_k^{RODS}	Aggregate choice proportions from the Rolling Origin Destination Survey (RODS) data for route k
n^{RODS}	Sample size of RODS data

Vector of variables

$\boldsymbol{\theta}_r$	Parameters of the statistical distribution for route r

[3] Based on timetables and station layouts (see **Section 3.6.2**).

$\boldsymbol{\vartheta}$	The collection of parameters $\boldsymbol{\theta}_r$ for all routes $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{N_R})$
$\boldsymbol{\mu}$	Random vector of all means $(\mu_1, \dots, \mu_{N_R})$ (minutes)
$\boldsymbol{\sigma}$	Random vector of all standard deviations $(\sigma_1, \dots, \sigma_{N_R})$ (minutes)
$\boldsymbol{\omega}$	Random vector of all component proportions $(\omega_1, \dots, \omega_{N_R})$

Functions

$\Pr(\cdot)$	Probability of an event
$\pi(\cdot)$	Probability function
$\ell(\cdot)$	Likelihood function
$\log \ell(\cdot)$	Log-likelihood function

3.2.1 Problem description

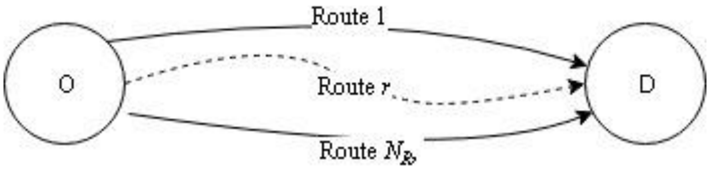


Figure 3-1 An OD pair with the set of reasonable routes

Given an OD pair (**Figure 3-1**), the route choice set between the origin and destination station is denoted by R, with r denoting a route, and the number of routes in R is equal to N_R. As the focus is on OD pairs with more reasonable routes, $N_R \geq 2$. In this setting R and N_R is assumed to be known as an input from the metro map of from surveys (i.e. Rolling Origin Destination Survey (RODS) for the LU).

The total statistical population of passengers willing to travel between the origin and destination station is denoted by Q, with q denoting an individual passenger and the number of passengers in Q is equal to N_Q. For simplicity, it is assumed, that each passenger q has the same route choice set.

The elementary event that passenger q have chosen route r is denoted by $choice_{qr}$. As the chosen route of individual passengers is unknown, it can be only described with probabilities as $\Pr(choice_{qr})$.

The dataset used for route choice estimation in this chapter is the distribution of Observed Journey Times extracted from smart card data (denoted by OJT). Firstly, it is checked, whether all records can be accepted as valid data by removing the outliers. Those entries are considered as outliers, which exceed the upper outer fence (i.e. three times interquartile range more than the third quartile) (Frigge et al., 1989). This valid dataset is denoted by OJT^0. Within this valid dataset, the OJT of passenger q is denoted by δ_q^{OJT}. The elementary event that the OJT of a passenger is δ_q^{OJT} is denoted by Δ_q.

One key point in Fu (2014) is to establish the connection between the events of $choice_{qr}$ and Δ_q in a Bayesian framework, working with conditional probabilities. $\Pr(choice_{qr}|\Delta_q)$ denotes the probability, that passenger q has chosen route r on condition that his/her journey time was δ_q^{OJT}. According to the Bayes theorem, this can be formulated as:

$$\Pr(choice_{qr}|\Delta_q) = \frac{\Pr(choice_{qr})\Pr(\Delta_q|choice_{qr})}{\Pr(\Delta_q)}. \tag{3-1}$$

In this setting, $\Pr(\Delta_q)$ is the total probability for each passenger q that his/her journey time is δ_q^{OJT} irrespective to his/her chosen route. According to the law of total probability it can be formulated as:

$$\Pr(\Delta_q) = \sum\nolimits_{r\in R} \Pr(choice_{qr})\Pr(\Delta_q|choice_{qr}). \tag{3-2}$$

In the Bayesian context $\Pr(choice_{qr})$ is called prior probability and it describes the probability that passenger q has chosen route r, without having any information on his/her journey time. This can be interpreted also as an average route choice probability.

$\Pr(\Delta_q|choice_{qr})$, is the likelihood function and it describes the probability, that the observed journey time of passenger q is δ_q^{OJT} (event Δ_q occurs) given the fact that he/she has chosen route r (event $choice_{qr}$ occurred). This corresponds to the probability density function of the journey time distribution for r.

The problem here is, that $\Pr(choice_{qr})$ and $\Pr(\Delta_q|choice_{qr})$ cannot be known explicitly for each individual passenger q. but a modelling approach is required for their estimation, which is discussed in **Section 3.2.2**.

3.2.2 Application of finite mixture models for route choice estimation

Finite mixture models have been applied in many fields of biological, physical and social science (McLachlan and Peel, 2000). The novelty in Fu (2014) is to apply it in the field of transport for the previously described route choice problem.

The statistical population of passengers (Q) can be decomposed to N_R number of subpopulations according to the number of routes in R. These subpopulations – denoted by Q_r – represent passengers on the same route r. The random variable of journey time produced by subpopulation Q_r, who travels through route r is denoted by δ_r. It follows a statistical distribution, which is denoted by $c_r(\delta)$, where letter "c" refers to the component distribution. In total there are N_R journey time distributions according to the number of routes in R.

In reality, when the routes are overlapping, the random variable of their journey time $(\delta_1, \dots, \delta_{N_R})$ are correlated, however for the simpler formulation of the model it is assumed that they are independent of each other. In this case, the random variable of the journey time for the OD pair can be called as δ and its distribution can be described as the joint probability distribution for all the route-specific journey times. It is denoted with $m(\delta)$ where letter "m" refers to the mixture distribution. It can be calculated as the weighted sum of the component distributions, $c_r(\delta)$ (Frühwirth-Schnatter, 2006):

$$m(\delta; \omega_1, \dots, \omega_{N_R}) = \sum\nolimits_{r \in R} \omega_r c_r(\delta) \,, \tag{3-3}$$

In this setting ω_r denotes the proportion of the component distribution, c_r in the mixture. It describes how likely it is that the probability of any individual's journey time observation, δ_q^{OJT} drawn from the statistical population of passengers (Q) may be within the probability domain of the component distribution.

The way how Fu (2014) applied the mixture model in the context of route choice is to define the connection between the elements of the mixture model and the previously described Bayesian probabilities (cf. **Section 3.2.1**). More specifically, he pointed out the similarity between the component distribution of a route and the corresponding likelihood function:

$$\Pr\left(\Delta_q \middle| choice_{qr}\right) \approx c_r\left(\delta = \delta_q^{\mathrm{OJT}}\right) \tag{3-4}$$

as well as between the component proportion and the prior probabilities of route choice:

$$\Pr\left(choice_{qr}\right) \approx \omega_r \,. \tag{3-5}$$

These formulae does not express equality, but similarity, explaining, that in a context where both probabilities of $\Pr\left(\Delta_q \middle| choice_{qr}\right)$ and $\Pr\left(choice_{qr}\right)$ are unknown, the outputs of the mixture model can be a potential tool to estimate them.

Bringing this analogy forward and substituting formulae (3-4) and (3-5), into equation (3-1), the route choice probability conditional on journey time $\Pr\left(choice_{qr} \middle| \Delta_q\right)$ can be expressed as:

$$\Pr\left(choice_{qr} \middle| \Delta_q\right) \approx \frac{\omega_r c_r\left(\delta = \delta_q^{\mathrm{OJT}}\right)}{m\left(\delta = \delta_q^{\mathrm{OJT}};\, \boldsymbol{\omega}\right)}\,, \tag{3-6}$$

$\Pr\left(choice_{qr} \middle| \Delta_q\right)$ for all r routes could be derived, if the mixture components $c_r(\delta)$ and the corresponding weights ω_r were known. This could be obtained by finding the solution for the parameterised equivalent of equation (3-2):

$$m(\delta; \boldsymbol{\omega}, \boldsymbol{\vartheta}) = \sum\nolimits_{r \in R} \omega_r c_r(\delta; \boldsymbol{\theta}_r)\,, \tag{3-7}$$

In this setting, each mixture component $c_r(\delta)$ can be characterised by the parameters of the statistical distribution of the corresponding route, which is denoted by $\boldsymbol{\theta}_r$. The collection of parameters $\boldsymbol{\theta}_r$ for all routes is denoted by $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{N_R})$. Additionally, $\boldsymbol{\omega}$ denotes the random vector of all component proportions: $\boldsymbol{\omega} = (\omega_1, \dots, \omega_{N_R})$.

There are different model specifications for the statistical distribution of the mixture components, such as the Gaussian or lognormal. It can be easily understood, that the mixture components – which correspond to the journey time distribution of the actual routes – usually are not symmetric, but skewed to the left; therefore in theory a lognormal distribution could be a better model specification. Fu (2014) in his work concluded that both model specifications could give a good match of the real-world routes; however the lognormal model showed better goodness-of-fit for mixtures with two components, while it showed other problems for mixtures with three or more components. In this thesis – in order to make model and the corresponding program code simpler – the Gaussian distribution was chosen. This decision is justified by the fact that also Lee and Sohn

(2015) worked with the Gaussian model specification. Furthermore, Wahaballa et al. (2017) suppose Gaussian distribution for the on-board, access and egress time of passengers. Working with Gaussian distribution, the random vector of parameters $(\boldsymbol{\theta}_r)$ include the mean (μ_r) and standard deviation (σ_r) of $c_r(\delta)$.

3.2.3 Deriving posterior probabilities from the dataset of journey times

The available, valid dataset of OJTs – denoted as $OJT^0 := \{\langle q, \delta_q^{\mathrm{OJT}} : q = 1, \dots, n^{OJT} \rangle\}$ – contains n number of records, where each record gives information on the journey time δ_q^{OJT} of individual passenger q. The route r chosen by individual q is not known from this dataset, it can be treated only as a categorical variable. It is denoted by $r^{[q]}$. The probability function of $r^{[q]}$ is denoted by $\pi(r^{[q]} = r)$. It corresponds to the previously discussed component proportions (ω_r) for which the similarity with the prior probability $(\Pr(choice_{qr}))$ was expressed in formula (3-5).

Based on this, for each q, the posterior route choice probability $\pi(r^{[q]} = r;\ \delta)$ – which is equivalent to $\Pr(choice_{qr} | \Delta_q)$ – can be derived given the available dataset (OJT^0) and the appropriate parameters $(\boldsymbol{\vartheta})$ and proportions $(\boldsymbol{\omega})$ of the mixture components.

$$\pi\left(r^{[q]} = r;\ \delta, \boldsymbol{\omega}, \boldsymbol{\vartheta}\right) = \frac{\omega_r c_r(\delta;\ \boldsymbol{\theta}_r)}{m(\delta; \boldsymbol{\omega}, \boldsymbol{\vartheta})} \tag{3-8}$$

The problem here consists in the fact that in equations (3-7) and (3-8) both the parameters $(\boldsymbol{\vartheta})$ and the corresponding proportions $(\boldsymbol{\omega})$ are unknown. Therefore this equation cannot be solved analytically, but a numerical method is required, which is discussed in **Section 3.2.4**.

3.2.4 Solving the finite mixture model with the Expectation-Maximisation algorithm

To solve the above described problem, the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) is used. It is a maximum-likelihood function, an iterative process, which searches a set of optimum values of the parameters $(\boldsymbol{\vartheta})$ and component proportions $(\boldsymbol{\omega})$ with respect to dataset OJT^0, and estimates them by maximising the log-likelihood of the data sample. In this specific case, the EM algorithm is implemented as it follows:

Let $\ell(\boldsymbol{\omega}, \boldsymbol{\vartheta}; OJT^0)$ denote the likelihood function of $(\boldsymbol{\omega}, \boldsymbol{\vartheta})$, given the data set (OJT^0). The corresponding log-likelihood function with respect to the mixture model can be specified as:

$$\log \ell \left(\boldsymbol{\omega}, \boldsymbol{\vartheta}; OJT^0\right) = \sum\nolimits_{q=1}^{n^{OJT}} \log\left(\sum\nolimits_{r=1}^{N_R} \omega_r c_r\left(\delta_q^{\mathrm{OJT}}; \boldsymbol{\theta}_r\right)\right), \qquad (3\text{-}9)$$

The EM algorithm can be formulated in the following steps (McLachlan and Peel, 2000):

(i) **Initialisation**: The initial values of the parameters ($\boldsymbol{\vartheta}$), and component proportions ($\boldsymbol{\omega}$) are denoted with $\boldsymbol{\vartheta}^{[0]}$ and $\boldsymbol{\omega}^{[0]}$ respectively. Similarly to Fu (2014), the K-means clustering algorithm (Forgy, 1965; MacQueen, 1967) is applied to produce $\boldsymbol{\vartheta}^{[0]}$ and $\boldsymbol{\omega}^{[0]}$ (see **Section 3.3.1**). These values are used as an input in the Expectation [E] step: $\boldsymbol{\omega}^{[\mathrm{E}]} = \boldsymbol{\omega}^{[0]}$ and $\boldsymbol{\vartheta}^{[\mathrm{E}]} = \boldsymbol{\vartheta}^{[0]}$

(ii) **Expectation** (E-step): Let $\widehat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}^{[\mathrm{E}]}$ and $\widehat{\boldsymbol{\omega}} = \boldsymbol{\omega}^{[\mathrm{E}]}$. Firstly, the posterior route choice probabilities $\pi(r^{[q]} = r;\ \delta)$ are calculated for each q individual passenger and r route with equation (3-8). Then, based on equation (3-9), the expectation of the log-likelihood will be:

$$E(\log \ell) = \sum\nolimits_{q=1}^{n^{OJT}} \sum\nolimits_{r=1}^{N_R} \pi\left(r^{[q]} = r;\ \delta_q^{\mathrm{OJT}}, \widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\vartheta}}\right)\left[\log \widehat{\omega}_r + \log c_r\left(\delta_q^{\mathrm{OJT}}; \widehat{\boldsymbol{\theta}}_r\right)\right] \qquad (3\text{-}10)$$

(iii) **Maximisation** (M-step): Look for optimum values of $\widehat{\boldsymbol{\vartheta}}$ and $\widehat{\boldsymbol{\omega}}$, which should increase the current expectation of the log-likelihood (Equation (3-10)). They are denoted as $\boldsymbol{\vartheta}^{[\mathrm{M}]}$ and $\boldsymbol{\omega}^{[\mathrm{M}]}$. To find these optimum values, the partial derivative of the log-likelihood function (3-8) needs to be taken with respect to each variable within $\boldsymbol{\vartheta}$ and $\boldsymbol{\omega}$, then those equations need to be set to zero. Depending on the type of distribution (e.g. Gaussian, lognormal) of $c_r(\delta)$, the partial derivative functions with respect to $\boldsymbol{\vartheta}$ can be calculated and solved and hence $\boldsymbol{\vartheta}^{[\mathrm{M}]}$ can be found. The optimum value for the component proportion ($\boldsymbol{\omega}^{[\mathrm{M}]}$) can be obtained with the following equation:

$$\omega_r^{[\mathrm{M}]} = \frac{\sum\nolimits_{q=1}^{n^{OJT}} \pi\left(r^{[q]} = r;\ \delta_q^{\mathrm{OJT}}, \widehat{\boldsymbol{\omega}}, \boldsymbol{\vartheta}^{[\mathrm{M}]}\right)}{n}, \qquad (3\text{-}11)$$

Once $\boldsymbol{\vartheta}^{[\mathrm{M}]}$ and $\boldsymbol{\omega}^{[\mathrm{M}]}$. are found they are used in the next iteration to update the estimates of the E-step: $\boldsymbol{\omega}^{[\mathrm{E}]} = \boldsymbol{\omega}^{[\mathrm{M}]}$ and $\boldsymbol{\vartheta}^{[\mathrm{E}]} = \boldsymbol{\vartheta}^{[\mathrm{M}]}$.

(iv) **Iteration and stopping**: Repeat the E and M step until the increase in the expectation of the log-likelihood ($E(\log \ell)$) is not greater than a tolerance threshold. The results will be the optimum values of $\widehat{\boldsymbol{\vartheta}}$ and $\widehat{\boldsymbol{\omega}}$. The iteration process may stop at either the global or a local maximum of the log-likelihood. This means, that the EM algorithm may find the global or a local optimum values of $\widehat{\boldsymbol{\vartheta}}$ and $\widehat{\boldsymbol{\omega}}$.

These values obtained with the EM algorithm are the mixture model estimates. In the particular case of normal distribution, $\widehat{\boldsymbol{\vartheta}}$ includes the mean (μ_r) and standard deviation (σ_r) of each route r; while $\widehat{\boldsymbol{\omega}}$ can be associated with the aggregate route choice probabilities.

3.3 Convergence of the estimation algorithm

As it was described earlier (cf. **Section 3.2.4**), the EM algorithm and hence the mixture model may converge either to a global optimum value of $\widehat{\boldsymbol{\vartheta}}$ and $\widehat{\boldsymbol{\omega}}$ or to a local optimum. Therefore, it is crucial to know, whether the solutions given by the model truly reflect the actual values in the metro network, or the algorithm converges to a different solution. This brings up the following questions:

Question 1: If there are more possible solutions, how to decide which one to choose?

Question 2: What influences the convergence of the mixture model?

Question 3: Is it possible to set them to ensure it converges to the solution, which reflects the actual values in the metro network?

The question of identifying the desired local optimum was discussed in McLachlan and Peel (2000) focusing this question on finding the global optimum among the possible solutions. Gan and Jiang (1999) investigated further this question and tested the consistency and asymptotical efficiency of the possible solutions stating, that the global optimum should hold this property. Finding the global optimum is a challenging question; however even if that is found, it would not necessarily mean that it also corresponds to the actual values of journey times and route choice probabilities captured from other data sources (i.e. timetables, travel surveys, see **Section 3.6**). Therefore, the model results need to be further evaluated. (**Question 1**).

As it is discussed in Fu (2014); it is the initial value and the tolerance threshold that gives the most significant influence on the results of the mixture model (**Question 2**). In this section trials are conducted with different settings of the mixture model for initial values (see **Section 3.3.1**) and tolerance thresholds (see **Section 3.3.2**) comparing these results with the actual LU values (**Question 3**).

3.3.1 Initialisation with the K-means clustering algorithm

For the $\boldsymbol{\vartheta}^{[0]}$ and $\boldsymbol{\omega}^{[0]}$ values in the EM algorithm (cf. **Section 3.2.4**) an initialisation technique should be used. These initial values, in fact, could be associated with the centroids of the data clusters that can be reproduced with an appropriate algorithm. Hong et al. (2017) presents a comprehensive literature review on different approaches, such as the K-means (Forgy, 1965; MacQueen, 1967), K-medoids (Kaufman and Rousseeuw, 2009), distribution based (McLachlan and Krishnan, 2007) or density based (Ester et al., 1996) clustering algorithms as well as a novel method, which performs the clustering by fast search and find of density peaks (Rodriguez and Laio, 2014). Although there are more advanced methods in literature, for the purpose of this thesis, the simpler K-means clustering algorithm is already adequate.

As Gaussian component distributions were assumed (cf. **Section 3.2.2**), the parameters for the mixture model ($\boldsymbol{\vartheta}^{[0]}$) include the mean ($\boldsymbol{\mu}^{[0]}$) and standard deviation ($\boldsymbol{\sigma}^{[0]}$) of the mixture components.

Running the K-means clustering algorithm, each value in the OJT^0 dataset is assigned to one of the N_R clusters according to the number of routes. This subset of OJTs is denoted as OJT_r^{KMS}. The initial value for the mean ($\boldsymbol{\mu}^{[0]}$) and standard deviation ($\boldsymbol{\sigma}^{[0]}$) for the mixture model is given as the mean and standard deviation of the OJT_r^{KMS} clusters (denoted as μ_r^{KMS} and σ_r^{KMS} respectively). The initial value of the component proportion ($\boldsymbol{\omega}^{[0]}$) produced by the K-means clustering algorithm corresponds to:

$$\omega_r^{KMS} = \frac{OJT_r^{KMS}}{OJT^0} \tag{3-12}$$

To get a better confidence that the initialisation gives acceptable values, the input parameters of the K-means clustering algorithm were set in the following way:

- The cluster centroids were initialised by K-means ++ algorithm (Arthur and Vassilvitskii, 2007). This more advanced technique was applied, because working with randomly chosen cluster centroids, the mixture model gave results, which were far from the expected values for the LU.
- The point-to-cluster-centroid distances were calculated according to the Euclidean square distance metrics.

- The online update phase was used in addition to the batch update phase to guarantee a solution that is a local optimum (Chamundeswari et al., 2012).

The settings of these parameters could improve the results obtained with the K-means clustering algorithm; but it still could not ensure that it converges to the same value in every iteration. This is due to the fact, that even though a more advanced method is used for initialising the cluster centroids with K-means ++ algorithm, the very first input for the centroid initialisation was generated randomly. In order to enumerate the possible results with the K-means clustering algorithm trials were conducted with different settings of the random number generator of the computer – they are called seeds – to

obtain the estimates of the K-means clustering algorithm. As it follows (**Section 3.3.2**), the mixture model is tested with these initial values for different tolerance thresholds.

3.3.2 Setting a tolerance threshold

Using different set of initial values produced by the K-means clustering algorithm (cf. **Section 3.3.1**). The mixture model was run with a range of tolerance thresholds (step (iv) in the EM algorithm, cf. **Section 3.2.4**) to see, which would give the best fit of the model. It is noteworthy that the mixture model starts finding solutions near the initial value; therefore if the tolerance threshold is set larger, it is likely to find a local optimum in the proximity of the initial values, while if it is set smaller, it is more likely that it converges to the global optimum. The task here is to find the tolerance threshold that is small enough to obtain the desired solution, but not too small, so that it increases unnecessarily the computational time of the algorithm. In the case studies trials were conducted with tolerance threshold values ranging (exponentially) between 1e-01 and 1e-10.

At this point the local optimum in the proximity of the results captured from other sources (i.e. timetables, travel surveys) was accepted as a solution of the mixture distribution problem instead of the global optimum. This is because it is expected that among all the possible solutions there exists at least one, which reflects the actual values of the metro network, although this may not be necessarily the global optimum.

3.4 Validation of the model results

Applying the finite mixture model on the OJT^0 distribution (cf. **Section 3.2**) with the appropriate settings for initial value and tolerance threshold (cf. **Section 3.3**) the results were obtained as the journey time distribution (with parameters μ_r^{MIX} and σ_r^{MIX}) and proportion (ω_r^{MIX}) of the mixture components. At this point it is still unknown, which mixture component (labelled with r) corresponds to which actual route (labelled with k). Therefore in **Section 3.4.1**, it is discussed how to match them, so that the journey time and route choice values of actual routes can serve for the validation of model results.

In this thesis the finite mixture model was applied on a much smaller data sample than in Fu (2014), with slightly different settings for the initial values. Therefore the results of that model can serve as another source for validation. This is presented in **Section 3.4.2**.

3.4.1 Matching the mixture results with actual LU routes

Fu (2014) proposed to match the results of the finite mixture model with the actual LU routes based on the following criteria:

- Match the mean journey time of mixture components with the reference time of the actual LU routes
- Calculate confidence intervals for the actual LU routes and check whether the mean of the mixture components falls into that confidence interval
- Check the proportion of the mixture components with RODS data on route choice

He modelled the confidence intervals for the actual routes based only on the difference in fail-to-board delays, and he assumed that for each route and each journey leg of that route half of the passengers can board the first, half of them the second train.. Looking at the characteristics of metro networks, it was understood that this assumption for the fail-to-board delays is not always realistic as it varies in time (e.g. within the AM peak) and within the metro network (e.g. city centre vs outskirts) (see **Section 7.6.2**). Therefore in this thesis a different approach is used.

As the first attempt, a simpler consideration is made, without yet considering confidence intervals and fail-to-board delays: The mixture components (r) are associated with the actual LU routes (k) only based on their journey time, matching the mean journey time for mixture components (μ_r^{MIX}) with the Scheduled Journey Time (SJT) of the actual LU routes, which is calculated as:

$$t_k^{SJT} = t_{1,k}^{acc} + \sum_{l=1}^{N_{L,k}} (t_{l,k}^{wait} + t_{l,k}^{ob}) + \sum_{s=1}^{N_{S,k}} t_{s,k}^{ic} + t_{L,k}^{egr} \qquad (3\text{-}13)$$

Based on (3-13), a route is defined as the sequence of the following journey segments between the entry and exit ticket gate: access to line 1 – wait for line 1 – on-board line 1 – interchange 1 – … – interchange $N_{S,k}$ – wait for line $N_{L,k}$ – on-board line $N_{L,k}$ – egress from line $N_{L,k}$, where $N_{L,k}$ and $N_{S,k}$ is the total number of lines and interchange stations respectively.

Once the journey time of the mixture results are matched with the actual LU routes, also the corresponding component proportions (ω_r^{MIX}) can be validated with the aggregate route choice results from RODS data (**Section 3.6.3**) (ω_k^{RODS}).

In reality, for the appropriate matching between the mixture results and the actual routes a more detailed consideration of confidence intervals would be necessary. Firstly – as it as mentioned above – fail-to-board delays vary within the AM peak and the metro network. Secondly, the confidence interval itself depends not only on the variance in fail-to-board delays, but also on the variance in the other time components (i.e. on-board, wait for the first coming service, access egress interchange). In this thesis, the problem of fail-to-board delays is introduced in **Chapter 6** and **Chapter 7**, and a different matching process accounting for these issues is discussed in **Section 7.3**. Another possible approach would be to see the confidence interval of the mixture results instead of the one of the actual routes. This approach is not discussed in this thesis, but could be subject of further research.

3.4.2 Comparison with Fu (2014)

The solution algorithm for the finite mixture distribution problem in Fu (2014) and in the proposed model is essentially the same (EM algorithm, cf. **Section 3.2.4**). Similarly also the algorithm used for initialisation is the same in the two models (K-means clustering algorithm, cf. **Section 3.3.1**), however the input parameters are slightly different (see

Table 3-1).

The main difference between the two models is that for this thesis only a smaller data sample was available The reason for this is that Transport for London (TfL) gave access only to the open data, which was available from their website (see **Section 3.6.1**): a 5% sample of the Oyster cardholders from a 1 week period (50-100 records per OD pair); while for Fu (2014) a larger sample of aggregate data could be provided as bespoke data: a 100% sample from a 40 week period (20000-30000 records per OD pair). Having smart card data collected over a longer time period has the obvious advantage of larger data sample and hence more reliable model estimates. However it is important to note that with the temporal aggregation, the modeller does not consider that in different time periods passengers may have different route choice patterns (e.g. term time vs summer holiday). Therefore working with data from shorter time periods has the value to give time period specific route choice estimates.

Applying the proposed model on the same case study OD pairs as Fu (2014) can be another source for validation. Through this, it is possible to understand, at what extent finite mixture models can estimate route choice when only a small data sample is available.

When comparing the results of the model implemented in this thesis and the results of Fu (2014), it is also important to note, that while his data was collected in 2011-2012; the data used in this thesis is from 2009. Therefore also the changes in people's route choice behaviour or in service provision (e.g. timetable changes, etc.) need to be considered.

Table 3-1 Comparison of Fu (2014) and the implemented model

Model input	**Fu (2014)**	**Implemented model**
Solution algorithm	Expectation-Maximisation algorithm	
Initialisation algorithm	K-means clustering algorithm	
Initialisation of cluster centroids	Random selection	K-means ++ algorithm
Point-to-cluster-centroid distance	Sum of absolute differences	Squared Euclidean distance
Update phase	Online	
Tolerance threshold	Trials with a range of values	
Dataset	100% sample, 40 weeks	5% sample, 1 week

3.5 Software implementation

The code for the implementation of the finite mixture model on the case study OD pairs was written in Matlab. The functions, input parameters and outputs for the initialisation

(K-means clustering, **Section 3.5.1**) and solution (Expectation-Maximisation, **Section 3.5.2**) algorithm are presented as it follows.

3.5.1 Initialisation: K-means clustering

As discussed in **Section 3.2.4** the initial values for the EM algorithm are produced by the K-means clustering algorithm. In Matlab 'kmeans'[4] is a built-in function that can be used to assign the given dataset to one of the K number of clusters. In this specific case, the input parameters for the function are the following:

- X: The dataset of journey time observations after removing outliers OJT^0.
- k: The number of known routes for the OD pair.
- **'Start'**: the option **'plus'**, is chosen, which means that the cluster centroids are initialised by the K-means ++ algorithm.
- **'Distance'**: the option **'sqeuclidean'** is chosen, which means that the point-to-cluster-centroid distances were calculated according to the Euclidean square distance metrics.
- **'OnlinePhase'**: the option **'on'** is chosen, which means that the online update phase is used in addition to the batch update phase.

The 'kmeans' function produces the output of a vector with the cluster labels, showing to which cluster the data entry is assigned. The OJTs assigned to each cluster correspond to OJT_r^{KMS}. From that μ_r^{KMS}, σ_r^{KMS} and ω_r^{KMS} is calculated as described in **Section 3.3.1**.

In order to conduct trials of the K-means clustering algorithm with different seeds, the random number generator of the computer was set to constant values with the 'rng' Matlab function.

3.5.2 Solution: Expectation-Maximisation

As discussed in **Section 3.2.4**, the finite mixture model is solved with the EM algorithm. In Matlab 'fitgmdist'[5] is a built-in function that is used to fit a Gaussian mixture distribution on a given dataset. In this specific case, the input parameters for the function are the following:

- X: The dataset of journey time observations after removing outliers OJT^0
- k: The number of known routes for the OD pair

[4] https://uk.mathworks.com/help/stats/kmeans.html

[5] https://uk.mathworks.com/help/stats/fitgmdist.html

- **'Start'**: Initial values ($\boldsymbol{\vartheta}^{[0]}$ and $\boldsymbol{\omega}^{[0]}$ step i), produced by the K-means algorithm (cf. **Section 3.5.1**)
- **'MaxIter'**:Maximum number of iterations, set to 10000
- **'TolFun'**:Threshold value (step iv), with different settings (cf. **Section 3.3.2**)

The fitgmdist function gives the following outputs, relevant to this thesis:

- Converged: Logical (0 or 1) to state, whether the EM algorithm could converge to the local optimum given the set tolerance threshold and maximum iteration values
- NumIterations: Number of iterations necessary for convergence
- mu: Mean of the Gaussian mixture components
- Sigma: Variance of the Gaussian mixture components
- NegativeLogLikelihood: Negative log-likelihood, showing how good match the estimates give

3.6 Data sources for the case studies

Principally, the main input for the finite mixture model is the dataset of OJTs understood from smart card data, which is presented in **Section 3.6.1**. Additionally, the data sources that are used for matching the mixture results with the corresponding values of the actual LU routes (i.e. Scheduled Journey Time of routes) are described in **Section 3.6.2**. Finally, the historical data for route choice is presented in **Section 3.6.3**.

3.6.1 Oyster data

The smart card for the LU network is called Oyster card, and the journey detail records extracted from that card are called Oyster data. The raw data is collected and processed by TfL and provided for researchers in different output forms, depending on the research objectives.

Disaggregate Oyster data contains detailed records on each smart card transaction, including the encrypted passenger ID, public transport subsystem (i.e. bus, LU, rail) entry/exit time and station as well as information on the ticket type and fare. Their advantage is that travel patterns of individual passengers, such as the day-to-day variation of entry/exit choice can be observed (see **Section 2.3** and **Section 5.2.2**). In case of London, around 200 million Oyster transactions are recorded in a 4 week period for the whole TfL network[6]. The storage and processing of such an amount of data is computationally expensive. Therefore, in some cases, – depending on the research objective – researchers are only provided with a smaller data sample (e.g. 5% sample of Oyster cardholders) for shorter time periods (e.g. 4 weeks).

In order to be able to analyse longer time periods with larger data sample at less computational cost researchers are provided with Oyster data in the form of aggregate data. There Oyster transactions are aggregated according to certain features (e.g. Observed Journey Times, OJT), and are filtered according to entry/exit station and time. The limitation of this dataset is, that due to data aggregation it is not possible to identify individual travel patterns (e.g. which journey time observations belong to the same passenger).

TfL released a 5% sample of disaggregate Oyster data from a 1 week period in November 2009 as open data[7]. In this dataset – to further comply with the privacy policy of the passengers – also the encrypted passenger ID column was removed, therefore individual travel patterns are unidentifiable. Data with the information on encrypted passenger IDs, with larger sample size or from longer time period needs to be requested as bespoke data. For this thesis, only the open data was available.

In the period of data collection (2009), Oyster card was used for 73% of the LU journeys[8]. Since April 2014, also contactless payment cards (i.e. bank cards) are accepted for fare payment the same way as the Oyster card. The total proportion of Oyster and contactless payments card usage is similar at the time of this study being conducted (2018)[9].

Oyster card is accepted on all public transport modes within Greater London. For rail modes (LU, LO, DLR, TfL rail, NR) passengers need to tap card both at the entry and at the exit station. This way, the Oyster card record includes:

- Day of the week (e.g. Monday)
- Subsystem (e.g. LU, NR)
- Entry/exit station and time

Additionally it gives information on the ticket type used by the passenger (pay-as-you-go or season ticket) and on the fare calculated.

[6] https://www.whatdotheyknow.com/request/oyster_card_usage
[7] https://tfl.gov.uk/info-for/open-data-users/
[8] https://www.whatdotheyknow.com/request/oyster_card_usage
[9] https://tfl.gov.uk/corporate/publications-and-reports/oyster-card
https://tfl.gov.uk/corporate/publications-and-reports/contactless-payment

In the model, the period of main interest is the weekdays AM peak (set for the time between 7:00 and 10:00 by TfL). Therefore as the first step, the Oyster dataset was filtered accordingly. This dataset contains for each observation (q) the time stamps at the entry (T_q^{entry}) and exit ticket gate (T_q^{exit}). However, the finite mixture model requires solely the OJTs of passengers as an input, which can be calculated as:

$$\delta_q^{\mathrm{OJT}} = T_q^{exit} - T_q^{entry} \tag{3-14}$$

It is important to note that as the Oyster data reveals T_q^{entry} and T_q^{exit} with the precision of 1 minute, also δ_q^{OJT} will be treated with that precision.

3.6.2 Scheduled journey time of routes

In addition to the OJT of passengers it is also necessary to gain further understanding on the journey time of the actual LU routes between the entry and exit ticket gates (t_k^{SJT}), which can be calculated using equation (3-13). Among the journey time components specified there, LU timetables are used for t_l^{ob} and $t_{l,k}^{wait}$ (see **Section 3.6.2.1**) and the access egress interchange (AEI) times ($t_{1,k}^{acc}$, $t_{L,k}^{egr}$ and $t_{s,k}^{ic}$) are estimated based on station layouts from The Nationwide Access Register (see **Section 3.6.2.2**).

3.6.2.1 On-board and wait times

The current (2018) LU timetables are available online from the TfL website[10]. From this, the on-board time for each journey leg ($t_{l,k}^{ob}$) can be obtained straightforward. Assuming high-frequency services (more than 4 trains/hr), the wait time for the first coming service can be assumed to be half of the frequency:

$$t_{l,k}^{wait} = \frac{1}{2 \cdot f_{l,k}} \cdot 60 \tag{3-15}$$

where $f_{l,k}$ is the frequency (trains/hour) on the given journey leg l, k, which can be captured from timetables as the number of trains in a given time period (e.g. hour). Using equation (3-15), $t_{l,k}^{wait}$ is obtained in minutes.

It is important to acknowledge that working with timetable data would suggest that all trains run on time, which would not stand especially under extreme crowding conditions in the peak of the AM peak (8:00-9:00). There delayed boarding time and not constant

[10] https://tfl.gov.uk/travel-information/timetables/

headways could be expected due to train bunching. For the objectives of this thesis, it is considered to be sufficient to work with the journey time values understood from timetables. Further research can address capturing data from the live departure board feeds of the LU, available also from the TfL open data website[11]. From that the mean and variance of $t_{l,k}^{ob}$ and $t_{l,k}^{wait}$ could be estimated, providing a better approximation for the actual journey time of routes.

3.6.2.2 Access Egress Interchange times

The Nationwide Access Register (a.k.a. Direct Enquires)[12] provides information on the layout of all LU stations in terms of the passageways from/to/between the platforms of the LU lines. For each passageway, the sequence of (ascending or descending) stairs, escalators, lifts and level walks are given with their corresponding length or number of steps. Knowing this, $t_{1,k}^{acc}$, $t_{L,k}^{egr}$ and $t_{s,k}^{ic}$ were estimated, supposing an average of 1.33 m/s walk speed in even passageways (Transport for London, 2010) and 2.77 steps/s for descending and 2.36 steps/s for ascending (Fujiyama and Tyler, 2010).

3.6.3 Understanding route choice set and validating route choice results

The set of chosen routes and the surveyed route choice proportions can be understood explicitly from RODS data (collected between 1998 and 2017). This data is used for the validation of the model results.

The RODS has been carried out by TfL since 1998 collecting data at 30-40 LU stations each year, where passengers fill out a questionnaire on their current journey. From this, route choice proportions can be calculated. At the same time also (manual and automatic) passenger count data is collected at stations, so that the RODS route choice results can be reconciled to the control totals, producing this way information on the on-board and AEI flows for each 15 minute period.

The main issue with RODS data is that although it has a large sample of journey records (i.e. 4.9 million questionnaires); all these records come from different years, and reflect only the month of the data collection (i.e. November), therefore they are unable to provide time period specific information. Furthermore, RODS data contains only trips on weekdays during normal operation of the LU, excluding engineering works and disruption (Chan, 2007).

[11] https://tfl.gov.uk/info-for/open-data-users/
[12] http://www.directenquiries.com/londonunderground.aspx

3.7 Case studies on the London Underground

The previously described finite mixture model is applied on the case study OD pairs of the LU. These are OD pairs with multiple reasonable routes. According to RODS data 83.7% of the OD pairs in the LU have only one observed route in the AM peak (Guo, 2008). Even though this percentage seems quite high, for the adequate passenger flow modelling, it is still important to know how passengers make choices on the remaining 16.3% of the OD pairs. Especially, because within Central London there are many lines and hence route options. Among the OD pairs with multiple routes those two of them were chosen for the case study (**Figure 3-2**):

- **Case 1**: **Victoria – Holborn**
- **Case 2**: **Liverpool Street – Green Park**

These OD pairs have relatively greater demand. Both of these cases represent route choice within the LU inner zone, where travel times are relatively short and interchange stations are quite complex. The main difference between these two cases is that, according to RODS data, while for **Case 1** there are two reasonable routes, for **Case 2** there are three. These case study OD pairs were also analysed in Fu (2014), where he compared the performance of the finite mixture model for two and three component mixture distributions. For the same purpose also in this chapter both of these cases are reported. Additionally, for the research problems discussed later in this thesis (**Chapter 5**, **Chapter 6** and **Chapter 7**) there is also a need to present more cases. This will be explained in the corresponding case studies.

Table 3-2 presents the case study OD pairs with the observed routes (according to RODS, cf. **Section 3.6.3**), as well as the available data for them:

- Travel time of routes (t_k^{SJT}), cf. **Section 3.6.2**)
- RODS route choice proportions (ω_k^{RODS}) and sample size (n^{RODS}, cf. **Section 3.6.3**)
- OJT sample size of OD pairs (n^{OJT}, cf. **Section 3.6.1**)

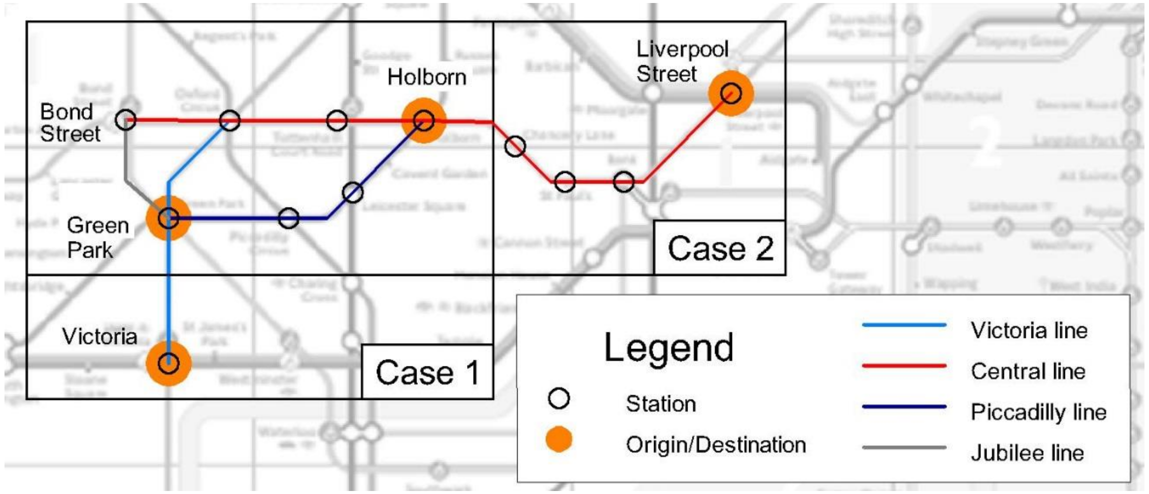


Figure 3-2 Overview of the case study OD pairs

Table 3-2 Resume of the case study OD pairs and their properties

Case	**OD pair**		**Route**			**Time**	**RODS**		**OJT**
	Origin	**Destination**	**Line 1**	**Interchange 1**	**Line 2**	(minutes)	**RC** (%)	**Sample**	**Sample**
			$l = 1$	$s = 1$	$l = 2$	t_k^{SJT}	ω_k^{RODS}	n^{RODS}	n^{OJT}
1	**Victoria**	**Holborn**	**Victoria (NB)**	**Oxford Circus**	**Central (EB)**	17.6	80.4%	561	54
			Victoria (NB)	**Green Park**	**Piccadilly (EB)**	20.4	19.6%		
2	**Liverpool Street**	**Green Park**	**Central (WB)**	**Oxford Circus**	**Victoria (SB)**	21.3	75.9%	917	30
			Central (WB)	**Holborn**	**Piccadilly (WB)**	24.0	12.2%		
			Central (WB)	**Bond Street**	**Jubilee (EB)**	23.2	11.9%		

Case 1 Victoria – Holborn

Looking at the map and RODS data, two reasonable routes can be identified for the **Victoria** – **Holborn** OD pair (see **Figure 3-3**):

1) **Victoria** – **Central** (via **Oxford Circus**)
2) **Victoria** – **Piccadilly** (via **Green Park**)

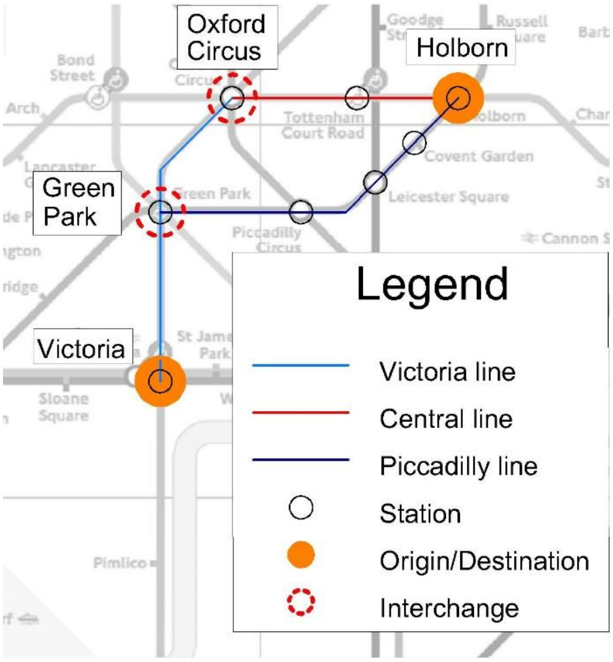


Figure 3-3 The **Victoria** – **Holborn** OD pair

From Oyster data (cf. **Section 3.6.1**) the OJT dataset is given for this OD pair, containing 54 transactions in the observation period (1 week in November 2009). Within this dataset all entries could be considered as valid data, because the upper outer fence (cf. **Section 3.2.1**) resulted 40 minutes, while the maximum OJT value is 31 minutes. This valid dataset is denoted by OJT^0 (**Figure 3-4**)

Having identified two reasonable routes, route choice is estimated as a two-component ($N_R = 2$) finite mixture distribution. Therefore, the K-means clustering algorithm was applied on the OJT^0 dataset with two clusters to produce the initial values for the EM algorithm. The previously described (cf. **Section 3.3.1**) settings were used for centroid initialisation (K-means ++), distances (Euclidean square) and update methods (online phase). Conducting trials with various seed values for the random number generator (1, 2, 3, etc.) the K-means clustering algorithm gave two possible solutions for μ_r^{KMS}, σ_r^{KMS} and ω_r^{KMS} (**Table 3-3**)

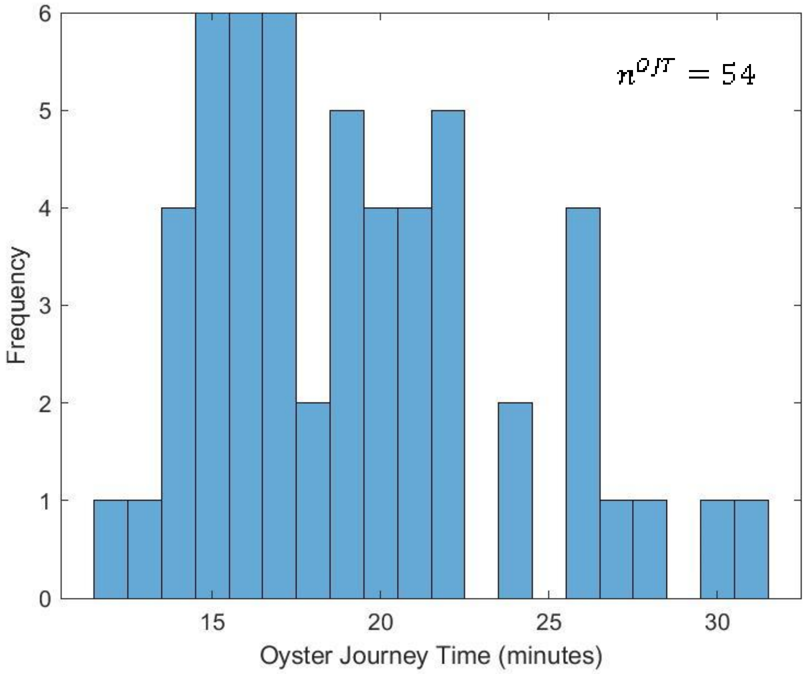


Figure 3-4 Distribution of Observed Journey Times for **Victoria** – **Holborn**

Table 3-3 Results of the K-means clustering algorithm with different seeds for **Victoria** – **Holborn**; -a) Seed=1, b) Seed=2

Label	K-means clustering		
r []	μ_r^{KMS} [min]	σ_r^{KMS} [min]	ω_r^{KMS} [%]
1	17.0	2.8	81.5%
2	26.0	2.3	18.5%

a)

Label	K-means clustering		
r []	μ_r^{KMS} [min]	σ_r^{KMS} [min]	ω_r^{KMS} [%]
1	16.0	1.9	57.4%
2	22.0	3.3	42.6%

b)

Using these initial values the EM algorithm was run with different settings for the tolerance threshold (cf. **Section 3.3.2**). **Figure 3-5** and **Figure 3-6** presents the estimated mean (μ_1^{MIX}) and proportion (ω_1^{MIX}) for mixture component labelled with $r = 1$. There, it is shown that when the tolerance threshold is 1e-05 or greater, the EM algorithm converges to a solution close to the initial value for seed 1. But when the tolerance threshold is 1e-06 or smaller, the EM algorithm converges to a solution around 15.5 minutes for the mean and 32.9% for the component proportion for both seeds.

Similar properties could be observed for the other mixture component (labelled with $r =$ 2).

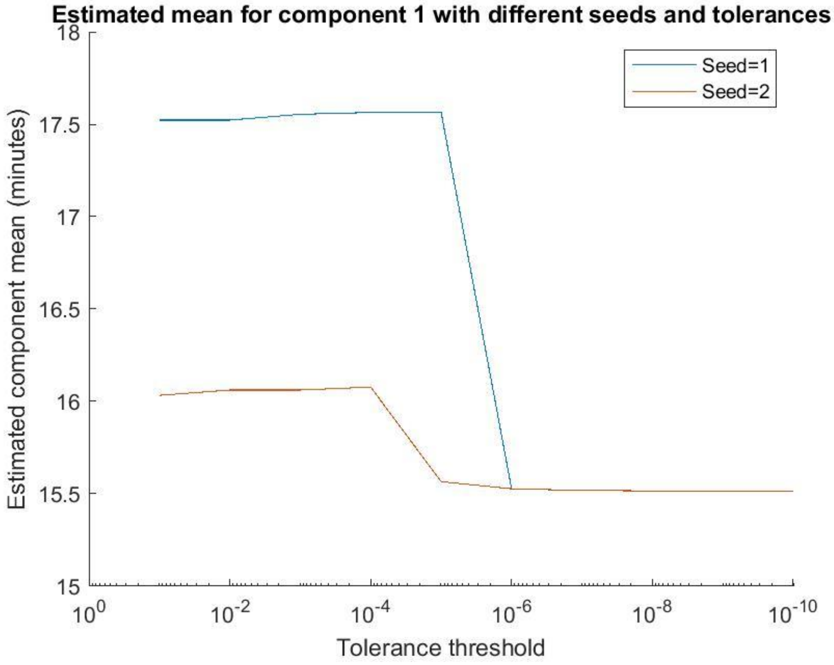


Figure 3-5 Estimated mean for mixture component 1, given different initial values and tolerance thresholds for **Victoria** – **Holborn**

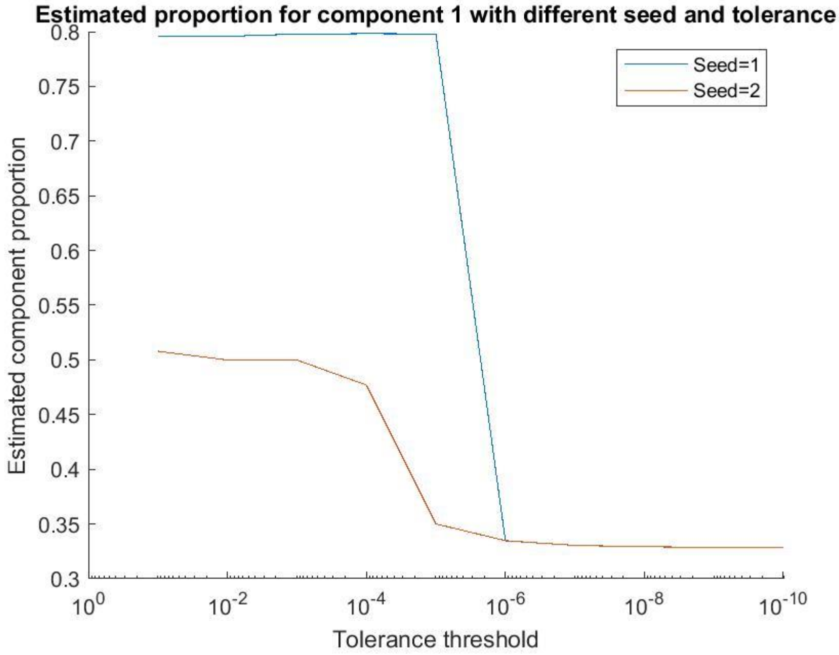


Figure 3-6 Estimated proportion for mixture component 1, given different initial values and tolerance thresholds for **Victoria** – **Holborn**

Figure 3-7 presents the log-likelihood (equation (3-9)) for each initial value (seed) and tolerance threshold. It shows a considerable jump in the log-likelihood between the tolerance threshold of 1e-05 and 1e-06, below which the EM converges to the mean of 15.5 minutes and proportion of 32.9% for component 1.

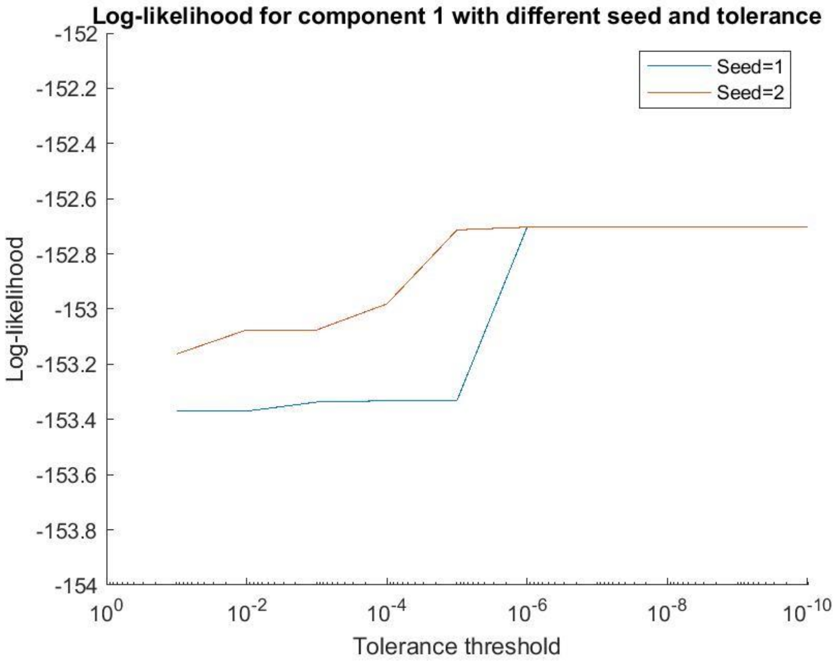


Figure 3-7 Log-likelihood, given different initial values and tolerance thresholds for **Victoria – Holborn**

According to RODS data, the route choice proportions for the two routes of the **Victoria – Holborn** OD pair are 80.4% and 19.6%. Among the estimates, the one with seed 1 and tolerance threshold 1e-05 gives the best approximation, therefore these settings were applied for the finite mixture model (**Table 3-4**).

Table 3-4 Finite mixture model results; with Seed: 1, Tolerance threshold: 1e-05 for **Victoria – Holborn**

Label	Mixture model		
r	μ_r^{MIX}	σ_r^{MIX}	ω_r^{MIX}
[]	[min]	[min]	[%]
1	17.6	2.9	79.8%
2	26.1	2.8	20.2%

Table 3-5 Journey time of actual London Underground routes for **Victoria – Holborn**

Route				Journey Time [min]							
k	$l=1$	$s=1$	$l=2$	$t_{1,k}^{acc}$	$t_{1,k}^{wait}$	$t_{1,k}^{ob}$	$t_{1,k}^{ic}$	$t_{2,k}^{wait}$	$t_{2,k}^{ob}$	$t_{2,k}^{egr}$	t_{k}^{SJT}
1	Victoria	Oxford Circus	Central	2.4	0.9	4.0	3.1	1.2	3.0	3.1	**17.6**
2	Victoria	Green Park	Piccadilly	2.4	0.9	2.0	4.1	1.3	6.0	3.8	**20.4**

Table 3-6 Matching mixture model results with the actual London Underground routes for **Victoria – Holborn**

Red: Mixture results of proposed model, **Yellow**: Fu (2014), **Green**: actual LU routes

Mixture label	Journey Time (min)			Route Choice (%)			Route label	Route Matched		
	Mixture		Timetable	Mixture		Timetable		Line 1	Interchange 1	Line 2
r	μ_r^{MIX}		t_k^{SJT}	ω_r^{MIX}		ω_k^{RODS}	k	$l=1$	$s=1$	$l=1$
	Proposed	Fu(2014)		Proposed	Fu(2014)					
1	17.6	16.6	17.6	79.8%	75.4%	80.4%	1	Victoria	Oxford Circus	Central
2	26.1	22.2	20.4	20.2%	24.6%	19.6%	2	Victoria	Green Park	Piccadilly

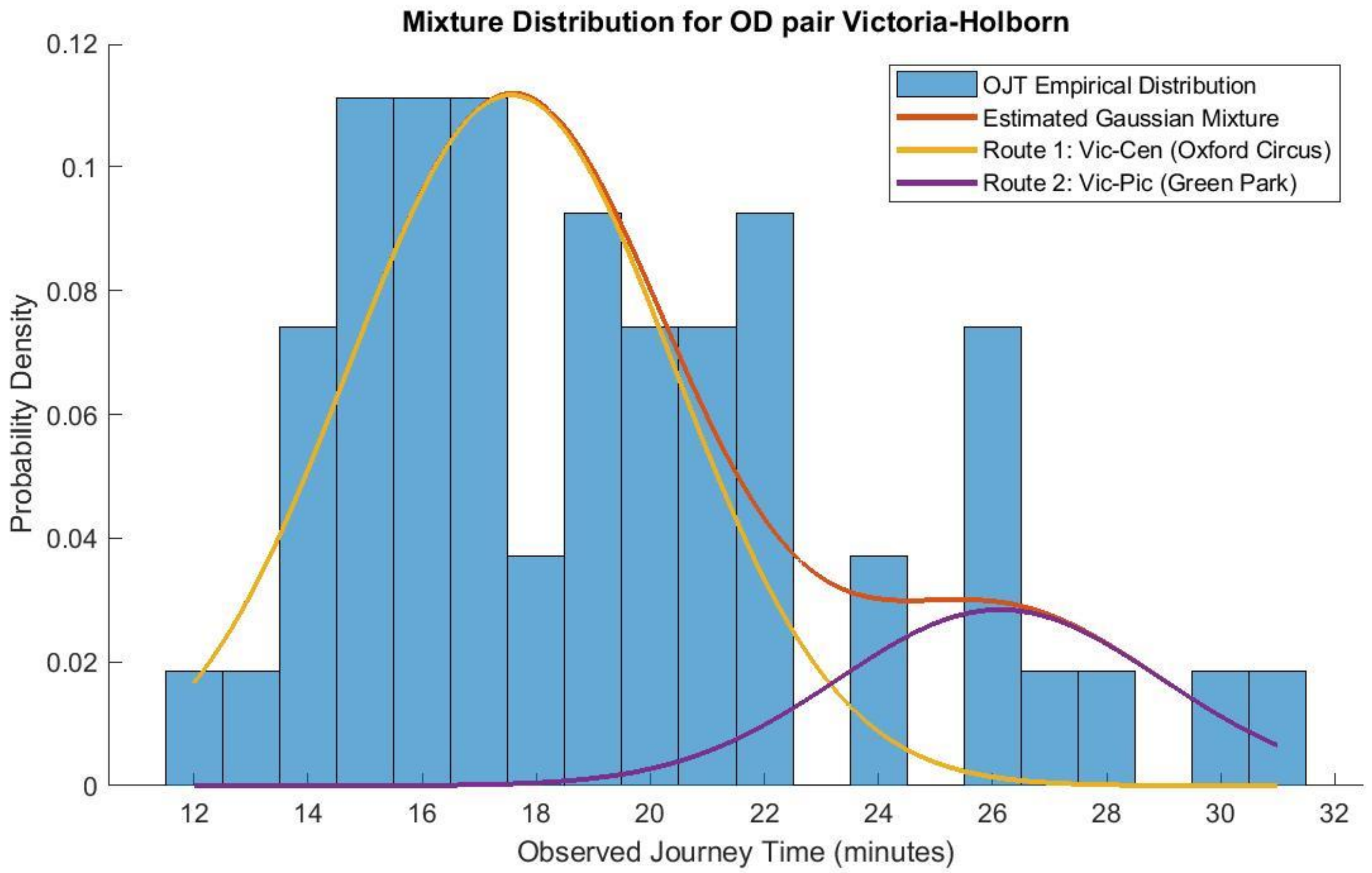


Figure 3-8 Estimated (Gaussian) journey time distribution of the routes for **Victoria** – **Holborn**

Following this, the results of the finite mixture model were matched with the actual LU routes (cf. **Section 3.4.1**). The total journey times of the actual LU routes (t_k^{SJT}) were calculated between **Victoria** and **Holborn** stations based on equation (3-13). The results are presented in **Table 3-5**. It is expected that the mixture component with the lower mean ($r = 1$) may correspond to the route with the shorter journey time ($k = 1$). Similarly the component with the higher mean ($r = 2$) to the route with the longer journey time ($k = 2$). The results for the finite mixture model and the values for the actual LU routes are summarised in **Table 3-6** together with the results of Fu (2014) for the same OD pair. **Figure 3-8** presents the probability density functions of the mixture distribution fit on the CCOJT dataset as well as of the mixture components matched with the actual LU routes.

Table 3-7 Mean journey times with different number of mixture components for **Victoria** – **Holborn**

# of mixture components	Mean journey time of mixture component			
N_R	μ_1^{MIX}	μ_2^{MIX}	μ_3^{MIX}	μ_4^{MIX}
1	19.3			
2	17.6	26.1		
3	15.6	20.4	26.5	
4	14.8	16.3	20.4	26.5

Table 3-8 Proportions with different number of mixture components for **Victoria** – **Holborn**

# of mixture components	Proportion of mixture component			
N_R	ω_1^{MIX}	ω_2^{MIX}	ω_3^{MIX}	ω_4^{MIX}
1	100.0%			
2	79.8%	20.2%		
3	48.4%	32.3%	19.3%	
4	21.0%	27.1%	32.4%	19.4%

Based on the results with the finite mixture model, the following issues were raised: Firstly, it is important to note that these results were obtained by setting the number of mixture components (N_R) to 2, as from the LU map and RODS data it was understood, that the **Victoria** – **Holborn** OD pair has two reasonable routes. To illustrate the importance of the correct specification of the route choice set, the same finite mixture model was run with different settings for the number of mixture components (i.e. 1, 2, 3, and 4). The corresponding component means (μ_r^{MIX}) and proportions (ω_r^{MIX}) are shown in **Table 3-7** and **Table 3-8** respectively.

Looking at the mean and proportion of the mixture component corresponding to the longest route ($\mu_{N_R}^{MIX}$ and $\omega_{N_R}^{MIX}$ respectively), it is observable that in all cases – whether N_R is chosen to be 2, 3 or 4 – they are around the same value (26 minutes and 20%). This is definitely not a good representation of the reality, because in actual metro networks it is expected that the next (i.e. 3rd and 4th) shortest route have a much longer journey time and a very small route choice proportion (less than 5%).

Another crucial point in the application of finite mixture models is that it gave essentially different results depending on the seeding of the random number generator and on the tolerance threshold, as the EM algorithm was converging to different local optima. Most notable among these results is the component proportion, which exhibits a significant jump from 79.8% to 33.5% between tolerance thresholds 1e-05 and 1e-06 for component 1 (cf. **Figure 3-6**) As the component proportion corresponds to route choice (formula (3-5)) it is crucial that the modeller could choose the proper seed value and tolerance thresholds. The reason for this big difference across the estimates could be explained with the fact that the OJT^0 dataset, on which the finite mixture model was applied had very small sample size ($n = 54$).

Finally, comparing the results of the proposed finite mixture model with the actual LU routes (cf. **Table 3-6**), one can see that the mean journey time of component 1 (μ_1^{MIX}) shows quite a good match to the actual LU route (t_1); however for component 2 the difference is quite notable (26.1 and 20.4 minutes respectively). The higher OJT values in the data sample, in fact, could not necessarily mean that the passenger has taken a longer route, but it could be also because he/she has experienced fail-to-board delays in any of the routes, as that the northbound platform of the Victoria line at Victoria station is extremely crowded in the AM peak.

Case 2 Liverpool Street – Green Park

Looking at the map and RODS data, three reasonable routes can be identified for the **Liverpool Street** – **Green Park** OD pair (see **Figure 3-9**):

1) **Central** – **Victoria** (via **Oxford Circus**)
2) **Central** – **Piccadilly** (via **Holborn**)
3) **Central** – **Jubilee** (via **Bond Street**)

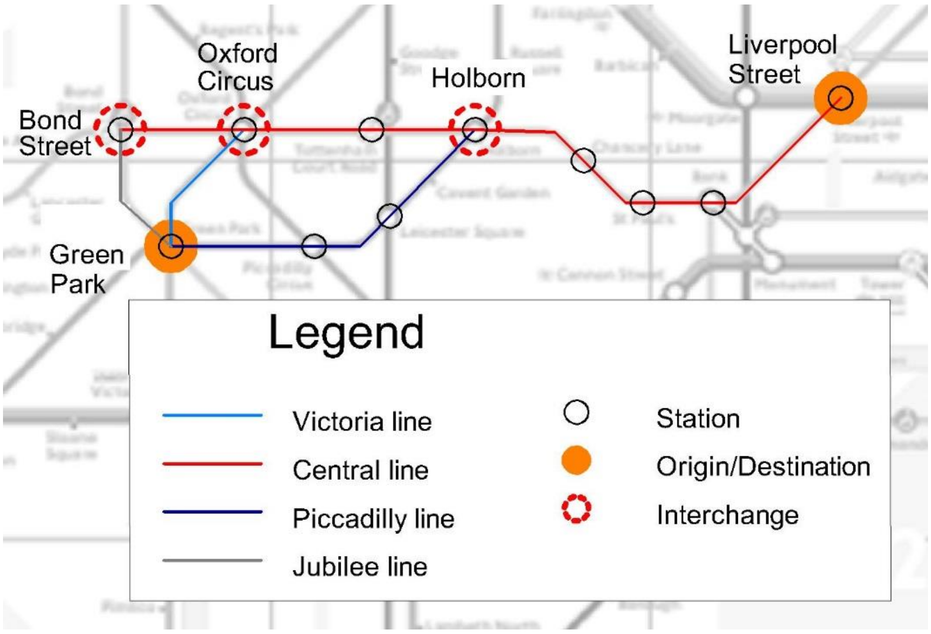


Figure 3-9 The **Liverpool Street** – **Green Park** OD pair

From Oyster data (**Section 3.6.1**) the OJT dataset is given for this OD pair, containing 30 transactions in the observation period (1 week in November 2009). Within this dataset all entries could be considered as valid data, because the upper outer fence (cf. **Section 3.2.1**) resulted 38 minutes, while the maximum OJT value is 36 minutes. This valid dataset is denoted by OJT^0 (**Figure 3-10**).

Having identified three reasonable routes on the map and from RODS data, route choice is estimated as a three-component ($N_R = 3$) finite mixture distribution. Therefore, the K-means clustering algorithm was applied on the OJT^0 dataset with three clusters to produce the initial values for the EM algorithm. The previously described (cf. **Section 3.3.1**) settings were used for centroid initialisation (K-means ++), distances (Euclidean square) and update methods (online phase). Conducting trials with various seed values for the random number generator (1, 2, 3, etc.) the K-means clustering algorithm gave two possible solutions for μ_r^{KMS}, σ_r^{KMS} and ω_r^{KMS} (**Table 3-9**).

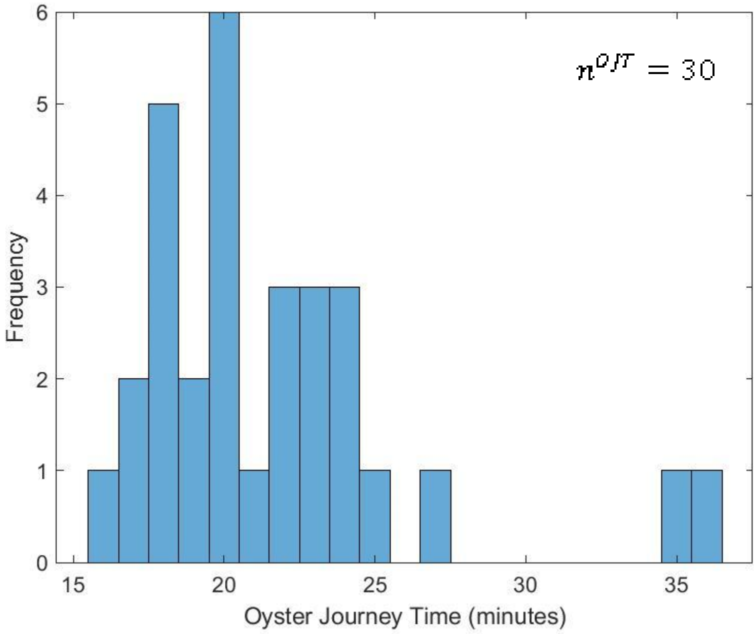


Figure 3-10 Distribution of Observed Journey Times for **Liverpool Street** – **Green Park**

Table 3-9 Results of the K-means clustering algorithm for **Liverpool Street** – **Green Park**

Label	K-means clustering		
r []	μ_r^{KMS} [min]	σ_r^{KMS} [min]	ω_r^{KMS} [%]
1	19.0	1.4	56.7%
2	23.0	1.5	36.7%
3	35.5	0.7	6.7%

Using these initial values the EM algorithm was run with different settings for the tolerance threshold (cf. **Section 3.3.2**). **Figure 3-11** and **Figure 3-12** presents the estimated mean (μ_1^{MIX}) and proportion (ω_1^{MIX}) for mixture component labelled with $r = 1$. There it is shown, that the EM algorithm converges to a solution close to the initial value; and it starts plateauing from the tolerance threshold of 1e-07 around the value of 18.6 minutes for the mean and 50.7%for the component proportion. Similar properties could be observed for the mixture component labelled with $r = 2$; while the mean and proportion of the third mixture component remains constant for all tolerance thresholds.

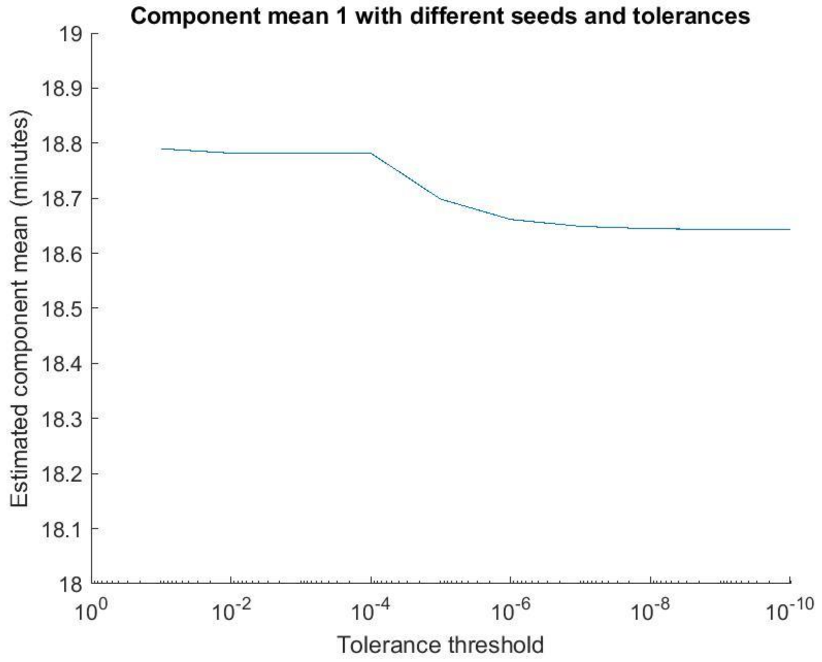


Figure 3-11 Estimated mean for mixture component 1, given different tolerance thresholds for **Liverpool Street** – **Green Park**

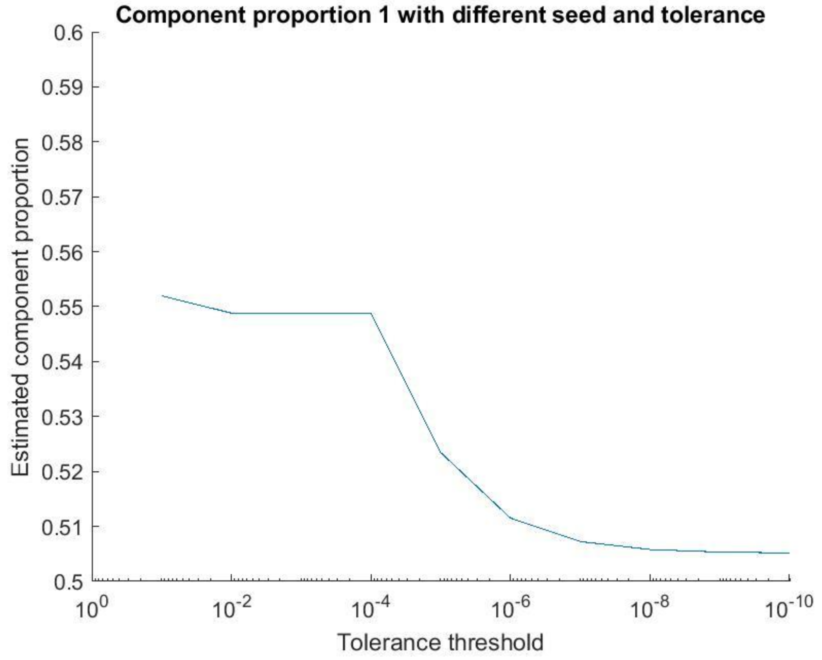


Figure 3-12 Estimated proportion for mixture component 1, given different tolerance thresholds for **Liverpool Street** – **Green Park**

Figure 3-13 presents the log-likelihood (equation (3-9)) for each initial value (seed) and tolerance threshold. It starts plateauing from the tolerance threshold of 1e-07. Due to these considerations the finite mixture model was applied with the tolerance threshold of 1e-07 (**Table 3-10**).

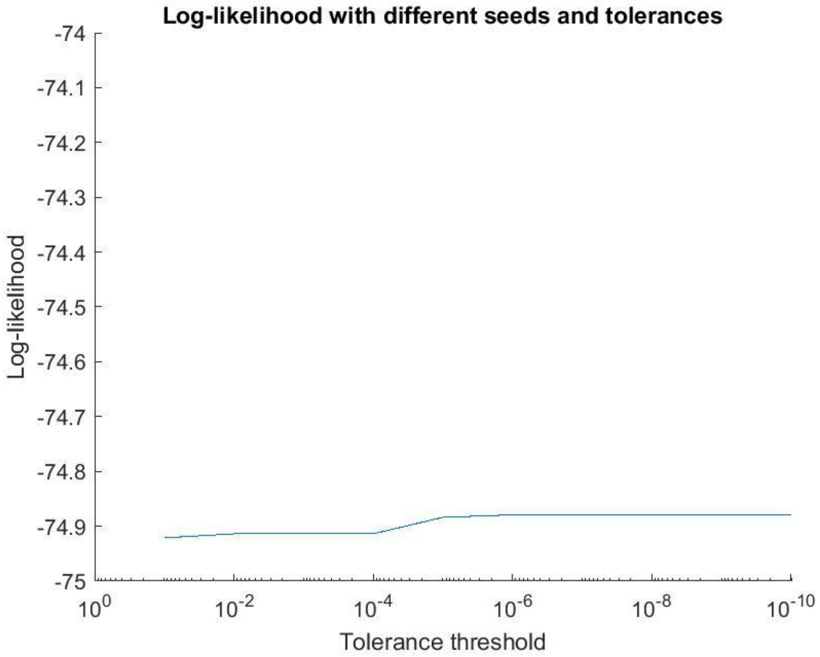


Figure 3-13 Log-likelihood, given different tolerance thresholds for **Liverpool Street – Green Park**

Table 3-10 Mixture model results; with tolerance threshold: 1e-07 for **Liverpool Street – Green Park**

Label	Mixture model		
r []	μ_r^{MIX} [min]	σ_r^{MIX} [min]	ω_r^{MIX} [%]
1	18.6	1.4	50.7%
2	23.0	1.9	42.6%
3	35.5	0.5	6.7%

Table 3-11 Journey time of actual London Underground routes for **Liverpool Street** – **Green Park**

Route				Journey Time [min]							
k	$l=1$	$s=1$	$l=2$	$t_{1,k}^{acc}$	$t_{1,k}^{wait}$	$t_{1,k}^{ob}$	$t_{1,k}^{ic}$	$t_{2,k}^{wait}$	$t_{2,k}^{ob}$	$t_{2,k}^{egr}$	t_{k}^{SJT}
1	**Central**	**Oxford Circus**	**Victoria**	2.6	0.9	10.0	2.9	1.0	2.0	1.9	**21.3**
2	**Central**	**Holborn**	**Piccadilly**	2.6	0.9	7.0	3.4	1.5	6.0	2.6	**24.0**
3	**Central**	**Bond Street**	**Jubilee**	2.6	0.9	11.0	3.2	1.0	1.0	3.5	**23.2**

Table 3-12 Matching mixture model results with the actual London Underground routes for **Liverpool Street** – **Green Park**

Red: Mixture results of proposed model, **Yellow**: Fu (2014), **Green**: actual LU routes

Mixture label	Journey Time (min)			Route Choice (%)			Route label	Route Matched		
	Mixture		Timetable	Mixture		Timetable		Line 1	Interchange 1	Line 2
r	μ_r^{MIX}		t_k^{SJT}	ω_r^{MIX}		ω_k^{RODS}	k	$l=1$	$s=1$	$l=1$
	Proposed	Fu(2014)		Proposed	Fu(2014)					
1	18.6	18.7	21.3	50.7%	35.9%	75.9%	1	Central	Oxford Circus	Victoria
2	23.0	22.0	23.2	42.6%	47.7%	11.9%	3	Central	Bond Street	Jubilee
3	35.5	27.6	24.0	6.7%	16.4%	12.2%	2	Central	Holborn	Piccadilly

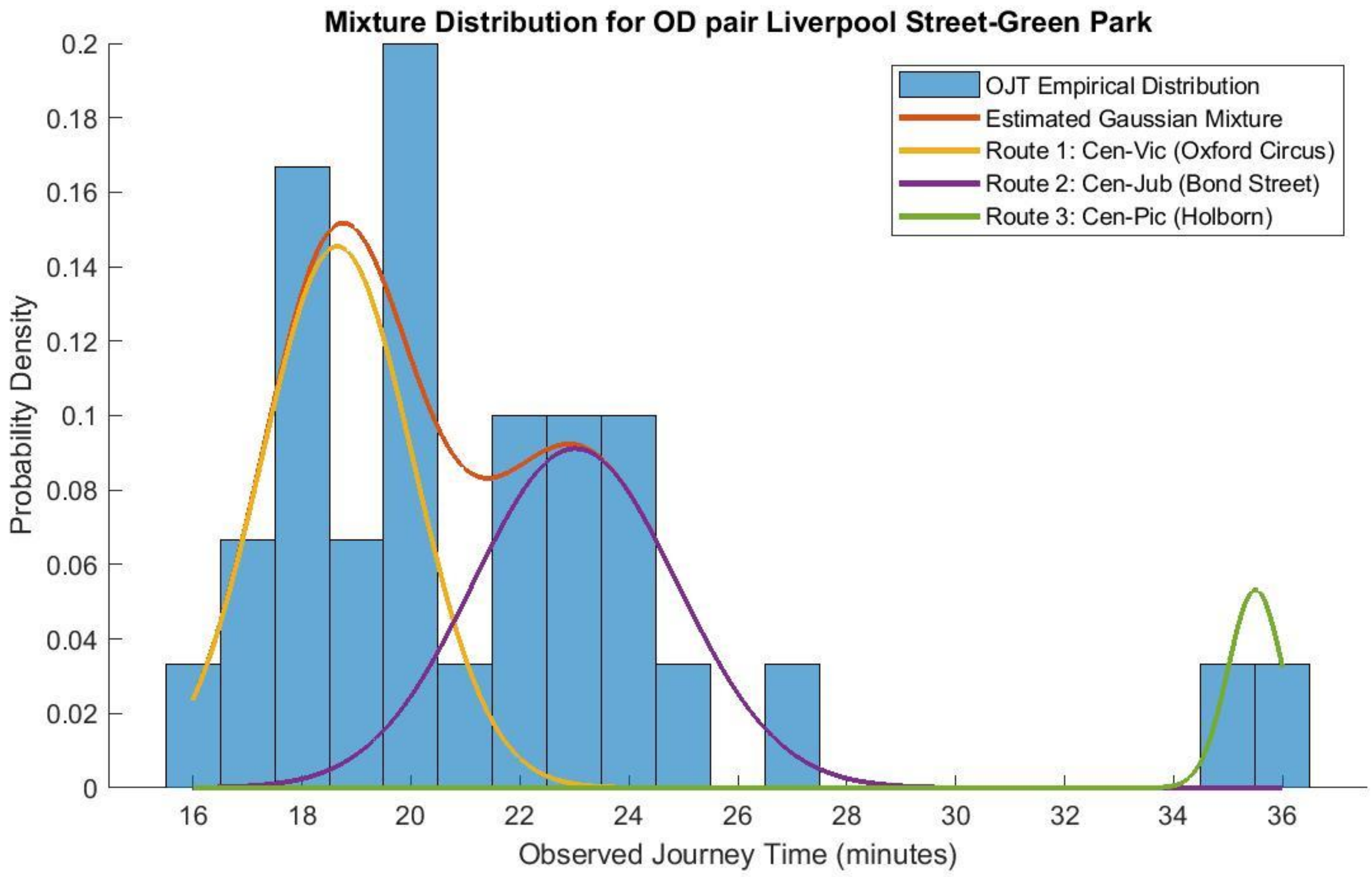


Figure 3-14 Estimated (Gaussian) journey time distribution of the routes for **Liverpool Street** – **Green Park**

Following this the results of the finite mixture model were matched with the actual LU routes (cf. **Section 3.4.1**). The total journey times of the actual LU routes (t_k^{SJT}) between the origin (**Liverpool Street**) and destination station (**Green Park**) were calculated based on equation (3-13). The results are presented in **Table 3-11**. The mixture components were matched with the actual LU routes in order of their journey times and the results are summarised in **Table 3-12** together with the results of Fu (2014) for the same OD pair. **Figure 3-14** presents the probability density functions of the mixture distribution fit on the CCOJT dataset as well as of the mixture components matched with the actual LU routes.

Based on the results with the finite mixture model the following issues were raised: Firstly, it is important to note that these results were obtained by setting the number of mixture components (N_R) to three as from the LU map and RODS data it was understood that the **Liverpool Street** – **Green Park** OD pair has three reasonable routes. To illustrate the importance of the correct specification of the route choice set the same finite mixture model was run with different settings for the number of mixture components (i.e. 1, 2, 3 and 4). The corresponding component means (μ_r^{MIX}) and proportions (ω_r^{MIX}) are shown in **Table 3-13** and **Table 3-14** respectively.

Table 3-13 Mean journey times with different number of mixture components for **Liverpool Street** – **Green Park**

# of mixture components	Mean journey time of mixture component		
N_R	μ_1^{MIX}	μ_2^{MIX}	μ_3^{MIX}
1	21.6	0.0	0.0
2	20.6	35.5	0.0
3	18.6	23.0	35.5

Table 3-14 Proportions with different number of mixture components for **Liverpool Street** – **Green Park**

# of mixture components	Mean journey time of mixture component		
N_R	ω_1^{MIX}	ω_2^{MIX}	ω_3^{MIX}
1	100.0%	0.0%	0.0%
2	93.3%	6.7%	0.0%
3	50.7%	42.6%	6.7%

The finite mixture model could produce results up to 3 components. When N_R was set 4, the EM algorithm could not converge as it created an ill-conditioned covariance at iteration 6. Additionally, in other cases of N_R (2 or 3), the mean and the proportion of the last component ($\mu_{N_R}^{MIX}$ and $\omega_{N_R}^{MIX}$) is around the same value (35.5 minutes and 6.7% respectively). As discussed earlier (**Case 1**), this is definitely not a good representation of the reality.

Comparing the results of the proposed finite mixture model with the actual routes (cf. **Table 3-12**) one can see that they do not show a good match. One reason for this could be due to the higher estimate for mixture component 2. Similarly to **Case 1**, also here the higher OJT values in the data sample would not necessarily mean that the passenger has taken a longer route, but it could be also because he/she has experienced fail-to-board delays in any of the routes, as the westbound platform of the **Central** line at **Liverpool Street** station is extremely crowded in the AM peak.

Another reason is, that applying the finite mixture model assuming three components may not give the best estimates. According to RODS data, in fact, there are three observed routes for this OD pair; however – looking at the map – the third shortest route (**Central** – **Jubilee**, via **Bond Street**) would be a sort of turning away from the destination (Dial, 1971). Furthermore, looking at the RODS data from other origin stations on the **Central** line (e.g. **Bethnal Green**), the option of interchanging at **Bond Street** does not appear among the reasonable routes. Therefore, it should be further examined whether assuming 2 or 3 routes reflects better the reality.

3.8 Issues with finite mixture models addressed in the thesis

Applying the finite mixture model on the LU network important issues were raised, which will be addressed in details in this thesis:

- The setting of the number of mixture components influences the model results (see **Section 3.8.1**)
- The finite mixture model may converge to different values depending on the setting of the initial values and tolerance thresholds (see **Section 3.8.2**)
- Longer OJT values may correspond to various reasons, not necessarily to the longer route (see **Section 3.8.3**)

3.8.1 Number of mixture components

Results showed that choosing the number of mixture components higher than the number of observed routes, the finite mixture model gives higher estimates of proportion for the mixture component with the highest mean (cf. **Table 3-8**), which would not be a true representation of the actual values.

Fu (2014) relied on pathfinding "by eye" from the LU map and used RODS data for determining the number of observed routes. Knowing the drawbacks of manual surveys (cf. **Section 3.6.3**) it would be advantageous to move away from them and to apply a route choice set generation algorithm that can automatically find the number of reasonable routes for a given OD pair.

Determining route choice set in complex metro networks, such as the LU is not a straightforward task; because there might be many physically possible routes for a given an OD pair, however only few of them are reasonable. The greatest challenge in this to find the cut-off value between reasonable and unreasonable routes and set a general rule for all OD pairs of the metro network. To address this issue, a pathfinding and attribute cut-off algorithm is proposed in **Chapter 4**.

3.8.2 Convergence of the finite mixture model

It is a known property of finite mixture models that depending on the initial values and tolerance thresholds they may converge to different local optima. In other terms, there are more possible solutions that can solve the mixture distribution problem. Based on the case study results (**Case 1**) it was initially illustrated that these solutions are not necessarily near each other (cf. **Figure 3-6**). As the finite mixture model is applied for route choice estimation, the modeller could not be confident which solution he/she could accept for route choice. Furthermore, in some cases (i.e. **Case 2**, with N_R=4) it can also happen, that the EM algorithm is unable to converge as it creates an ill-conditioned covariance. This could be also attributed to the small sample size of OJTs.

Although it is the initial value and the tolerance threshold that influences the most the convergence and the results of the finite mixture model (cf. **Section 3.3**), it can be logically understood that it is also related to the OJT sample size: the bigger the OJT sample is, the more regular is its distribution is and the better the convergence of the finite mixture model is. However to formulate this relationship exactly would require more advanced modelling, which is beyond the scope of this thesis.

One core objective of this thesis is to provide a framework for obtaining larger sample of OJTs for more reliable route choice estimates. To obtain this, in **Chapter 5** it is proposed to group those OD pairs, which have similar route choice patterns.

3.8.3 Reasons for the variation of the Observed Journey Times

The results of the finite mixture model were matched to the actual LU routes in a way that the mean journey time of the former (μ_r^{MIX}) were matched with the uncongested journey time of the latter (t_k^{SJT}) (cf. **Section 3.4.1**). In the case studies (e.g. **Case 1**) results illustrated that the journey time of the mixture component with the higher mean does not show a good match to the uncongested journey time of the corresponding route. Practically the longer journey time could be due to various reasons, such as fail-to-board delays at the origin or interchange stations, service delays or passenger carrying a heavy luggage. For the sake of simplicity, this thesis focuses only on the variable of fail-to-board delay in the model.

Fu (2014) accounted for fail-to-board delays in the matching process, but he made the simplified assumption that for each journey leg of each route half of the passengers can board the first train, half of them the second train. In reality, however, crowding levels on different routes and journey legs may significantly vary. Therefore in **Chapter 6** and **Chapter 7** a more detailed model is introduced for the consideration of fail-to-board delays. Furthermore in **Chapter 7**, the question of matching mixture results with actual LU routes are further discussed.

Chapter 4
Route choice set generation in complex metro networks

4.1 Introduction

In the general choice modelling context the appropriate consideration of the choice set is a prerequisite for the correct estimation of choice probabilities. This was further illustrated through the application of finite mixture models in complex metro networks. Case studies on the origin destination (OD) pairs of the London Underground (LU) has shown that setting the number of mixture components higher than the number of reasonable routes would give results, which are not a true representation of the actual values (cf. **Section 3.8.1**). Therefore it is crucial for the application of finite mixture models that the number of the mixture components – which corresponds to the number of reasonable routes of an OD pair – could be determined adequately.

Finding the route choice set in complex metro networks, such as the LU is a challenging task as – in theory – there could be several physically possible routes, however only a few of them are considered in the choice set of passengers.

There are several approaches to obtain information on route choice set. One possible approach is to conduct a survey on passengers' route choice, such as the Rolling Origin Destination Survey (RODS) in the context of the LU (cf. **Section 3.6.3**). The problem with this approach is that this survey might not be representative as the data is collected only from a few sample of the whole population on certain days of the year at certain stations. Furthermore – strictly speaking – from these surveys only the historical route choices of passengers can be understood, not the set of considered routes. In reality, route choice set of passengers remains unobserved (Bergantino et al., 2019).

Another approach is to infer from the available smart card records of Observed Journey Times (OJT) the number of mixture components as a prior step within he finite mixture model. To solve this problem Lee and Sohn (2015) proposed a reversible-jump Markov chain Monte Carlo simulation following the concept in Richardson and Green (1997) (cf. **Section 2.3.2**). They inferred the number of mixture components only based on the OJT dataset without actually considering the actual network properties. In reality, for the correct inference of the route choice set it is advantageous to use both sources of information. Therefore in this thesis, instead of following their approach, a simpler finite mixture is applied (Fu, 2014) together with a route choice set generation algorithm based on the metro network properties (i.e. journey times, interchanges) examining also scenarios, where the OJT records can serve as an additional information. The objective is to develop an algorithm that can automatically generate the route choice set based on the available data for most of the OD pairs of the metro network.

The rest of this chapter is structured as follows: Firstly, in **Section 4.2** existing literature is reviewed on route choice set generation methods including pathfinding and attribute cut-off; then in **Section 4.3**, the modelling challenges are discussed that arise in complex metro networks. Following this, in **Section 4.4** the representation of metro networks is discussed. Once the link times and the corresponding weights are known, the times and generalised costs of the routes can be obtained as explained in **Section 4.5**. After this, a pathfinding algorithm is applied to find a set of shortest routes. **Section 4.6** describes these algorithms in details focusing on their implementation for complex metro networks. Once a certain number of shortest routes were found, the main challenge is to narrow down this set to the set of reasonable routes by applying the appropriate cut-off criteria, which will be further explored in **Section 4.7**. **Section 4.8** concludes the chapter by summarising the findings and the possible extensions of the model.

This chapter builds on Nádudvari et al. (2016) following the concept of pathfinding algorithms, but proposing more detailed analysis on attribute cut-off methods.

4.2 Literature review on route choice set generation

In the general choice modelling context it is desirable that modellers could have adequate information or assumptions on the choice set for the correct estimation of choice probabilities (Swait and Ben-Akiva, 1987; Bovy, 2009; Bergantino et al., 2019). In the specific case of route choice in complex metro networks it is a challenging question, because for many OD pairs a very large number of physically possible routes are available, however only a few of them are considered by the passengers. This set is called the “reasonable route choice set”.

In literature there are approaches that interpret choice set consideration and choice estimation as a two-stage process (Manski, 1977; Gaundry and Dagenais, 1979; Başar and Bhat, 2004; Cantillo and Ortúzar, 2005). Firstly a set of attractive alternatives are selected from the universal choice set, and then the choice probabilities are estimated among those alternatives. On the other hand, there are also those, who argue that also the selection of the choice set is also an indicator of preferences, therefore it should be modelled in one stage with the choice (Horowitz and Louviere, 1995; Cascetta and

Papola, 2001; Swait, 2001; Martínez et al., 2009; Watling et al., 2018). Among them Watling et al. (2018) highlighted that in case of congested transport networks, not only route choice, but also the choice set may depend on the link flows.

Most of the route choice set generation methods can be summarised in two steps: The first step is a pathfinding algorithm to generate a certain number of routes for an OD pair (see **Section 4.2.1**); while the second step is the application of the attribute cut-off to find the set of reasonable routes among them (see **Section 4.2.2**). Additionally, there are other link-based approaches, which does not explicitly generate routes for finding the reasonable route choice set (see **Section 4.2.3**). After having presented these approaches, it is discussed, how the chosen route choice set generation methods could be applied in this thesis (see **Section 4.2.4**).

4.2.1 Pathfinding algorithms

The simplest approach for pathfinding is to search through all possible routes in a certain order and then select the adequate one among them (e.g. Brute-force, Breadth-first and Depth-first). While these approaches can be well applied in smaller networks; they reach their limitation for larger networks, such as the LU. This necessitates the application of efficient pathfinding algorithms.

Looking at literature reviews on pathfinding algorithms (Ramming, 2002; Fiorenzo-Catalano et al., 2004; Bekhor et al., 2006; Guo, 2008; Prato, 2009), the deterministic shortest path based methods were already proved to be adequate for the set objectives; therefore this literature review focuses on those methods.

The first step within the pathfinding algorithm is to find the shortest route for a given OD pair. The fundamentals for these algorithms started in the 1950s (Ford, 1956; Bellman, 1958; Dijkstra, 1959). The shortcoming of these methods is that they have higher computational time as they search the routes in all directions. To address this issue, more advanced algorithms have been developed, such as the A* (Hart et al., 1968), which starts searching routes only in promising directions.

Once the shortest route was found, the next step is to make a slight modification to the transport network and to find the shortest route on that modified network with one of the previously described shortest path algorithms. The modification to the network means eliminating one (i.e. K shortest path (Yen, 1971)) or more links (i.e. link elimination (Azevedo et al., 1993)) or increasing their link cost (link penalty (de la Barra et al., 1993)).

In the context of metro networks some studies worked with simpler search methods (see **Table 4-1**). Sun et al. (2015) applied Brute-force search in Singapore; while Sun et al. (2017); Xu et al. (2018) applied Depth-first search in Shanghai and Beijing respectively. Other studies applied efficient pathfinding algorithms. Hörcher et al. (2017) chose the K shortest path algorithm. Zhu and Xu (2016) implemented an improved Deletion Algorithm based on Depth-first search (Azevedo et al., 1990).

4.2.2 Attribute cut-off

While in the context of road networks it is sufficient to make simpler considerations for the reasonable route choice set by setting thresholds for their generalised costs; for metro networks it is a more complex question due to their special properties, such as interchanges, perception of the metro map and crowding (Raveau et al., 2014).

To account for these, most studies applied some heuristics as attribute the cut-off in addition to the threshold for the generalised costs of routes (see **Table 4-1**). Zhu and Xu (2016) – based on a travel survey in the Shanghai metro – considered a route reasonable, if its generalised cost is not more than 1.6 times or 10 minutes higher than the shortest route. Hörcher et al. (2017) worked with travel times of routes instead of generalised costs, and they considered a route reasonable if its travel time is not more than 1.5 times the shortest route. Furthermore – working on a relatively simple network of the Hong Kong metro – they searched only up to the second shortest path. Sun et al. (2017) worked with the natural logarithm of distances and applied the following heuristics to further filter the routes: (1) they should contain no loop, (2) if origin and destination station is on the same line, there is only 1 reasonable route for that OD pair, and (3) transfer time cannot be longer than the one third of the shortest route's travel time.

Xu et al. (2018) used additional constraints from smart card data and timetables. They calculated the longest possible journey time for each route of an OD pair, considering the worst case when the passenger is able to board only the third train at each journey leg. Among these routes they considered feasible those, whose longest possible journey time is shorter, than the maximum Observed Journey Time (OJT) value from smart card data. Additionally, they made the assumption that reasonable routes can have maximum 4 journey legs.

4.2.3 Link-based approaches

The concept of link-based methods (i.e. obviating the explicit enumeration of the routes) for route choice set generation started from Dial (1971), who set the criteria for a route to be considered reasonable, if every link in it:

1) Has its initial node closer to the origin node than is its final node (no turning back)
2) Has its final node closer to the destination node than its initial node (no turning away)

Table 4-1 Review on pathfinding algorithms and the attribute cut-off in metro networks

Reference	Method		Case study
	Pathfinding	**Attribute cut-off**	
Guo (2008)	Labelling + Optimal strategies		London
Sun et al. (2015)	Brute-force search		Singapore
Zhu and Xu (2016)	Deletion Algorithm	Gen. cost diff. (10 min) Gen. cost prop. (1.6)	Shanghai
Sun et al. (2017)	Depth-first search	Logarithm of distance No loop OD pairs on same line Transfer time prop. (1/3))	Shanghai
Hörcher et al. (2017)	K shortest path	Travel time prop. (1.5) Up to second shortest route	Hong Kong
Xu et al. (2018)	Depth-first search	Longest possible journey time vs OJT Max 3 interchanges	Beijing

Another possible approach is the concept of labelling (Ben-Akiva et al., 1984), which was also applied on the LU network for route choice set generation by Guo (2008). There, the labels correspond the weighting factors of time and interchange attributes. He followed the concept of optimal strategies (Spiess and Florian, 1989) for pathfinding, considering that a line segment going out of a station is utilised only, if its addition to the optimal strategy will reduce the total expected cost from that station to the destination. Once the set of reasonable routes were generated for different labels, they were compared with the set of used routes from RODS data; and those labels were proposed, which gave the best match between the two sets.

4.2.4 Discussions

Route choice set generation algorithms are discussed in this thesis; because the model that is applied to estimate route choice from smart card data (Fu, 2014) requires the number of reasonable routes as an input. From this, it logically follows, that the application of the two-stage approach (i.e. modelling route choice set and route choice as two distinct sub-models) would be more straightforward.

To comply with this objective, any pathfinding algorithm can be used, not necessarily the computationally most efficient one. Therefore the Dijkstra's algorithm was chosen for finding the shortest route and the K shortest path algorithm to generate a set of shortest routes as their program code was easily applicable (see **Section 4.6**). This choice was confirmed by the fact, that there are also other studies in literature, which apply the same pathfinding method (Hörcher et al., 2017).

In order to give the correct number or reasonable routes it is particularly important that the applied attribute cut-off method gives reliable results. Therefore it needs to be further examined (see **Section 4.7**), whether the existing attribute cut-off approaches can be applied with confidence also for the LU network, or additional criteria is required.

4.3 Modelling challenges in the London Underground

The proposed route choice set generation algorithm is applied on a subnetwork of the London Underground[13], which is the oldest and probably the most complex metro network of the world. This complexity requires several modelling challenges for network representation.

[13] See http://content.tfl.gov.uk/standard-tube-map.pdf for the map of the London Underground

London (i.e. Greater London, which includes the City of London and the 32 London boroughs) is a home to 8.8 million inhabitants[14] and is hosting around 20 million international visitors per year[15] as well as many more commuters and visitors from within the United Kingdom. To accommodate such a great demand an extensive public transport network has been developed since the mid of the 19th century, which includes the LU, London Overground (LO), Docklands Light Railway (DLR), Transport for London (TfL) rail, National Rail (NR) services, London Buses, London Trams and London River Services. Most of these transport subsystems (except for NR) are under the responsibility of TfL.

This thesis focuses on the LU, however it can be easily understood that modelling passenger flow in the LU is not an isolated problem, because at many stations it is connected with other rail subsystems (LO, DLR, TfL rail, NR, see **Figure 4-1**). In fact, the LU itself is a very complex transport system. It has 11 colour coded lines, however many of them have branches (e.g. **District**, **Northern** lines), short runs (e.g. **Victoria**, **Bakerloo** lines) or express services (e.g. **Metropolitan** line). Therefore, from the point of view of the modeller, it would mean much more than 11 lines.

Similarly, also modelling a station is a complex task. There are station complexes where more LU stations with different names are physically connected (common ticket gates), therefore passengers entering at one station can take lines from the other one (i.e. **Bank/Monument** station complex). On the other hand, there are stations, which are physically not connected (distinct ticket gates), but they have the same name (i.e. **Edgware Road**, **Paddington**, **Hammersmith** stations) (see **Figure 4-1**). Furthermore, some of the stations have more entrances, which are quite distant from each other, and also within a station there are multiple possible passageways between platforms.

Modelling passenger flows in the LU is a very challenging task for the following reasons: On the one hand, passenger flow of the LU needs to be modelled considering a very large network (beyond the LU network). On the other hand, due to the complexity of stations (multiple entrances and passageways) a more detailed understanding is necessary. To build a model for the entire LU and rail network of Greater London is definitely beyond the scope of this thesis. Therefore only a smaller problem, a subnetwork of the LU will be analysed, making the appropriate assumptions.

[14] https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland
[15] https://www.visitbritain.org/latest-quarterly-data-area

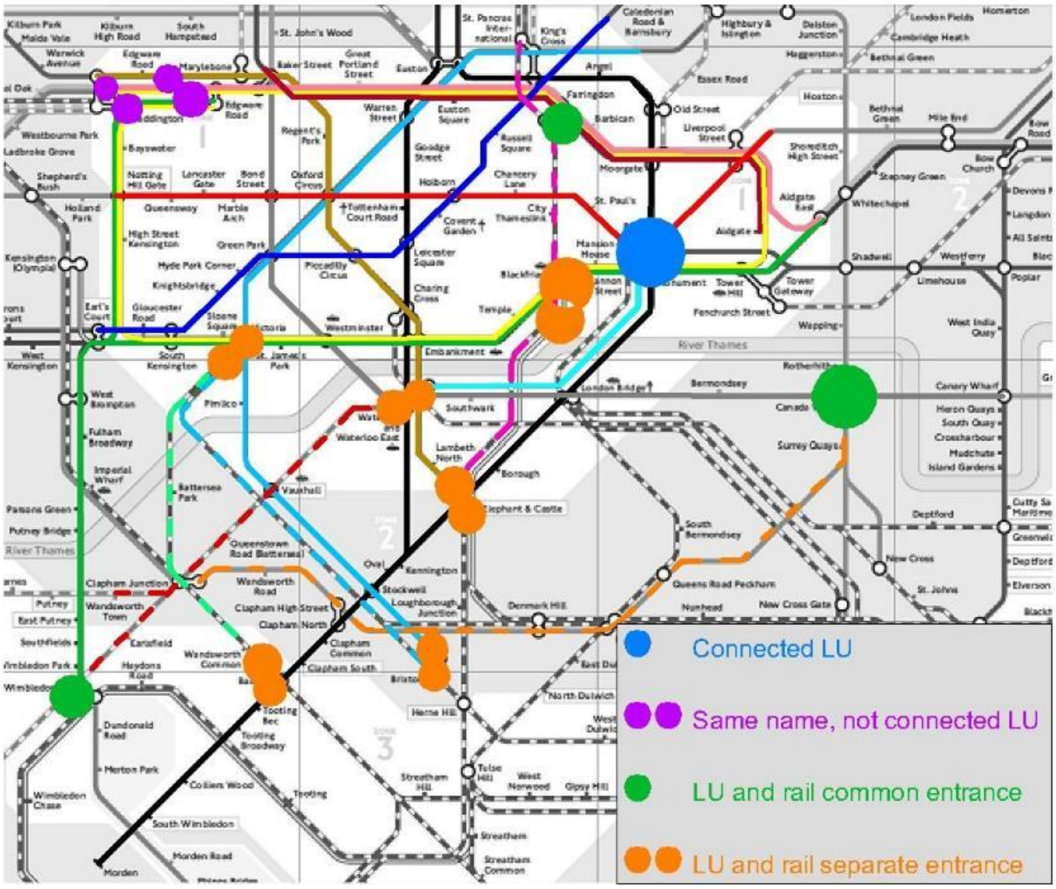


Figure 4-1 Examples for stations with special properties:

Connected LU stations (**blue**),

LU stations with the same name, but not connected (**purple**),

LU and rail stations with common ticket gate (**green**),

LU and rail stations with separate ticket gate (**orange**)

4.4 Representation of metro networks

In this chapter – in addition to what was presented in the previous chapters – the following notation is used:

Variable identifiers

a	Index of a link (arc)
i, j	Index of origin and destination station

Sets

$G(P, A)$	Metro network (graph) consisting of nodes (P) and links (A)
P	Set of nodes (points) in the metro network
A	Set of links (arcs) in the metro network
A^{ob}	Set of on-board links
A^{wait}	Set of wait links
A^{algiht}	Set of alight links
A^{acc}	Set of access links
A^{egr}	Set of egress links
A^{ic}	Set of interchange links
A, k, ij	Set of links on route k of OD pair ij
K_{ij}^{uni}	Universal route choice set for OD pair ij
K_{ij}^{gen}	Set of shortest routes generated for OD pair ij
K_{ij}	Set of reasonable routes for OD pair ij
K_{ij}^{obs}	Set of observed routes for OD pair ij

Variables

t_a	Travel time on link a (minutes)
t_{ij}^{min}	Minimum journey time for OD pair ij (minutes)
$c_{k,ij}$	Generalised cost of route k of OD pair ij (minutes)
$c_{k,ij}^{AEI}$	Total access egress interchange (AEI) cost of route k of OD pair ij (minutes)

$N_{K,ij}^{uni}$	Number of all theoretically possible routes for OD pair ij
$N_{K,ij}^{gen}$	Number of shortest routes generated for OD pair ij
$N_{K,ij}$	Number of reasonable routes for OD pair ij
$N_{K,ij}^{obs}$	Number of observed routes for OD pair ij
N_{OD}	Number of case study OD pairs
$N_{L,ij}^{min}$	Number of journey legs for the route with minimum number of journey legs for OD pair ij
$N_{D,i}$	Number of available directions at origin station
$N_{D,j}$	Number of available directions at destination station
OJT_{ij}^{max}	Maximum Observed Journey Times (OJT) record for OD pair ij (minutes)
ρ	Attribute cut-off
ρc	Attribute cut-off according to generalised cost proportion
$\rho c_{k,ij}$	Generalised cost proportion of route k of OD pair ij with respect to the shortest route
$\rho c_{ij}^{max,obs}$	Generalised cost proportion of the longest observed route
$\rho c_{ij}^{min,unobs}$	Generalised cost proportion of the shortest unobserved route

Parameters

w^{wait}	Weight of wait time
w^{AEI}	Weight of access egress interchange (AEI) time
w_a^{ic}	Weight having an interchange on link a (minutes)

ψ_a^{ob} 1 if $a \in A^{ob}$, otherwise 0

ψ_a^{wait} 1 if $a \in A^{wait}$, otherwise 0

ψ_a^{AEI} 1 if $a \in \left(A^{acc} \cup A^{egr} \cup A^{ic}\right)$, otherwise 0

ψ_a^{IC} 1 if $a \in A^{ic}$, otherwise 0

Functions

$f(\ \ )$ General notation for function

4.4.1 Definition of nodes and links

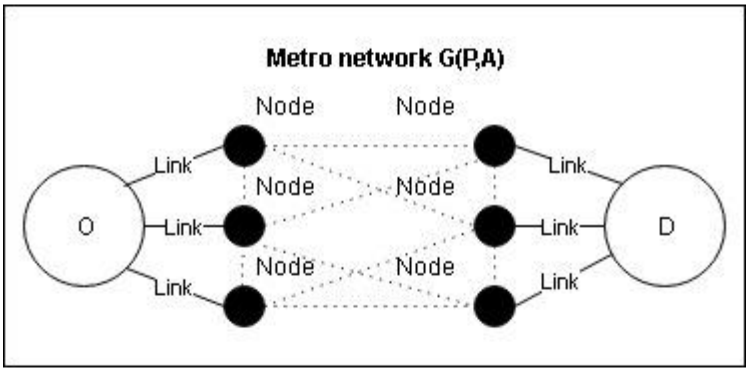


Figure 4-2 An OD pair in a metro network

Given an OD pair, ij in a metro network, $G(P, A)$, where P denotes the set of nodes (points, vertices) and A the set of links (arcs, edges) (**Figure 4-2**). Metro networks are specific as a journey between the origin and destination station consists of different characteristics of passenger movement (i.e. access from the ticket gate to the platform, waiting for the metro service, on-board travel, interchange between platforms and egress from the platform to the ticket gate). To account for this, in this thesis the following node types defined (**Figure 4-3**):

- *On-board node*: one node for each line at a station
- *Platform node*: one node for each pair of platforms at a station
- *Ticket gate node*: one node for each station

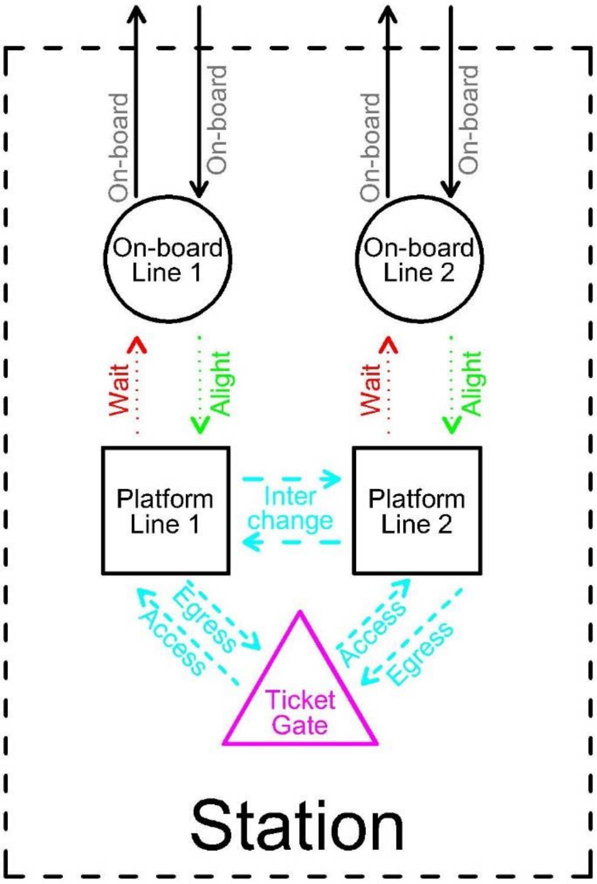


Figure 4-3 Definition of nodes and links at a station of a metro network

The reason why ticket gate nodes are defined is that the start and end of the journey of a passenger can be associated with the entry and exit smart card transaction at the ticket gate. For simplicity in this network model, all stations are considered with one ticket gate.

The links connecting these nodes and the corresponding subsets of links are:

- *On-board link* (A^{ob}): between on-board nodes of adjacent stations
- *Wait link* (A^{wait}): from platform node to on-board node of the same platform
- *Alighting link* (A^{algiht}): from on-board node to platform node of the same platform
- *Access link* (A^{acc}): from ticket gate node to platform node of the same station
- *Egress link* (A^{egr}): from platform node to ticket gate node of the same station
- *Interchange link* (A^{ic}): between platform nodes of the same station

For each link $a \in A$ its travel time is given and denoted with t_a. The times on on-board ($t_a \ if \ a \in A^{ob}$) and wait ($t_a \ if \ a \in A^{wait}$) links are taken from timetables (cf. **Section 3.6.2.1**), while the times on access ($t_a \ if \ a \in A^{acc}$), egress ($t_a \ if \ a \in A^{egr}$) and interchange ($t_a \ if \ a \in A^{ic}$) links are calculated based on station layouts known from the Nationwide Access Register (cf. **Section 3.6.2.2**). The times on alighting links are zero.

Table 4-2 Allocation of link types in the matrix of link times

	On-board node	**Platform node**	**Ticket gate node**
On-board node	**On-board links**	**Alighting links**	
Platform node	**Wait links**	**Interchange links**	**Egress links**
Ticket gate node		**Access links**	

These values are stored in the matrix of link times (**Table 4-2**). The reason why alighting links are defined is to make the allocation within the matrix symmetric. **Appendix A** presents the Matlab code to produce the matrix of link times automatically from the input data described above.

4.4.2 Case study network: the London Underground inner zone network

The route choice set generation algorithm is applied on a part of the LU network within Central London. Transport for London (TfL) defined fare zones for the LU and rail network (see **Figure 4-4**), where the network within Central London is called Zone 1. It includes the stations inside the **Circle** line plus some other stations (e.g. **Waterloo**, **London Bridge** and **Angel** stations).

In the case study of Schmöcker (2006), the term "inner zone" was used. This network, in addition to including most of the LU stations in Zone 1, it represents also the LU stations in other outer zones as "line specific stations" at the two ends of each LU line (see **Section 5.2.3**). Throughout the case studies of this chapter, this network consideration is followed; and the network is referred as the "LU inner zone network" (see **Figure 4-6**).

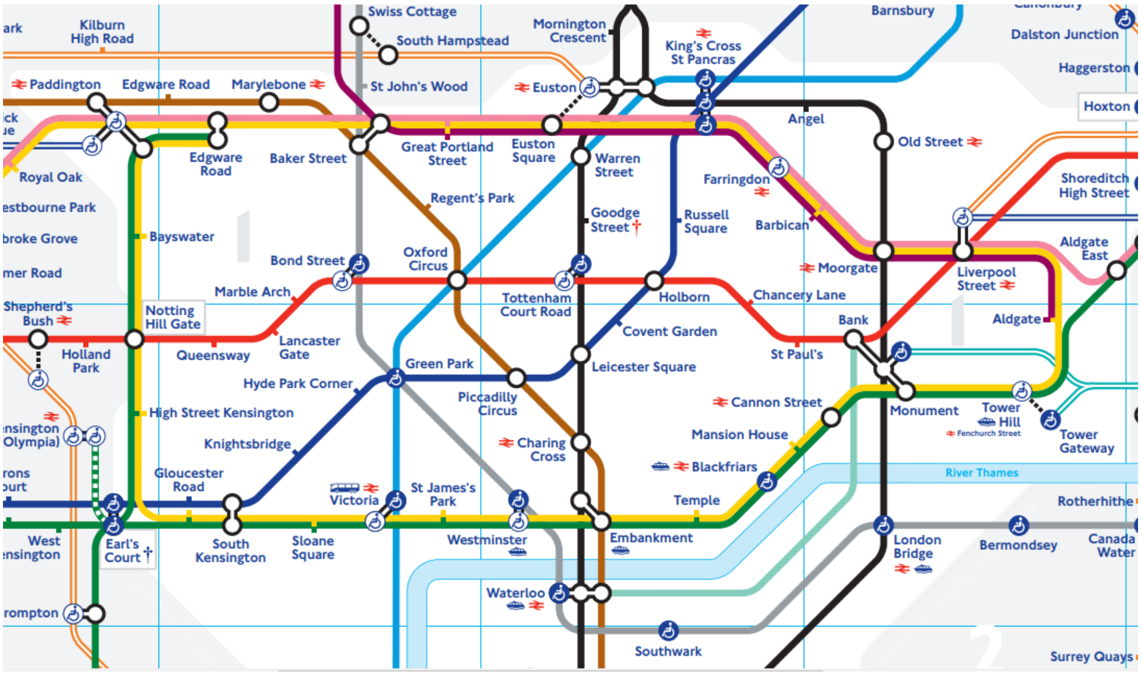


Figure 4-4 The London Underground network in Central London

The white background corresponds to Zone 1 in the fare scheme of TfL

source: http://content.tfl.gov.uk/standard-tube-map.pdf

4.4.3 Consideration of common lines

In addition to the principal rules for network representation, it is important to note how common lines are included in the model. The common line problem (Chriqui, 1975; Nguyen and Pallottino, 1988; Spiess and Florian, 1989) in metro networks occurs, when different lines depart from the same or from adjacent platforms. In this case passengers instead of choosing one line at a boarding platform, they may have a set of attractive lines (and hence the corresponding routes), and they board the line which arrives first within this set. This choice problem is called the choice of optimal strategies (Spiess and Florian, 1989) and the set of attractive routes are called hyperpaths (Nguyen and Pallottino, 1988).

In the LU inner zone network there are several line segments, where the common line problem occurs, due to the fact that some LU lines (i.e. **Circle**, **District**, **Hammersmith & City** and **Metropolitan** lines) share their track at a considerable length (**Figure 4-4**). Additionally, there are also cases, where at a station different LU lines depart from adjacent platforms (e.g. **Victoria** and **Bakerloo** lines at **Oxford Circus**) (**Figure 4-5**).

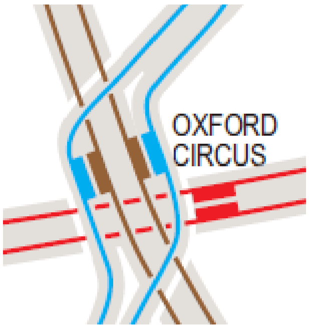


Figure 4-5 Common line problem, LU lines departing from adjacent platforms source: https://www.whatdotheyknow.com/request/track_map_london_undergound

Schmöcker (2006) stated that the consideration of common lines is important in the context of the LU inner zone. To account for this he used the link-based approach in the pathfinding algorithm without the explicit enumeration of each route (cf. **Section 4.2.3**). As the key objective of this chapter is to generate the set of reasonable routes as an input for the finite mixture model (cf. **Chapter 3**), it is required to follow the route-based approach for pathfinding (cf. **Section 4.2.4**). Enumerating all possible routes within the hyperpaths, would make the problem exceedingly complex (Nguyen and Pallottino, 1988), which is beyond the scope of this thesis. Therefore, at this point, the focus is still on the pure route choice problem in metro networks without yet considering optimal strategies.

In order to model the LU inner zone without the consideration of the common line problem, the following simplifications were made: Among the LU lines that share their track (i.e. **Circle**, **District**, **Hammersmith & City** and **Metropolitan** lines), only the **Circle** line was included with a frequency of 20 trains/hour as at most of its length (i.e. **Gloucester Road** – **Tower Hill** segment, where it shares the track with the **District** line; as well as the **Liverpool Street** – **Baker Street** segment, where it shares the track with the **Hammersmith & City** and **Metropolitan** lines) the combined frequency is around that value. In order to account for the segments, where only the **Circle** line is available (i.e. **Tower Hill** – **Aldgate**[16] and **High Street Kensington** – **Gloucester Road** links), an adjustment of 3.5 minutes was made to the corresponding on board links as passengers travelling on that route have a an average wait time of 5 minutes instead of 1.5 minutes (cf. (3-15) for the relationship between service frequency and wait time). In reality, also

[16] It is also known, that the trains on the **Circle** line stopping at **Aldgate** station wait for a longer time to keep themselves to the schedule. This is included in the timetable data, which was used for the analysis. Therefore no further adjustments were required for this.

along the **Baker Street** – **High Street Kensington** segment, the combined frequency is less than 20 trains/hour (around 12 trains/hour) as there the **Circle** line shares its track only with the **Hammersmith & City** and with the **Edgware Road** branch of the **District** line; however this was not considered in the network model (**Figure 4-6**).

4.4.4 Network size

Having made the above described considerations, the network model of the LU inner zone has the following characteristics: In the LU, there are 11 colour coded lines; however due to the fact, that many of these lines have branches, short runs or express services, the number of the lines for the network model would be more than that (cf. **Section 4.3**). In the LU inner zone, the two branches of the **Northern** line (via **Bank** and via **Charing Cross** stations) would count as two distinct lines. Regarding common lines (cf. **Section 4.4.3**) instead of the LU lines that partially share their track (i.e. **Circle**, **District**, **Hammersmith & City** and **Metropolitan** lines), only the **Circle** line is considered. With these considerations, 9 lines are included in the network model. Among the 68 stations, 55 are stations of the LU inner zone and 13 are line specific stations at the two ends of the LU lines. The reason why this number is odd, because the two **Northern** line branches has the same line specific station on the north (**Figure 4-6**).

Following the definition of nodes and links in **Section 4.4.1**, the case study network is represented with 280 nodes in total, among which 106 are on-board node, 106 platform node and 68 are ticket gate node. These nodes are connected with 722 links.

4.4.5 Case study origin destination pairs

The route choice set generation algorithm is applied on the OD pairs presented in **Table 4-3** and **Figure 4-6**. For this case study those OD pairs were chosen that have two or more observed routes according to the Rolling Origin Destination Survey (RODS, cf. **Section 3.6.3**). For all of these OD pairs, the nature of passenger choice is a pure route choice problem, without facing the problem of common lines. (cf. **Section 4.4.3**). Although from the smart card dataset provided for this research (cf. **Section 3.6.1**), it seems that there is no considerable demand for **OD 4**, **OD 6** and **OD 7** (see **Table 4-7**); looking at Fu (2014) it was understood that there are still passengers travelling between those origins and destinations; therefore analysing the route choice set for those OD pairs still makes sense.

Table 4-3 The case study OD pairs in the LU inner zone network

OD pair	Origin	Destination
1	**Victoria**	**Holborn**
2	**Euston**	**St. James's Park**
3	**Victoria**	**Liverpool Street**
4	**Angel**	**Waterloo**
5	**Liverpool Street**	**Green Park**
6	**Euston**	**South Kensington**
7	**Victoria**	**Waterloo**

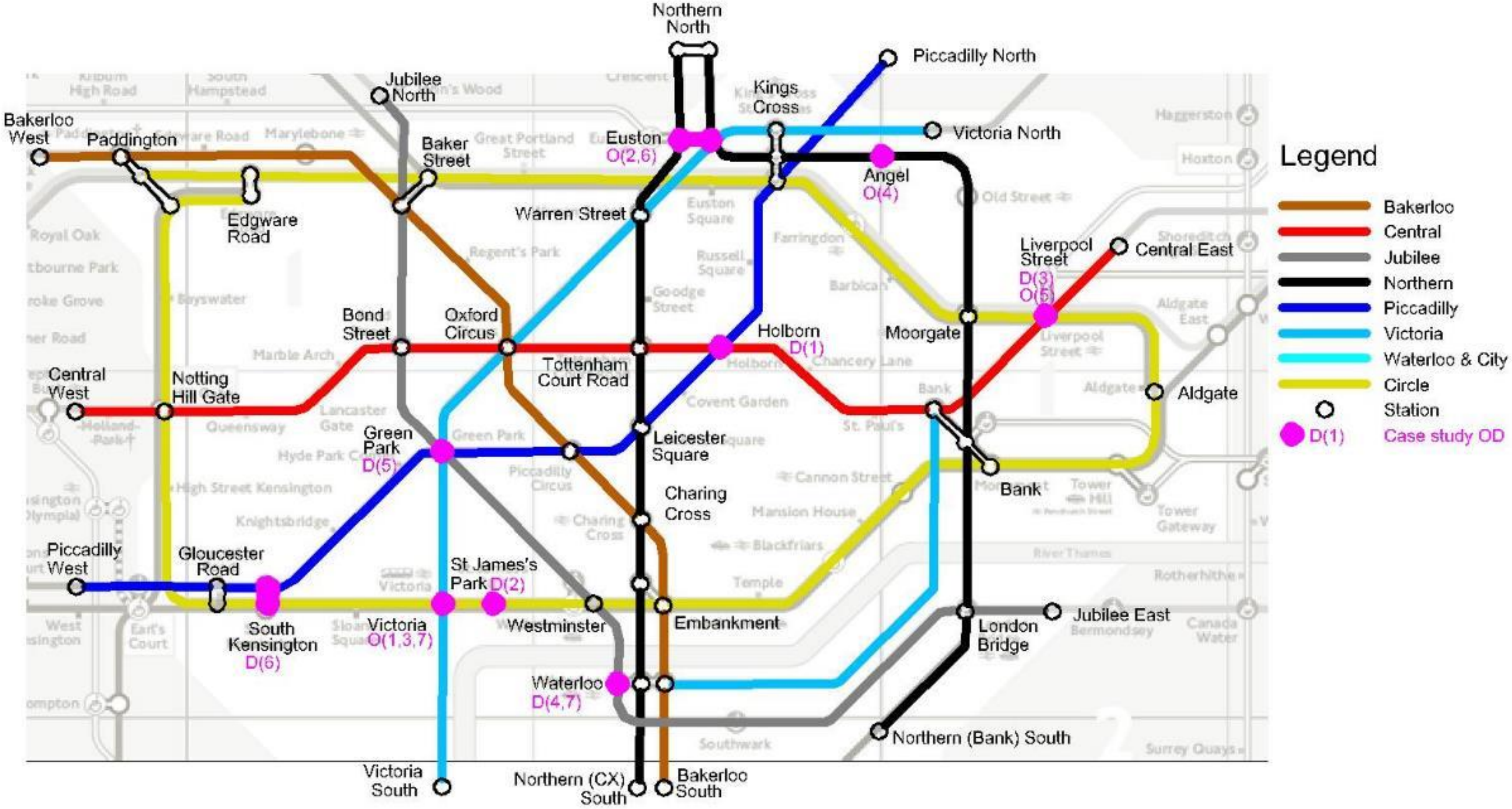


Figure 4-6 Case Study (London Underground inner zone) network and OD pairs

4.5 Journey time and generalised costs of routes

Given the representation of metro network, $G(P, A)$ with the set of nodes (P) and links (A), route k of OD pair ij can be defined as a sequence of links between origin station i and destination station j. Let A, k, ij denote the set of links on route k of OD pair ij.

Knowing all the link times (t_a) from the available data sources (cf. **Section 3.6.2**), the total (scheduled) journey time of route k of OD pair ij is:

$$t_{k,ij}^{SJT} = \sum_{a \in A,k,ij} t_a \tag{4-1}$$

Given the fact that different types of movements are perceived differently by passengers (cf. **Section 2.2**), the generalised cost of routes can be defined to take into consideration the journey time components (i.e. on-board, wait, AEI) with their corresponding weights (w):

$$c_{k,ij} = \sum_{a \in A,k,ij} t_a \cdot \left(\psi_a^{ob} + w^{wait} \cdot \psi_a^{wait} + w^{AEI} \cdot \psi_a^{AEI}\right) + w_a^{ic} \cdot \psi_a^{ic} \tag{4-2}$$

The weight of wait time (w^{wait}) expresses that according to the perception of passengers, one minute of wait time is equivalent to how many minutes of on-board time. Similar explanation can be made for the weight of AEI time (w^{AEI}). For these weights, the values were taken from an earlier study applied on the LU network (Raveau et al., 2014). There, they calibrated the parameters of a C-Logit model using RODS data for the route choice observation and obtained the results for w^{wait} and w^{AEI} (**Table 4-4**). These values refer to the perception of passengers on weekdays, morning peak assuming that trips were done with restrictive purpose.

At this point it is important to note that for the correct estimation of the weights (w^{wait} and w^{AEI}) the calibration should be done with the same model specification as (4-2) (i.e. Multinomial Logit, MNL), not with the C-logit. However, not finding an adequate MNL model calibration for the LU network, it was chosen to apply the values understood from an LU specific study. This was also justified by the fact that the numerical values of these weights (1.93 and 1.30 respectively) seems to be a good description of passengers' perception.

Another issue is that in order to construct the simplified network model – which does not consider the common line problem, but still counts for the different frequencies along the

Circle line – an adjustment had been applied to the corresponding on-board links (cf. **Section 4.4.3**), which are in reality the differences in the wait time. To account for this, in the process of coding the adjustment was made not only to the link times, but also the corresponding generalised costs (analogously to the code reported in **Appendix A**).

Table 4-4 Weighs of wait and access egress interchange (AEI) time, based on Raveau et al. (2014)

Weight	Value
w^{wait}	1.93
w^{AEI}	1.30

In addition to the journey times, there is an additional term expressing that the fact of having an interchange is equivalent to how many minutes of on-board time (w_a^{IC}). In the context of LU, the type and size of interchanges significantly vary: There are simpler cases of interchanges, where passengers need to move only between adjacent platforms (e.g. between the **Victoria** to **Bakerloo** lines at **Oxford Circus** station, cf. **Figure 4-5**). At the same time, there are complex stations, where passengers need to walk up to 6 minutes between far away platforms (e.g. **Bank**/**Monument** station complex). To account for this, Raveau et al. (2014) defined w_a^{IC} in function of the level (i.e. ascending, even and descending) and assistance (i.e. assisted, semi-assisted and non-assisted) of the interchange movement; and they obtained the results presented in **Table 4-5**.

Table 4-5 Weights of the fact of having an interchange (minutes) in function of the station characteristics (level and assistance), based on Raveau et al. (2014)

Characteristics		w_a^{IC}
Level	Assistance	[min]
Ascending	Assisted	5.71
	Semi-Assisted	6.84
	Non-Assisted	7.32
Even	N/A	2.39
Descending	Assisted	4.87
	Semi-Assisted	5.97
	Non-Assisted	6.49

4.6 Pathfinding algorithm

It has been previously explained (cf. **Section 4.5**) that the generalised cost of route k of OD pair ij ($c_{k,ij}$) can be calculated as the weighted sum of the link times (t_a, cf. equation (4-2)). However the question still remains, how these routes can be found between the origin (i) and destination station (j), given the transport network ($G(P,A)$).

The universal route choice set for OD pair ij can be denoted as K_{ij}^{uni}. It includes all, $N_{K,ij}^{uni}$ number of theoretically possible routes. To find K_{ij}^{uni}, simple search methods could be applied (e.g. Brute-force search, cf. **Section 4.2.1**). However the problem is, that in complex metro networks, $N_{K,ij}^{uni}$ can be very large, therefore the computational time would be exceedingly high. For example, for certain OD pairs in the case study network of the LU inner zone (cf. **Section 4.4.5**), there can be up to thousands of theoretically possible routes and to find all of them with simple search methods would be computationally expensive.

Therefore, instead of finding all possible routes for an OD pair ij, the aim here is to generate a sufficiently large set of shortest routes (K_{ij}^{gen}), for which it can be ensured, that it contains all reasonable routes (K_{ij}):

$$K_{ij} \subseteq K_{ij}^{gen} \subseteq K_{ij}^{uni} \tag{4-3}$$

For the number of routes in these route choice sets, the following inequality holds:

$$N_{K,ij} \leq N_{K,ij}^{gen} \leq N_{K,ij}^{uni} \tag{4-4}$$

Following the concept of formulae (4-3) and (4-4), the set of reasonable routes can be obtained in two steps

1) Generate a sufficiently large set of shortest routes (K_{ij}^{gen})
2) Narrow down this set to the set of reasonable routes (K_{ij})

This section focuses on the first step: pathfinding; while in **Section 4.7**, the second step: the attribute cut-off is discussed. Following the literature review in **Section 4.2.1**, the K shortest path algorithm was chosen to be applied for pathfinding (Yen, 1971) together with the Dijkstra (1959) algorithm for finding the shortest route.

In this section the K shortest path algorithm is described in details (see **Section 4.6.1**); and it is further discussed what modifications are necessary when it is implemented for complex metro networks (see **Section 4.6.2**). Finally, the algorithm is applied on the case study OD pairs and the results for the set of shortest routes are presented (see **Section 4.6.3**). The detailed description of the Dijkstra (1959) algorithm is presented in **Appendix B**.

4.6.1 The K shortest path algorithm

The K shortest path algorithm (Yen, 1971) can be described with the following steps (see **Figure 4-7**):

1. Find the shortest route (i.e. 1-2-4-6) using the Dijkstra (1959) algorithm.
   Set it as the current path.
2. Find the next shortest routes
   2.1. Set the first node of the current path (i.e. 1) as the deviation vertex.
   Eliminate the link on the current path (i.e. 1-2-4-6)
   which starts from the deviation vertex (i.e. 1-2).

2.2. Find the shortest path from the deviation vertex to destination node on this modified network (i.e. 1-3-5-6) using the Dijkstra (1959)

2.3. Set the next node on the current path (i.e. 2) as the deviation vertex

repeat steps 2.2-2.3 to find the next shortest routes

until the last node on the current path is reached

2.4. Select the shortest route among the newly found routes

Set it as the current path

3. Repeat step 2 until the set number of shortest routes ($N_{K,ij}^{gen}$) are found

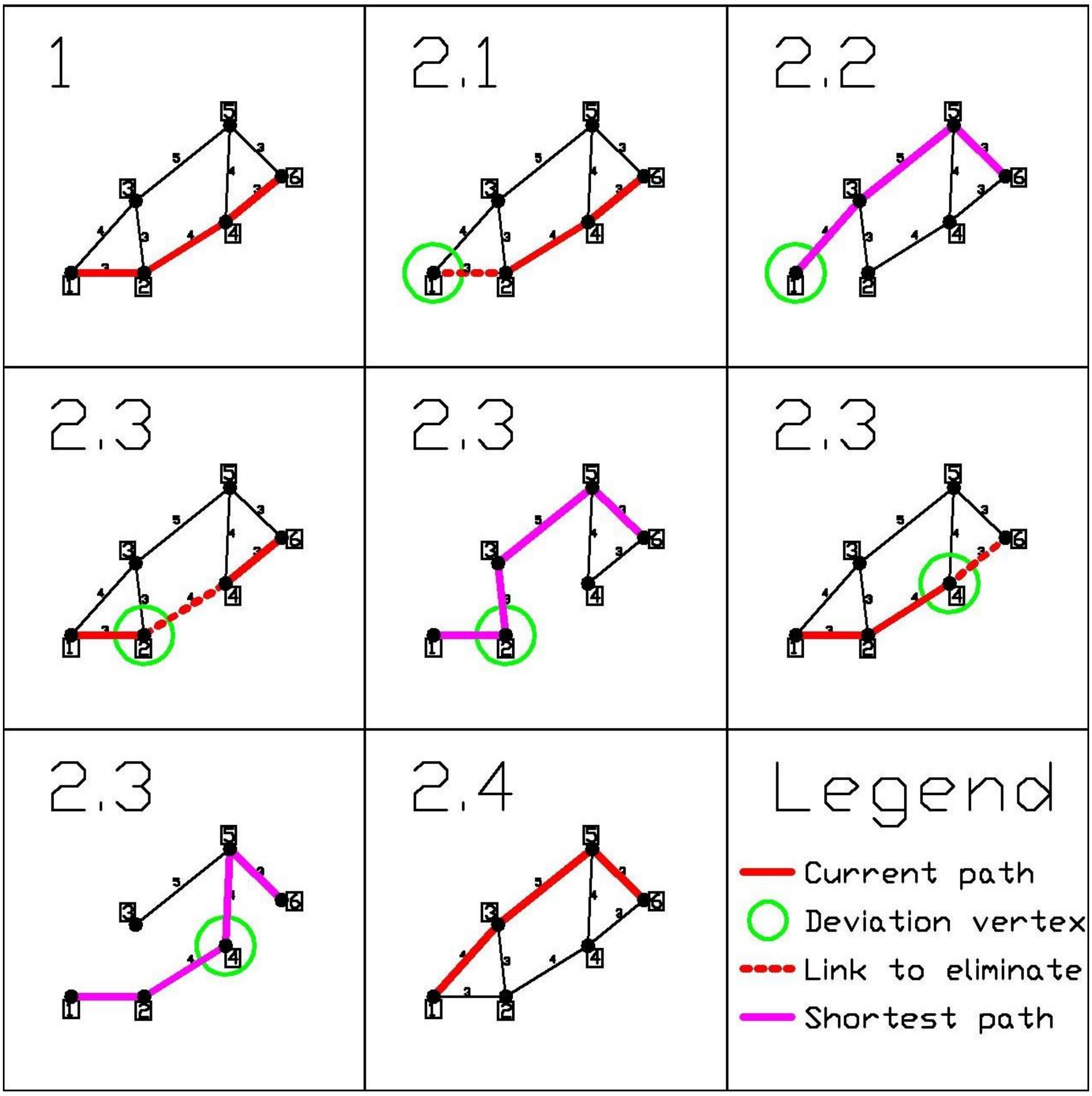


Figure 4-7 Illustration of the K shortest path algorithm on a small example network

It is important to note that as the K shortest path algorithm uses the Dijkstra algorithm in every iteration to find the shortest path between the deviation vertex and the destination node (step 2.2) it requires high computational time. There are faster methods in literature, however the K shortest path algorithm was proved to be adequate for the purpose of this thesis (cf. **Section 4.2.4**).

The K shortest path algorithm was applied in Matlab on the network model of the LU inner zone (cf. **Section 4.4.1**) for the case study OD pairs (cf. **Section 4.4.5**). It is set to search for and order the routes based on their generalised costs ($c_{k,ij}$, cf. **Section 4.5**). The program code is available from the Matlab file exchange website[17]. As it follows, it is discussed what modifications are made to the original program code in order that it could be implemented for metro networks (see **Section 4.6.2**).

4.6.2 Proposed modifications to account for multiple passageways within stations

Running the K shortest path algorithm (cf. **Section 4.6.1**) on the LU inner zone network, the results would contain many route variants, which differ only in their AEI movements within the stations. For example, such route variant would be, when a passenger at the origin station accesses the chosen line via the platform of another line. Similarly, another route variant could be when at the interchange station he/she walks to the chosen line via the ticket gate (see **Figure 4-8** a). In order to avoid finding these route variants, it is necessary that the algorithm could eliminate some of the AEI links automatically depending on the OD pair and the route. Therefore the following modifications are proposed for the K shortest path algorithm:

1) For the current OD pair
   Eliminate (**x**) interchange links at origin and destination stations
   Eliminate (**x**) access and egress links at all other stations (see **Figure 4-8** b).
2) At every iteration,
   if the deviation vertex is set at the platform node of a station,
   eliminate (**x**) interchange links from other platforms (see **Figure 4-8** c).

The program code for these modifications are presented in **Appendix C**.

[17] http://uk.mathworks.com/matlabcentral/fileexchange/32513-K_shortest-path-yen-s-algorithm

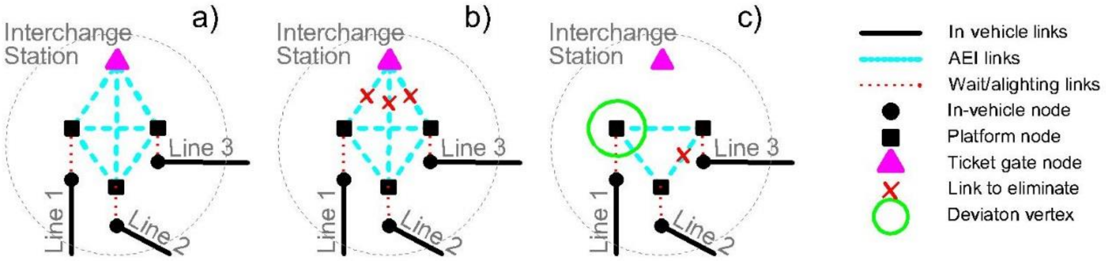


Figure 4-8 Proposed modification to the K shortest path algorithm

4.6.3 Set of shortest routes for the case study OD pairs

Once the proposed modifications were made to the K shortest path algorithm (cf. **Section 4.6.2**), it was applied on the 7 case study OD pairs of the LU specified in **Section 4.4.5**. In order to ensure that that the set of reasonable routes (K_{ij}) can be a subset within the set of generated routes (K_{ij}^{gen}) (cf. formulae (4-3) and (4-4)) the number of generated routes ($N_{K,ij}^{gen}$) was chosen sufficiently large (i.e. 10); in accordance with Guo (2008), who showed that in the LU 99% of the OD pairs has up to 4 observed routes. At this point, the number of generated routes was set independently from the type of the OD pair. As it follows, it will be examined how the number of reasonable routes depend on the OD attributes (see **Section 4.7.4** and **Section 5.3.3**).

Figure 4-9 presents the 10 shortest routes generated with the K shortest path algorithm for the **Victoria** – **Holborn** OD pair. **Table 4-6**, describes these routes with their journey time ($t_{k,ij}^{SJT}$, cf. equation (4-1)) and generalised costs ($c_{k,ij}$, cf. equation (4-2)). To get a better understanding on the interchange attributes of these routes also their total interchange time ($t_{k,ij}^{ic}$) and the total AEI cost ($c_{k,ij}^{AEI}$) is reported, which can be obtained as:

$$t_{k,ij}^{ic} = \sum_{a \in A,k,ij} t_a \cdot \psi_a^{ic} \tag{4-5}$$

and

$$c_{k,ij}^{AEI} = \sum_{a \in A,k,ij} t_a \cdot w^{AEI} \cdot \psi_a^{AEI} + w_a^{ic} \cdot \psi_a^{ic} \tag{4-6}$$

respectively.

Furthermore, for each route k, the proportion of their generalised cost with respect to the shortest route $(c_{1,ij})$ is calculated as:

$$\rho c_{k,ij} = \frac{c_{k,ij}}{c_{1,ij}} \tag{4-7}$$

In the same table the set of observed routes (K_{ij}^{obs}, known from RODS data, cf. **Section 3.6.3**) is highlighted with green. The results for the other OD pairs are presented in **Appendix D**.

In addition to these route attributes there are also OD specific properties to serve as an input for setting the attribute cut-off criteria:

- Number of observed routes according to RODS data $(N_{K,ij}^{obs})$
- Journey time of the route with minimum journey time (t_{ij}^{min})
- Generalised cost of the shortest route $(c_{k,ij})$
- Sample size of Observed Journey Times (OJT) (n_{ij}^{OJT})
- Maximum Observed Journey Times (OJT) record (OJT_{ij}^{max})
- Number of journey legs of the route with minimum number of journey legs $(N_{L,ij}^{min})$
- Number of available directions at origin $(N_{D,i})$ and destination $(N_{D,j})$ station

Number of available directions at a station means the number of directions that stays within the LU inner zone. For example, at **Victoria** station, $N_{D,i} = 3$, because the **Victoria** line can be taken only northbound as the southbound direction does not stay within the LU inner zone, but leads to the **Victoria South** line specific station. Additionally, the **Circle** line in both directions stays within the LU inner zone. Therefore in total there are 3 available directions (cf. **Figure 4-6)**.

These properties are resumed in **Table 4-7**. The applied method for the attribute cut-off is discussed in **Section 4.7**.

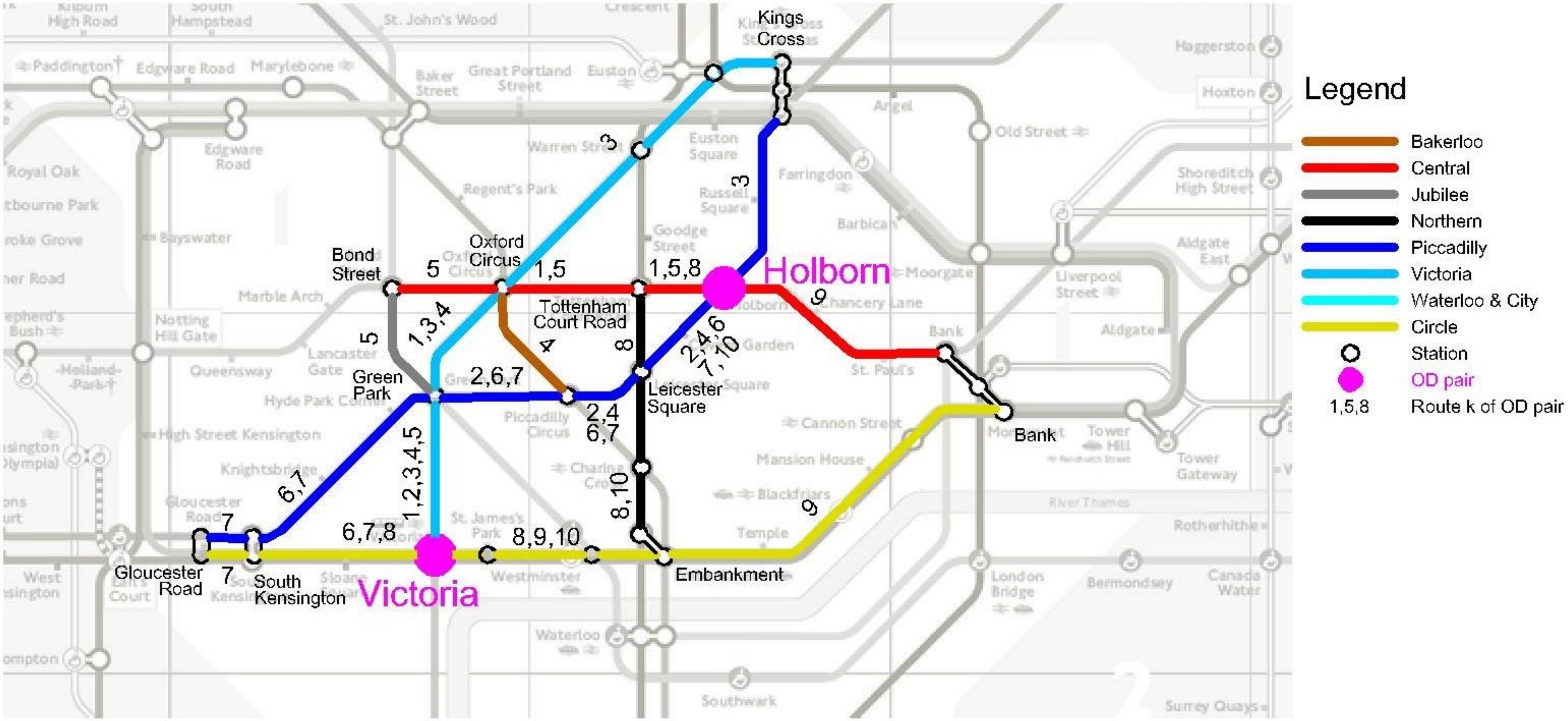


Figure 4-9 The 10 shortest routes for **Victoria** - **Holborn**

Table 4-6 The 10 shortest routes for **Victoria** - **Holborn** with their journey time and generalised cost, observed routes (Rolling Origin Destination Survey, RODS) are highlighted with green

Route						**Time**		**Generalised cost**		
ID	**Line 1**	**IC 1**	**Line 2**	**IC 2**	**Line 3**	**Total**	**IC**	**Total**	**AEI**	**Proportion**
k,ij $ij=1$						$t^{SJT}_{k,ij}$ [min]	$t^{ic}_{k,ij}$ [min]	$c_{k,ij}$ [min]	$c^{AEI}_{k,ij}$ [min]	$\rho c_{k,ij}$
1	**Victoria**	**Oxford Circus**	**Central**			17.2	3.4	28.3	18.4	1.00
2	**Victoria**	**Green Park**	**Piccadilly**			19.5	3.7	29.7	17.8	1.05
3	**Victoria**	**Kings Cross**	**Piccadilly**			24.2	3.4	34.2	17.4	1.21
4	**Victoria**	**Oxford Circus**	**Bakerloo**	**Piccadilly Circus**	**Piccadilly**	23.2	2.9	38.5	20.8	1.36
5	**Victoria**	**Green Park**	**Jubilee**	**Bond Street**	**Central**	20.9	5.0	38.6	25.8	1.37
6	**Circle**	**South Kensington**	**Piccadilly**			28.7	3.2	40.1	17.3	1.42
7	**Circle**	**Gloucester Road**	**Piccadilly**			30.6	2.1	41.7	15.9	1.47
8	**Circle**	**Embankment**	**Northern (CX)**	**Tottenham Court Rd**	**Central**	24.3	4.2	43.6	24.9	1.54
9	**Circle**	**Bank**	**Central**			32.4	5.8	44.0	20.2	1.56
10	**Circle**	**Embankment**	**Northern (CX)**	**Leicester Square**	**Piccadilly**	24.9	2.8	44.0	23.3	1.56

Table 4-7 Summary of the OD pairs and their properties

Index	Origin	Destination	# Observed Routes	Min time	Min gen. cost	# OJT records	Max OJT record	Min # journey legs	Origin # directions	Destination # directions
ij	i	j	$N_{K,ij}^{obs}$	t_{ij}^{min}	$c_{1,ij}$	n_{ij}^{OJT}	OJT_{ij}^{max}	$N_{L,ij}^{min}$	$N_{D,i}$	$N_{D,j}$
1	Victoria	Holborn	2	17.2	28.3	54	31	2	3	4
2	Euston	St James's Park	2	16.7	27.7	30	27	2	4	2
3	Victoria	Liverpool Street	2	23.2	34.1	43	36	1	3	3
4	Angel	Waterloo	3	23.5	33.9	7	34	2	2	5
5	Liverpool Street	Green Park	3	21.4	32.8	30	36	2	3	6
6	Euston	South Kensington	4	20.1	31.5	5	42	2	4	4
7	Victoria	Waterloo	5	15.6	24.0	4	21	2	3	5

4.7 Attribute cut-off based on generalised costs of routes

4.7.1 Definition of the attribute cut-off

As the result of the route choice set generation algorithm (cf. **Section 4.6**) a set of shortest routes (K_{ij}^{gen}) were generated for each OD pair ij, where in each set, there are $N_{K,ij}^{gen}$ number of generated routes. Now, the challenge is to find the set of reasonable routes (K_{ij}) among them (cf. formulae (4-3) and (4-4)).

The aim here is to be able to decide for each route in the set of generated routes ($k, ij \in K_{ij}^{gen}$), whether they are also in the set of reasonable routes ($k, ij \in K_{ij}^{gen}$). For this, the main data source is the metro network ($G(P, A)$) with its link times (t_a) (cf. **Section 4.4**); however it is also discussed, whether the additional information from smart card data (OJT, cf. **Section 3.6.1**) could provide a better understanding.

Attribute cut-off (ρ) is defined; so that a given route k can be considered reasonable ($k \in K_{ij}$), if the certain route and OD attributes – function of the metro network properties and other data sources (e.g. smart card) ($f(G(P, A), OJT)$) – are under that limit:

$$k \in K_{ij}, \qquad if\ f(G(P, A), OJT) \leq \rho \tag{4-8}$$

4.7.2 Selection of the attribute cut-off method

Through the literature review in **Section 4.2.2** and through making trials with those approaches it was understood that it is favourable to use a cut-off rule, which considers both journey time and interchange attributes. For that purpose in equation (4-2) the generalised cost of routes was formulated and the corresponding weights of the attributes were adapted from LU specific studies (cf. **Section 4.5**). Following this logical stream it was chosen to use the generalised costs of routes as attribute cut-offs.

Regarding the use of OJTs for attribute cut-offs it is important to consider its sample size. As the OJT sample provided for this research is quite small (cf. **Section 3.6.1**), it would not be representative, because the maximum OJT understood from the data sample may not reflect truly the maximum journey time that could be used for the attribute cut-off. Therefore, in this research attribute cut-offs were set based only on the generalised costs of routes, without considering the additional information on the OJT distribution.

Zhu and Xu (2016) defined attribute cut-off both in function of proportion and absolute difference in generalised costs. These considerations are useful, when the OD pairs of the network are of a significantly different scale. Regarding the case study OD pairs, as all of them are within the LU inner zone network (cf. **Section 4.4.5**), they could be considered of similar scale in terms of their minimum journey time (t_{ij}^{min}), which is between 16 and 23 minutes (cf. **Table 4-7**). Therefore in this case, working with proportions or with absolute differences would not give significantly different results. For the easier comparison, it was selected to work with proportions of generalised costs ($\rho c_{k,ij}$, cf. equation (4-7)).

The work in this section builds on Nádudvari et al. (2016), with the difference that, there the attribute cut-off was examined separately for each journey time component.

4.7.3 Attribute cut-off based on proportions of generalised costs

Setting the attribute cut-off in terms of proportions of generalised costs (ρc) the general definition (formula (4-8)) can be written as:

$$k, ij \in K_{ij}, \qquad if\ \rho c_{k,ij} \leq \rho c \tag{4-9}$$

The generalised cost proportion of each route ($\rho c_{k,ij}$) is calculated with formula (4-7).

To obtain the attribute cut-off (ρc) it is necessary to have information on the observed route choice set of passengers (K_{ij}^{obs}), and make ρc, so that the results for the reasonable route choice set (K_{ij}) could reproduce that set. For K_{ij}^{obs}, the route choice observations from RODS data (cf. **Section 3.6.3**) were used. The observed routes were highlighted in green among the results of the shortest routes (**Table 4-6**, **Table D-1**, **Table D-2**, **Table D-3**, **Table D-4**, **Table D-5** and **Table D-6**).

The generalised cost proportion of the longest route among the observed routes can be written as:

$$\rho c_{ij}^{max,obs} = max\left(\rho c_{k,ij} | k, ij \in K_{ij}^{obs}\right) \tag{4-10}$$

Similarly, the generalised cost proportion of the shortest route among the unobserved routes is:

$$\rho c_{ij}^{min,unobs} = min\left(\rho c_{k,ij} | k, ij \in K_{ij}^{gen} \backslash K_{ij}^{obs}\right) \tag{4-11}$$

Having calculated $\rho c_{ij}^{max,obs}$ and $\rho c_{ij}^{min,unobs}$ as confining values; it is expected that for most OD pairs ij, ρc will be between these limits.

$$\rho c_{ij}^{max,obs} \leq \rho c \leq \rho c_{ij}^{min,unobs} \tag{4-12}$$

In principle, one could have $\rho c_{ij}^{max,obs} \geq \rho c_{ij}^{min,unobs}$ (i.e. observed routes with higher generalised costs than the unobserved routes), however in this model specification this would not likely occur. The reason for this is that the weights of the generalised cost equation (4-2) are taken from Raveau et al. (2014), which is calibrated with RODS data; therefore when the routes are generated based on their generalised costs, they appear in the order that the observed routes from the same dataset (i.e. RODS data) have lower generalised costs than the unobserved routes.

In order to fulfil the conditions of equation (4-12) , a possible match for most of the ij, could be to use the method of least squares, where the objective function is:

$$min\left(\sum_{ij}^{N_{OD}}\left(\rho c - \rho c_{ij}^{max,obs}\right)^2 + \left(\rho c - \rho c_{ij}^{min,unobs}\right)^2\right) \tag{4-13}$$

This can be solved and simplified as

$$\rho c = \frac{\sum_{ij}^{N_{OD}}\left(\rho c_{ij}^{max,obs} + \rho c_{ij}^{min,unobs}\right)}{2 \cdot N_{OD}} \tag{4-14}$$

4.7.4 Classification of OD pairs based on case study results

Figure 4-10 presents the generalised cost proportions ($\rho c_{k,ij}$) for each generated route ($k \in K_{ij}^{gen}$) of all OD pairs ij (cf. **Table 4-3**). The values in the observed route choice set ($k \in K_{ij}^{obs}$) are labelled with green filled circles (●), while those in the unobserved route choice set ($k \in K_{ij}^{gen} \backslash K_{ij}^{obs}$) with red cross (**x**). Applying equation (4-14), $\rho c = 1.18$ was obtained, which means that a route k is reasonable, if its generalised cost ($c_{k,ij}$) is less or equal than 1.18 times than the generalised cost of the shortest route. ($c_{1,ij}$) (cf. equation (4-9)). The value of ρc is labelled with blue vertical line (**I**).

The attribute cut-off results (ρc) can reproduce the observed route choice set (K_{ij}^{obs}), when both of the following conditions hold. The generalised cost proportion is below the cut-off value for all observed routes:

$$\rho c_{k,ij} \leq \rho c \;\; for\ all\ k \in K_{ij}^{obs} \tag{4-15}$$

and it is above the cut-off value for all unobserved routes:

$$\rho c_{k,ij} > \rho c \; for \; all \; k \in K_{ij}^{gen} \backslash K_{ij}^{obs} \tag{4-16}$$

Based on these conditions, the case study OD pairs can fall into one of the following categories:

- Attribute cut-off results reproduce the set of observed routes (both conditions (4-15) and (4-16) hold)
- Unobserved routes are below the cut-off value (condition (4-16) does not hold)
- Observed routes are above the cut off value (condition (4-15) does not hold)

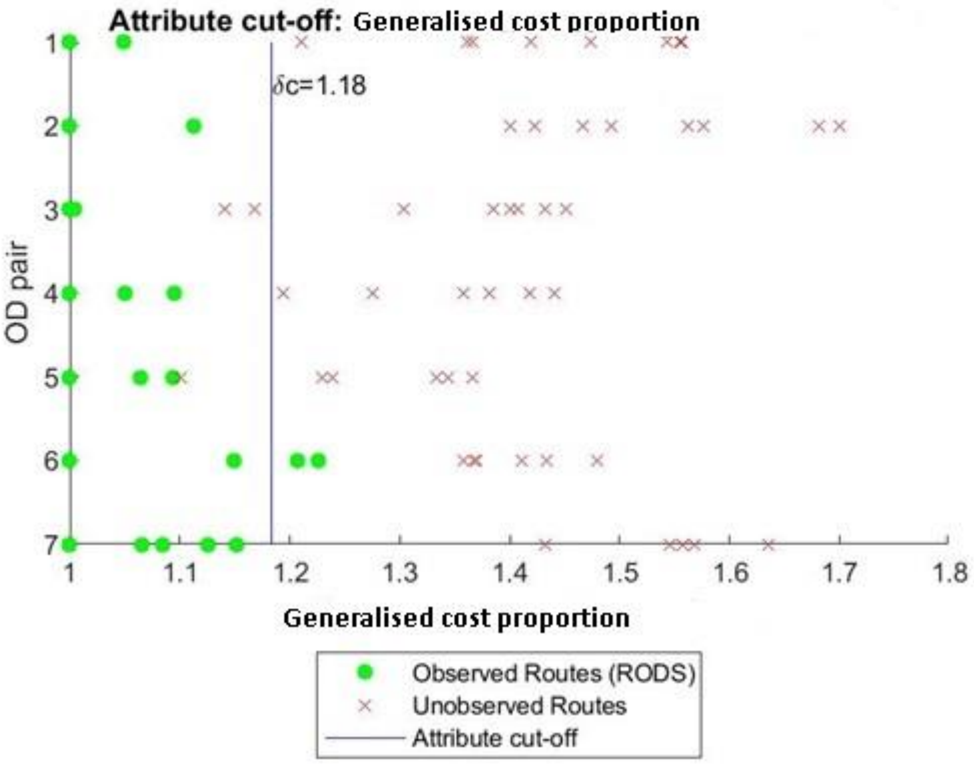


Figure 4-10 Attribute cut-off according to generalised cost proportions

4.7.4.1 Attribute cut-off results reproducing the set of observed routes

For OD pairs $ij = \{1, 2, 4, 7\}$ (cf. **Table 4-7**) both conditions ((4-15) and (4-16)) hold. For each of them, all observed routes (K_{ij}^{obs}) have two journey legs (**Figure 4-9**, **Table 4-6**, Table D-1, **Table D-3** and **Table D-6**). Those routes, which are not observed ($K_{ij}^{gen} \backslash K_{ij}^{obs}$) either have 3 or more journey legs or they make an excessive detour (mentioned as turning back and turning away in Dial (1971)). Although the journey time and interchange properties of the routes across the OD pairs differ, the same attribute cut-off value $\rho c =$ 1.18 could work for them.

4.7.4.2 Unobserved routes below cut-off value

For OD pairs $ij = \{3, 5\}$ (cf. **Table 4-7**) condition (4-16) does not hold, which means that there are unobserved routes ($K_{ij}^{gen} \backslash K_{ij}^{obs}$), with a lower generalised cost proportion ($\rho c_{k,ij}$) than the attribute cut off value (ρc). The reason for this could be found in the specific properties of these OD pairs

- Presence of direct routes
- Number of available directions at origin and destination station

Presence of direct routes

OD pair $ij = 3$ (**Victoria** – **Liverpool Street**, cf. **Table D-2**) has a direct route (i.e. **Circle** line, cf. **Table 4-7**). Having a direct route is so attractive to passengers, so that they would consider indirect routes, only if they are much better in other attributes. In this particular example, apart from the direct route there is also an indirect route (**Victoria** – **Central** via **Oxford Circus**) in the observed set (K_{ij}^{obs}). This route is attractive, because its total journey time ($t_{k,ij}^{SJT}$) is 5.3 minutes shorter. Furthermore, both the **Victoria** and **Central** lines are very frequent services (with 2 minutes of headway), while the circle line is an infrequent service (with 10 minutes of headway). Therefore, neither of the alternative routes dominate each other, which is also expressed in the similarity of their generalised costs.

Looking at the third (**Victoria** – **Circle** via **King's Cross**) and fourth shortest route (**Circle** – **Central** via **Bank**), it can be observed, that even though $\rho c_{k,ij}$ is only 1.14 and 1.17 minutes respectively (cf. **Figure 4-10**), they are not in the observed set. This is because these indirect routes do not have any attributes in which they dominate the direct route: Their total journey time is similar to the direct route and it involves interchanges through large station complexes (i.e. **King's Cross** and **Bank** stations).

Through the results for OD pair $ij = 3$ (**Victoria** - **Liverpool Street**) the following was observed: If an OD pair has a direct route, ρc is expected to be lower than for those OD pairs which only have routes with two or more journey legs. Therefore, ρc is not only a function of route attributes, but also of the property of the OD pair (i.e. presence of direct routes).

Number of available directions at the origin and destination station

OD pair $ij = 5$ (**Liverpool Street** – **Green Park**, cf. **Table D-4)** has many routes with two journey legs (i.e. route $k = \{1, 2, 3, 4, 5, 6, 10\}$) and for the shortest unobserved route (**Circle** – **Victoria** via **King's Cross**) the generalised cost proportion is still not that high ($\rho c_{ij}^{min,unobs} = 1.10$).

This could be explained with the many available directions (cf. **Table 4-7**) at the origin (i.e. 3: **Central** line westbound, and **Circle** line in both directions at **Liverpool Street** station) as well as at the destination station (i.e. 6: **Victoria**, **Jubilee** and **Piccadilly** lines from both directions at **Green Park** station). Due to the high number of available lines, there are many route options with two journey legs, which have similar generalised costs. The relationship between the number of available directions at the origin and/or destination station and the number of observed routes was also discussed in Guo (2008).

In this case a lower ρc value is expected than in other cases. Therefore ρc is not only a function of route attributes, but also of the property of the OD pair (i.e. number of directions available at origin and destination station).

4.7.4.3 Observed routes above the cut-off value

For OD pair $ij = 6$ (**Euston** – **South Kensington**, cf. **Table D-5**) condition (4-15) does not hold, where even though the third (**Northern (CX)** – **Piccadilly** via **Leicester Square**) and fourth shortest route (**Northern (CX)** – **Circle** via **Embankment**), are in the observed set (K_{ij}^{obs}), they have a $\delta c_{k,ij}$ value of 1.21 and 1.23 respectively, which is higher than the attribute cut-off ($\rho c = 1.18$).

4.8 Summary, discussion and proposed extension of the model

The purpose of this chapter was to address the issue that finite mixture models applied for route choice estimation require as an input the number of mixture components, which corresponds to the number of reasonable routes of an OD pair. For this, the question of route choice set generation was discussed. As the first step, a pathfinding method (i.e. K shortest path) was applied to find a set of shortest routes. In order to implement this algorithm for metro networks and to avoid that it gives route variants which differ only in their AEI movements within the stations, certain modifications were proposed to the algorithm. Following this, the attribute cut-off was set, based on the generalised cost proportions of routes to narrow down this set to the set of reasonable routes. The objective was to find the value, below which the set of routes can reproduce the observed route choice set for most of the OD pairs.

Results showed that the generalised cost proportion of 1.18 gives the best match for the seven case study OD pairs of the LU. This means, that a route is considered reasonable, if its generalised cost is not more than 1.18 times the generalised cost of the shortest route. This value is much lower than the results in Zhu and Xu (2016), which stated that in the Shanghai metro passengers consider a route reasonable up to 1.60 times the shortest route. This result could actually reproduce the observed route choice set for four out of the seven OD pairs. Two OD pairs there had unobserved routes with generalised costs below this cut-off value; and one OD pair had observed routes with generalised cost above the cut-off.

Based on this, it was understood that applying only a single attribute cut-off value cannot find the reasonable route choice set for all types of OD pairs, but it should be defined as a function of OD specific attributes. Among these OD specific attributes, two of them were highlighted through the case studies. One of these OD specific attributes was the presence of a direct route ($N_{L,ij}^{min}$). Results for the **Victoria** – **Liverpool Street** OD pair showed that if there is a direct route, passengers consider indirect routes only if they are dominant in other attributes (i.e. journey time, headway). Therefore, routes with generalised cost proportion of 1.14 or 1.17 were not in the observed set. The other OD specific attribute was the number of available directions at the origin ($N_{D,i}$) and the destination ($N_{D,j}$) station. Results for the **Liverpool Street** – **Green Park** OD pair showed that as there are many direction available at the destination station, there are many routes with two journey legs. Among them the route with the generalised cost proportion of 1.10 was not in the observed set.

As these OD specific attributes are proved to be important, the criteria set in formula (4-9) can be extended to the following:

$$k, ij \in K_{ij}, \qquad if\ \rho c_{k,ij} \leq \rho c\left(N_{L,ij}^{min}, N_{D,i}, N_{D,j}\right) \tag{4-17}$$

Formula (4-17) indicates that the attribute cut-off (ρc) is not a constant value valid for all OD pairs, but it is a function of the above discussed OD specific properties.

In order to obtain the actual function $\rho c\left(N_{L,ij}^{min}, N_{D,i}, N_{D,j}\right)$ it would be necessary to apply the method on more OD pairs of the LU network . As the program codes are already ready for all modelling steps discussed this chapter, it could be easily extended to apply it automatically for all OD pairs of the case study network. However, when applying for all OD pairs, there are other cases, which needs to be further examined. One of these cases are the routes with only one observed route. Learning those OD pairs can give a better information how to find the cut-off between the observed and unobserved routes. Additionally, there are also OD pairs, which have observed routes with three or more journey legs. It is expected that the attribute cut-off would be different also in those cases.

One possible limitation of the model applied here is that it considers only travel time components and interchange experience for the generalised costs, however it is acknowledged that also the perception of the map is an important attribute to consider, especially for the London Underground where the map is quite distorted (Guo, 2011).

Another limitation of the model is, that it used RODS data for the observed route choice set; and the weights, used in the generalised cost function also come from a calibration based on RODS data (Raveau et al., 2014). As it was expressed previously (cf. **Section 2.2.5**), one aim in this thesis is to move away from methods that uses results from manual surveys and to rely on automatically collected data sources. Therefore it could be further examined – if the results of the TfL WiFi survey (Transport for London, 2017) would be available – whether those could serve as a better source for validation.

An improvement of this route choice set generation model is further presented in **Chapter 5**, where the influence of additional OD specific attributes on the cut-off values is discussed. The purpose there is to identify those OD pairs, which have similar route choice patterns as well as to find the exact number of components for the applied finite mixture model.

Chapter 5
The superstation representation of metro networks to overcome data availability issues of station-to-station OD pairs

5.1 Introduction

As it was elucidated in **Section 3.8.2** one of the main limitations of finite mixture models to be applied for route choice estimation is the problem of data availability: While a large sample of smart card data is available for the whole network, for single station-to-station origin destination (OD) pairs this sample size is very few, therefore modellers often do not have sufficient data for their analysis. Through the case studies (cf. **Section 3.7**) it was further illustrated that when the finite mixture model is applied on a very small sample of Observed Journey Times (OJT) it either cannot find a solution (ill-conditioned covariance) or may converge to multiple possible solutions and the difference between these solutions is very large.

In case modellers can have access only to the open data sources, the question of data availability is even more crucial. For example in the context of the London Underground (LU) the open data contains only a 5% sample of Oyster cardholders for a 1 week period (cf. **Section 3.6.1**). Supposing that a larger sample of Oyster data could have been provided for this research from a longer period as a bespoke data – as it was in Fu (2014) (100% of data from a 40 week period) – temporal aggregation would have been a possible approach. While this could show success in overcoming the data availability issues, at the same time it loses the advantage that smart card data was intended to bring: time period specific estimates of route choice (cf. **Section 2.3**). Aggregating several months of data it is not possible to capture the day-to-day variation of travel patterns.

In this chapter the question of data aggregation is approached from another angle. Instead of working with data from longer time periods (i.e. temporal aggregation), it is explored how the data of OD pairs with similar properties can be aggregated (i.e. spatial aggregation). For this, firstly the origin and destination stations needs to be grouped according to certain rules. These groups of stations are called “superstations” throughout this thesis. Following this, for each superstation the centroids can be selected, so that the OJTs of each station-to-station OD pair can be adjusted there; and hence they could be aggregated, this way obtaining a larger sample of Centroid-to-Centroid OJTs (CCOJT) for the superstation-to-superstation OD pairs.

Applying the finite mixture model (**Chapter 3**) on the larger dataset of CCOJTs of superstation-to-superstation OD pairs, and comparing the results with those for station-to-station OD pairs (**Section 3.7**) can give an evaluation at what extent the superstation representation could overcome the previously mentioned data availability issues.

The rest of this chapter is structured as follows. In **Section 5.2**, earlier studies on station grouping are reviewed to understand for what purpose modellers define these larger network elements. Following this, focusing on station grouping for the purpose of route choice estimation the definition of superstations is presented together with their properties in **Section 5.3**. Once the superstations are created, in **Section 5.4** the methodology for adjusting OJTs of station-to-station OD pairs to superstation centroids is presented. **Section 5.5** presents the application of the mixture model on the CCOJTs of station-to-station OD pairs. These methods are illustrated through the case studies in **Section 5.6**. Finally, in **Section 5.7**, the benefits and the limitations of the superstation representation are summarised.

The concept of superstations and the application of finite mixture models for superstation-to-superstation OD pairs were initially presented in Nádudvari et al. (2015). This chapter brings forward the original idea, giving a more detailed formulation and examining more adequate case studies for passenger route choice.

5.2 Existing approaches for grouping stations

In the general transport modelling context zones are defined, when the inclusion of each basic network element (e.g. household, junction, stop, station) would require a too detailed, hence computationally expensive network model (Ortúzar and Willumsen, 2011; Connors and Watling, 2014). In these cases a zoning scheme is developed to aggregate these basic network elements into larger entities, such as traffic analysis zones or statistical/administrative wards.

For public transport networks, these larger entities would correspond to groups of stations and/or stops[18]. It can be easily understood, that different research problems within the field of public transport modelling may require different rules for station grouping. Therefore, the literature review in this section is arranged according to the research area, for which the station grouping methodology was proposed for:

[18] In this literature review both the terms of “stations” and “stops” are used, the former refers to metro or rail stations, while the latter to bus or tram stops.

- Overcoming data availability issues for OD matrix estimation (see **Section 5.2.1**)
- Considering choice between nearby stations for OD matrix or mode choice estimation (see **Section 5.2.2)**
- Reducing network complexity (see **Section 5.2.3**)

These three problems are not completely distinct research areas, there is overlap among them. The purpose for this literature review is to identify, whether any of these methods can be applied for the research problem addressed in this thesis (see **Section 5.2.4**).

5.2.1 Overcoming data availability issues for OD matrix estimation

The issue of data availability for station-to-station OD pairs was also mentioned in Cui (2006) and applied for the London bus network. He interpreted this problem as the objective to achieve a balance between accuracy and processing practicality. He introduced the concept of segments and aggregated the smart card transactions along them. He defined segments in the following way (**Figure 5-1**):

a) An interchange station (D) is a segment by itself.
b) Stations between two consecutive interchange stations (E and F) are defined as a segment.
c) Stations between the terminus and first interchange station (A, B and C) are defined as a segment.

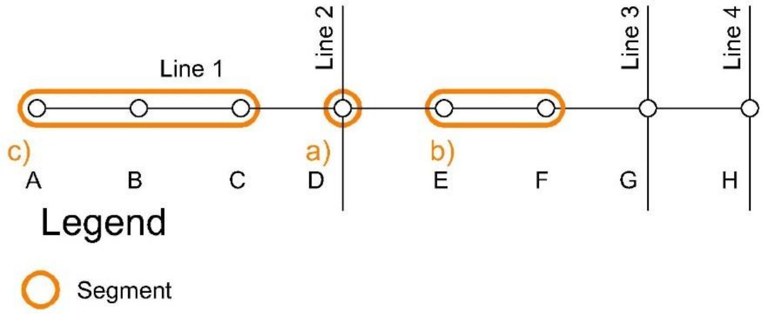


Figure 5-1 Concept of segments, based on Cui (2006)

Cui (2006) proposed the concept of segments for aggregating smart card transactions for OD demand matrices. This analogy is not fully applicable in the context of route choice estimation, for the following reasons:

a) If there are no attractive connections at an interchange station (D) towards the destination (Z); the route choice patterns may be similar from the segments before (A, B and C) and after the interchange station (E and F) as well as from the segment of the interchange station itself (D) (see **Figure 5-2**).

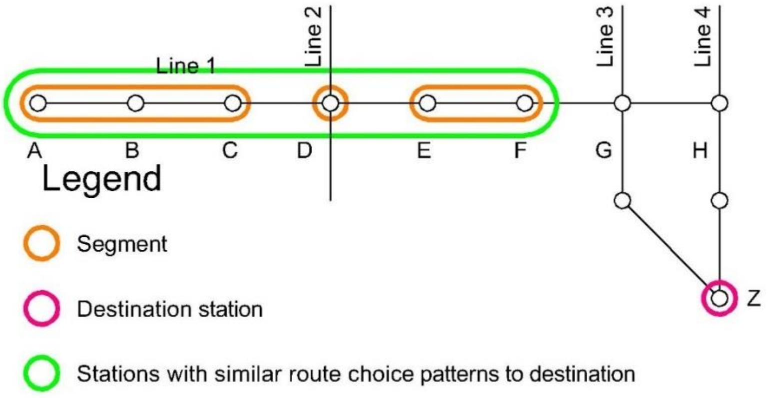


Figure 5-2 Segments with similar route choice patterns towards the destination

b) Given OD pairs (E-Y and F-Y), where the first journey leg of the routes are on the same line, but in the opposite direction (1D2 and 1G3); the route choice patterns from the stations of one segment (E and F) are different (see **Figure 5-3**).

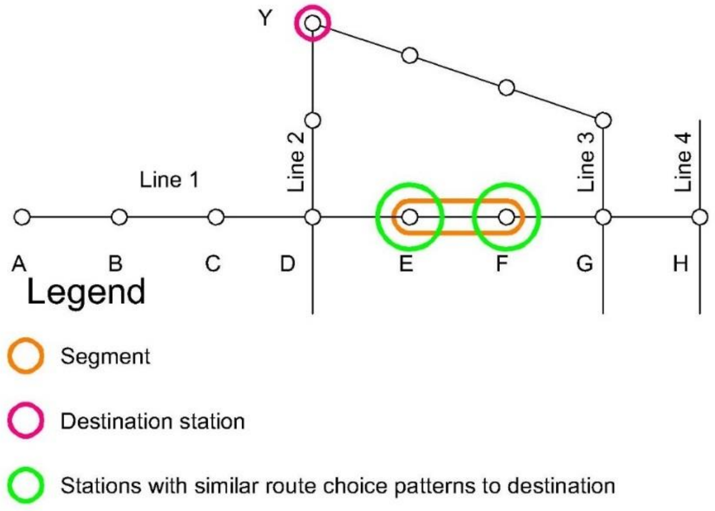


Figure 5-3 Different route choice patterns towards the destination within a segment

5.2.2 Considering choice among nearby stations for OD demand matrix or mode choice estimation

Another important and well explored research area for station grouping is to identify the set of attractive entry/exit stations (boarding/alighting stops) near the true origin/destination of the passenger. It can be understood, that the entry/exit station to a public transport (i.e. metro, bus) network does not necessarily reflect the true origin/destination of the passenger as he/she may have accessed this stop by other transport modes (e.g. walk, bike, car, taxi). This type of station grouping is generally applied to give a clearer picture on the OD demand matrix and mode choice of passengers.

Identifying stops in the catchment area of activities

The most straightforward station grouping approach is to associate them with the catchment area of the points of activities or transport hubs (see **Table 5-1**). Chu and Chapleau (2010) – working on the bus network of Gatineau and Ottawa, Canada – called these points of activities as anchor points. They represent places that a person repeatedly visits in short term (e.g. home, work, study) or long term (e.g. place of worship, visiting friends). They were identified from multiday smart card records in the following way: In case of student cardholders, they looked at boarding records after the end of the teaching and found the corresponding educational establishment from a georeferenced database. For other trips (e.g. home and work based trips), they envisioned a kernel density analysis to associate the range of activity locations with a probability. Following this, trip ends (i.e. first/last boarding/alighting stops) were linked to these anchor points if they are located within 500 m of the anchor point.

Similarly, Lee et al. (2013) – focusing on the Minneapolis-St. Paul metropolitan area, USA – grouped those stops, which have 50 m distance between each other as well as those, which have identical or similar stop name. Furthermore, they also examined the special cases, when there is a stop only in one direction of the bus line and found the matching stop in the opposite direction. The purpose for station grouping in this case was to understand transit demand at an aggregate level and land use patterns.

Table 5-1 Review of station grouping methods based on physical proximity – catchment area of activities

Reference	Purpose	Method	Mode	Case study
Chu and Chapleau (2010)	OD demand Travel behaviour	Trip end within 500 m of anchor points	bus	Gatineau, Ottawa, Canada
Lee et al. (2013)	OD demand Land use	Stops 50 m to each other Similar name Opposite direction	bus	Minneapolis-St. Paul, USA

Clustering algorithms

A more advanced approach for station grouping is to apply clustering algorithms (see **Table 5-2**).

Kieu et al. (2015b) – using the public transport network of Brisbane, Australia as a case study – applied the Density-Based Spatial Clustering of Application with Noise algorithm (Ester et al., 1996) for grouping the last alighting stops and then the first boarding stops of public transport trips known from smart card data. This algorithm uses two parameters to distinguish the least dense cluster of stations from the noise: the maximum density reach distance and the minimum number of points. This was further improved in Kieu et al. (2015a) to reduce the time complexity of the algorithm and called it as the Weighted-Stop Density-Based Spatial Clustering of Application with Noise.

Viggiano et al. (2016) – analysing London's multimodal public transport network – grouped the nearby stations and stops with the purpose to gain a better understanding on public transport mode choice (i.e. rail and bus). They called these group of stations as zones. They set the number of zones to 1000 and used the K-means clustering algorithm (Forgy, 1965; MacQueen, 1967) to allocate each stop and station in these clusters. In their model rail stations were weighted 10 times as much as bus stops.

Similarly, Luo et al. (2017) – doing the case study on the public transport network of The Hague, Netherlands – applied a K-means based station aggregation method for the purpose of obtaining transit OD demand matrices at a zonal level. They tested the K-means clustering algorithm for a range of cluster numbers (between 2 and 30) and for each value they calculated the spatial distance and passenger flow related metrics to find the optimum value (12 for the case study area).

Table 5-2 Review of station grouping methods based on physical proximity – clustering algorithms

Reference	Purpose	Method	Mode	Case study
Kieu et al. (2015b)	Transit passenger market segmentation	Density-Based Spatial Clustering of Application with Noise	Bus, rail, ferry	Brisbane, Australia
Viggiano et al. (2016)	Public transport mode choice	K-means clustering with 1000 clusters	Bus, metro, rail	London, UK
Luo et al. (2017)	OD demand at zonal level	K-means based station aggregation, trials with 2-30 clusters	Bus, Tram	The Hague, Netherlands

Logit allocation models

Another advanced approach for station grouping is to use logit allocation models, which means to estimate for each station the probability that it belongs to a certain zone. This approach is especially useful, when the station groups have to coincide with the existing zoning system (see **Table 5-3**).

Kuhlman (2015) – focusing on the public transport network of Amsterdam, Netherlands – applied a logit allocation model with the objective to construct purpose-specific OD demand matrices. Using smart card data he identified the trip ends (i.e. first boarding and last alighting stops of trips) and for each trip end, he estimated the probability with the Multinomial Logit (MNL) model that it belongs to a predefined traffic analysis zone. In this model specification, the alternatives are the traffic analysis zones nearby the trip ends and the attributes of the utility functions are the share of the catchment area, stop density and urbanisation level.

Tamblay et al. (2016) proposed the grouping of stations and stops for the purpose of developing a public transport planning computational tool for Santiago, Chile. They estimated the probabilities that the boarding/alighting stop of an observed trip from smart card data has its true origin/destination in a predefined census zone. They used a disaggregated logit model with the attributes of the access/egress times between the zone centroid and the given stop. This method was further developed in Tamblay et al. (2018).

Table 5-3 Review of station grouping methods based on physical proximity – logit allocation models

Reference	Purpose	Method	Mode	Case study
Kuhlman (2015)	Purpose-specific OD demand matrices	Probability that a trip end belongs to a traffic analysis zone (MNL)	Bus, tram, metro	Amsterdam, Netherlands
Tamblay et al. (2016)	Public transport planning tool	Probability that a trip end has its true origin/destination in a census zone	Bus, metro	Santiago, Chile
Young and Blainey (2017)	Improve catchment area representation of rail stations	MNL, mixed logit	rail	Wales, Scotland, UK

Young and Blainey (2017) focused on railway station choice in Wales and Scotland for the purpose of improving the representation of catchment areas of railway stations. For each origin/destination (i.e. postcode), they defined the choice set of the 10 nearest railway stations, ensuring that the major railway station is also included in that set. Following this, they estimated the railway station choice both with MNL and with mixed logit model, in which – apart from the access journey characteristics – they included also the attributes of the station facilities, service frequency and train journey.

Other approaches

Additionally, there are various other approaches in literature for station grouping according to physical proximity (see **Table 5-4**). Nassir et al. (2015b) – working on the public transport network of Brisbane, Australia – modelled the boarding stop choice set of a passenger in light of his/her route choice set between the true origin and destination. They used smart card data together with the information on the public transport and walkway network and determined a set of shortest routes with the K shortest path algorithm (Yen, 1971). Following this, they narrowed down this set based the criteria on the maximum acceptable access, egress (2 km) and interchange (1 km) distance and wait time (1 hour). Additionally, they also set a threshold for the travel time of a route and for the maximum number of interchanges (3). In this setting, the boarding choice set corresponds to the first boarding stops of the routes included in the route choice set. This method was further applied in (Hassan et al., 2016) for understanding passengers public transport stop choice behaviour and in (Nassir et al., 2016) to define a utility-based travel-impedance measure for public transport network accessibility.

Guo and Lu (2016) – focusing on the London Underground – applied the concept of neighbourhood centrality (Opsahl et al., 2010) to define the neighbourhoods that are centred in the statistical/administrative wards of Greater London. In their study, distances and path lengths between two stations correspond to the number of intermediate stations. The purpose of this study was to relate complex network properties to human geographical features in the city, such as age demographics, mode choice and housing.

Table 5-4 Review of station grouping methods based on physical proximity – other approaches

Reference	Purpose	Method	Mode	Case study
Nassir et al. (2015b)	Public transport network accessibility	K shortest path, narrow down set: max walk distance and wait time	Bus, rail, ferry	Brisbane, Australia
Guo and Lu (2016)	Relate network properties to age, mode choice and housing	Neighbourhood centrality; distances as number of stations	metro	London, UK

Further applications

The concept of working with groups of nearby stations have been also applied in the field of transport hub location problem (see **Table 5-5**). To address this issue, Yu et al. (2013) proposed a two-phase optimisation approach and applied for the Dalian, China. In the first phase, candidate nodes are selected among all stops based on passenger attraction, which is the function of the accessibility and connectivity of the stop. Following this, in the second phase, a location model is applied on the candidate nodes to find the optimal hub location among them, based on its largest serviced population, minimum overlap and least construction cost.

Furthermore, station grouping was applied not only for planning new hubs, but also for detecting the dynamics of urban structure. In this context, Zhong et al. (2014) analysed the smart card data of Singapore from 3 consecutive years (2010-2012) and constructed a weighted directed graph for each year. In this graph, nodes corresponded to urban areas, links to the possibility to travel between these areas and the weight of links to the volumes of travel. This graph was used to gain a better view on the travel demand, urban centres, transport hubs, neighbourhoods and borders; as well as on their dynamics over the years.

Table 5-5 Review of station grouping methods based on physical proximity – further applications

Reference	Purpose	Method	Mode	Case study
Yu et al. (2013)	Transport hub location	Candidate nodes: accessibility and connectivity, hub location model	bus	Dalian, China
Zhong et al. (2014)	Detecting the dynamics of urban structure	Weighted directed graph for 3 consecutive years	Bus, metro	Singapore

Applicability for the research problem

The main limitation of applying the previously presented station grouping concept for the research problem of this thesis can be illustrated by the following (see **Figure 5-4**): As for entry/exit station choice the requirement is, that the candidate stations are in physical proximity, in most cases these station are on different lines (C, D, E, J, K and L). In contrast, the purpose in this thesis is to group the stations according to similar route choice patterns, which requires, that they should be on the same line (A, B, C, D, E and F).

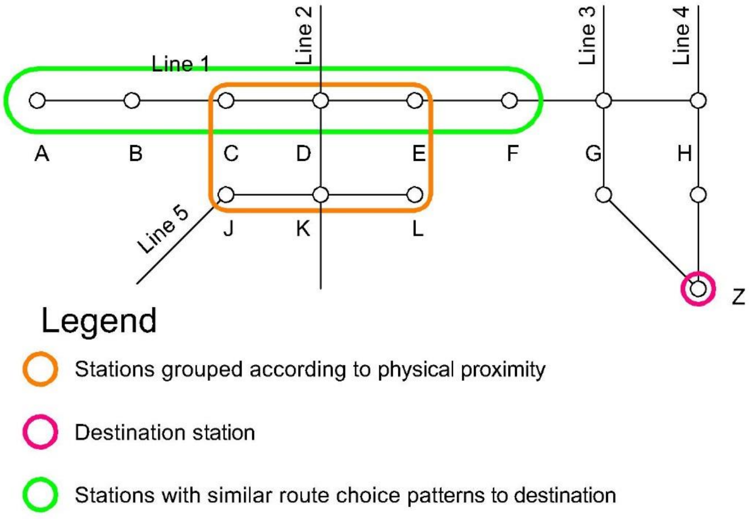

Line 1
Line 2
Line 3
Line 4
Line 5
A
B
C
D
E
F
G
H
J
K
L
Z
Legend
Stations grouped according to physical proximity
Destination station
Stations with similar route choice patterns to destination

Figure 5-4 Difference between grouping stations according to physical proximity and similar route choice patterns towards the destination

5.2.3 Reducing network complexity

The question of reducing the number of nodes and hence computational time has been widely explored in the context road traffic assignment and mentioned as "network aggregation" (Connors and Watling, 2014) or "network contraction" (Jafari and Boyles, 2016).

In the context of public transport networks, reducing the number of nodes can be achieved by defining network entities that can represent a group of stations with similar properties. Schmöcker (2006) developed a transit assignment model for the London Underground inner zone network. In order to represent the demand coming in the network from the outer zones, he defined line specific stations at the end of each LU line (cf. **Section 4.4.2**).

The concept of line specific stations is in connection with the purpose of this thesis; however the definition of station groups cannot be just limited to the question of inner and outer zones (see **Figure 5-5**). This is because for certain OD pairs with their origin in the outer zones (D-Z), passengers may have more reasonable routes (1H4 and 2M6) to enter in the LU inner zone via different lines (1 and 6).

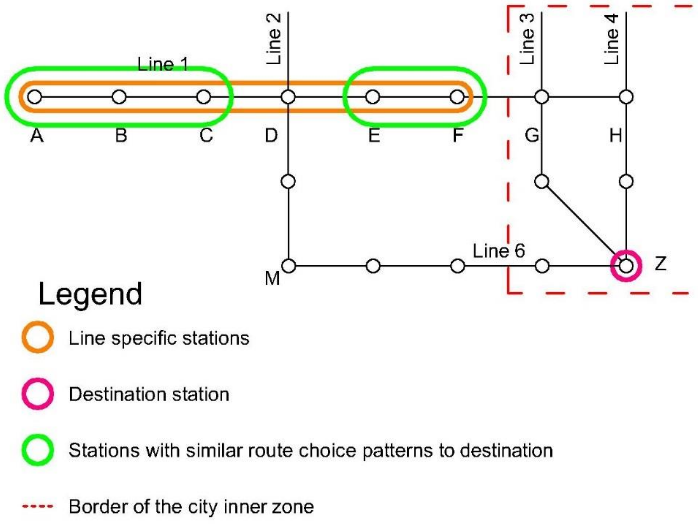

Line 2
Line 3
Line 4
Line 1
A
B
C
D
E
F
G
H
Line 6
Z
M
Legend
Line specific stations
Destination station
Stations with similar route choice patterns to destination
Border of the city inner zone

Figure 5-5 Difference between grouping stations as line specific stations (Schmöcker, 2006) and according to similar route choice patterns towards the destination

5.2.4 Gap in research

Having reviewed the literature on various studies for station grouping, it was established; that as they were applied for different modelling purposes, their concept cannot be directly implemented for the objective of this thesis: aggregating data of station-to-station OD pairs for better route choice estimates. As if follows, a new approach is proposed for grouping stations according to similar route choice patterns, by setting the definition and rules; as well as describing the method for finding those stations (see **Section 5.3**).

5.3 The concept of superstations

In this chapter – in addition to what was presented in the previous chapters – the following notation is used:

Variable identifiers

I, J Index of origin and destination superstation

Ii i-th station of origin superstation I

Jj j-th station of destination superstation J

Ic Centroid station of origin superstation I

Jc Centroid station of destination superstation J

Variables

$CCOJT$ The adjusted value of Observed Journey Times to superstation centroids (minutes)

N_I Number of stations in origin superstation I

N_J Number of stations in destination superstation J

n_{IJ}^{CCOJT} Sample size of $CCOJT_{IJ}$

$CCOJT_{r,IJ}^{KMS}$ Subset of $CCOJT_{IJ}$ produced by the K-means clustering algorithm (minutes)

5.3.1 Definition of superstations

In order to overcome the previously described data availability issues of station-to-station OD pairs for route choice estimation in metro networks (cf. **Section 5.1**), the concept of superstations is introduced:

Definition: *A group of stations on the same line from/to which passengers are expected to have the same route choice set and similar route choice probabilities*

Strictly speaking, according to the earlier definition of routes (3-13), the route choice set is not exactly the same for the candidate stations of the origin and destination superstation as they differ in their access, wait and egress journey segments, as well as in their the on-board segments from origin to interchange station and from interchange to destination station. Therefore, in a broader sense, for the sake of superstation definition, a route can be interpreted as the sequence of the following segments: line 1 – interchange 1 – … – interchange $N_{S,k}$ – line $N_{L,k}$. The adjustments according to the differences in these journey segments is discussed in **Section 5.4.2**.

5.3.2 Properties of superstations

In **Section 5.2.1**, the concept of segments (Cui, 2006) was presented and it was pointed out, that it cannot be directly applied for route choice estimation. Therefore, superstations are not identical to segments, due to the following properties:

Property 1: *Stations on the same segment can be grouped as a superstation only if all routes depart/arrive to/from the same direction on that line* (cf. **Figure 5-3**)

Property 2: *Multiple segments can be grouped as a superstation, if passengers from/to the stations of those segments are expected to have the same route choice set and similar route choice probabilities* (cf. **Figure 5-2**)

In the light of **Property 2**, it may occur that passengers have the same route choice set and similar route probabilities from stations before and after an interchange station, but from the interchange station itself the route choice set is different, because passengers may find there other attractive lines. To account for this, superstations have an additional property:

Property 3: *Superstations can include non-consecutive stations* (cf. **Figure 5-5**)

Furthermore, as some lines have short runs (see **Figure 5-6**), they have the following property:

Property 4: *Stations on the same line, but with different service frequency (short runs) can be included in the superstation, but adjustments need to be made according to the difference in wait time*

These adjustments are explained in **Section 5.4**.

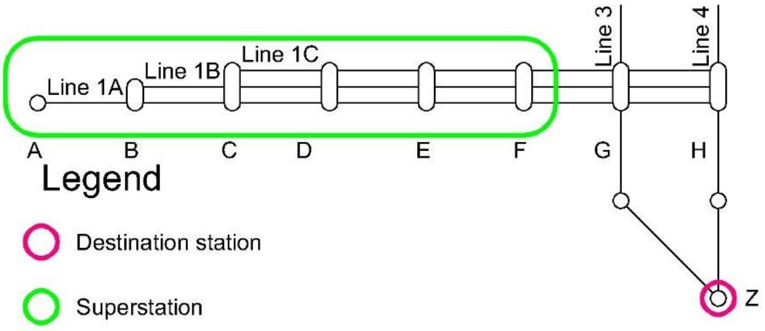


Figure 5-6 Property to create superstations in case of lines with short runs

5.3.3 Finding OD pairs with similar route choice patterns

According to the **Definition** (cf. **Section 5.3.1**), superstations are group of stations from/to which passengers are expected to have the same route choice set. This implies, that in order to find the stations of the origin and destination superstation, it is necessary to search for those OD pairs which have the same route choice set. To perform this, a route choice set generation algorithm needs to be applied on the extended LU inner zone network (see **Figure 5-8**) for the OD pairs composed by the candidate stations of the origin and destination superstation.

The results of the previously proposed method (cf. **Chapter 4**) showed that – in average – a route is considered reasonable if its generalised cost is less or equal than 1.18 times the generalised cost of the shortest route ($\rho c = 1.18$). However, it was also observed that this cut-off value did not work for all types OD pairs. Therefore it was concluded that ρc is also a function of OD specific attributes, such as:

- Presence of a direct route
- Number of available directions at the origin and destination station

Following this logical stream, it was understood that is not possible to use a single ρc value for all the case study OD pairs of this chapter, but this should be determined in function of the OD specific attributes. In addition to the previously listed attributes, the characteristics of these OD pairs requires to consider the following:

- OD minimum travel time
- Presence of an express line

The relationship between the OD minimum travel time and the number of observed routes was also discussed in Guo (2008), where he categorised the stations according to their location within Greater London, such as “Central”, “North”, “South”, “East”, “West” and “Outside”.

In each case study (cf. **Section 5.6**) it is described how these OD specific attributes influence the cut-off value. The explicit formulation of the route choice set generation algorithm is not discussed here as it is beyond the scope of this thesis.

5.4 Adjustment of the Observed Journey Times to superstation centroids

Once the origin and destination superstations are defined (cf. **Section 5.3**), it is possible to adjust the OJTs of station-to-station OD pairs to superstation centroids and hence to aggregate them spatially. This way a larger sample of centroid-to-centroid OJTs can be obtained for superstation-to-superstation OD pairs, proposing a solution for the previously mentioned data availability issues (cf. **Section 3.8.2**). This section presents the methodology for the selection of superstation centroids (see **Section 5.4.1**) as well as for the OJT adjustment (see **Section 5.4.2**) and aggregation (see **Section 5.4.3**).

5.4.1 Selection of superstation centroids

In the process of OJT adjustment the first step is to select the centroid for the origin (Ic) and destination (Jc) superstation. This can be any station of the superstation at the modeller’s convenience, not necessarily the geometrical centroid (see **Appendix E**). As in this chapter the cases bring forward the ones mentioned in **Chapter 3**, it is convenient to choose those stations as centroid, which were the origin and destination stations there (cf. **Section 3.7**).

5.4.2 Adjustment of the Observed Journey Times to superstation centroids

It is necessary to adjust the OJTs due to the different journey time components of the different station-to-station OD pairs (cf. equation (3-13)). At the origin superstation the following times need to be considered:

- On board time from entry station to origin superstation centroid $(t^{ob}_{(Ii)(Ic)})$
- Difference between access time at origin superstation centroid (t^{acc}_{Ic}) and at entry station (t^{acc}_{Ii})

- Difference between the wait time at origin superstation centroid (t_{Ic}^{wait}) and at the entry station (t_{Ii}^{wait}) (in case of short runs, cf. **Property 4**)

Given the entry time stamp at station i of the origin superstation I (T_{Ii}^{entry}), it is possible to obtain the equivalent entry time stamp at the superstation centroid ($T_{(Ii)(Ic)}^{entry}$) with the following adjustment of the journey time components (see **Figure 5-7**):

$$T_{(Ii)(Ic)}^{entry} = T_{Ii}^{entry} + t_{Ii}^{acc} + t_{Ii}^{wait} + t_{(Ii)(Ic)}^{ob} - t_{Ic}^{wait} - t_{Ic}^{acc} \tag{5-1}$$

Substituting $\Delta t_{(Ii)(Ic)} = t_{(Ic)} - t_{(Ii)}$ for all time components, this will be:

$$T_{(Ii)(Ic)}^{entry} = T_{Ii}^{entry} + t_{(Ii)(Ic)}^{ob} - \Delta t_{(Ii)(Ic)}^{acc} - \Delta t_{(Ii)(Ic)}^{wait} \tag{5-2}$$

At the destination superstation the following times need to be considered:

- On board time from exit station to destination superstation centroid ($t_{(Jj)(Jc)}^{ob}$)
- Difference between egress time at destination superstation centroid (t_{Jc}^{egr}) and at the exit station (t_{Jj}^{egr})
- Difference between the wait time on the last journey leg for the service that brings to the destination superstation centroid (t_{Jc}^{wait}) and the one that brings to the exit station (t_{Jj}^{wait}) (in case of short runs, cf. **Property 4**)

Given the exit time stamp at station j of the destination superstation J (T_{Jj}^{exit}) it is possible to obtain the equivalent exit time at the superstation centroid ($T_{(Jj)(Jc)}^{exit}$) with the following adjustment of the journey time components[19] (see **Figure 5-7**):

$$T_{(Jj)(Jc)}^{exit} = T_{Jj}^{exit} - t_{Jj}^{wait} - t_{Jj}^{egr} + t_{(Jj)(Jc)}^{ob} + t_{Jc}^{wait} + t_{Jc}^{egr} \tag{5-3}$$

Substituting $\Delta t_{(Jj)(Jc)} = t_{(Jc)} - t_{(Jj)}$ for all time components, this will be:

$$T_{(Jj)(Jc)}^{exit} = T_{Jj}^{exit} + t_{(Jj)(Jc)}^{ob} + \Delta t_{(Jj)(Jc)}^{wait} + \Delta t_{(Jj)(Jc)}^{egr} \tag{5-4}$$

The Observed Journey Times are calculated as the time difference between the entry and exit time stamps (cf. equation (3-14)):

$$OJT_{(Ii)(Jj)} = T_{Jj}^{exit} - T_{Ii}^{entry} \tag{5-5}$$

[19] As discussed earlier, t_{Jj}^{wait} and t_{Jc}^{wait} are the wait times for the service that brings to the exit station Jj and Jc respectively. For the correct illustration these times should be drawn at the last interchange station; however here for the sake of simplicity, they were drawn to the exit station.

Following this analogy, the Centroid-to-Centroid adjusted OJTs (CCOJT) corresponding to each station-to-station OD pair can be calculated as:

$$CCOJT_{(Ii)(Jj)} = T^{exit}_{(Jj)(Jc)} - T^{entry}_{(Ii)(Ic)} \tag{5-6}$$

Substituting equations (5-2) and (5-4) into (5-6):

$$\begin{aligned} CCOJT_{(Ii)(Jj)} &= T^{exit}_{Jj} + t^{ob}_{(Jj)(Jc)} + \Delta t^{egr}_{(Jj)(Jc)} + \Delta t^{wait}_{(Jj)(Jc)} - T^{entry}_{Ii} \\ &\quad - t^{ob}_{(Ii)(Ic)} + \Delta t^{acc}_{(Ii)(Ic)} + \Delta t^{wait}_{(Ii)(Ic)} \end{aligned} \tag{5-7}$$

Substituting equation (5-5) into (5-7):

$$\begin{aligned} CCOJT_{(Ii)(Jj)} &= OJT_{(Ii)(Jj)} + t^{ob}_{(Jj)(Jc)} - t^{ob}_{(Ii)(Ic)} + \Delta t^{acc}_{(Ii)(Ic)} \\ &\quad + \Delta t^{egr}_{(Jj)(Jc)} + \Delta t^{wait}_{(Ii)(Ic)} + \Delta t^{wait}_{(Jj)(Jc)} \end{aligned} \tag{5-8}$$

Equation (5-8) is applied in the case studies (see **Section 5.6**). There the OJTs of the station-to-station OD pairs ($OJT_{(Ii)(Jj)}$) are known from Oyster data (cf. **Section 3.6.1**). The necessary adjustments for on-board ($t^{ob}_{(Ii)(Ic)}$ and $t^{ob}_{(Jj)(Jc)}$) and wait ($\Delta t^{wait}_{(Ii)(Ic)}$) times are calculated using LU timetables (cf. **Section 3.6.2.1**). The corresponding access ($\Delta t^{acc}_{(Ii)(Ic)}$) and egress ($\Delta t^{egr}_{(Jj)(Jc)}$) times were estimated based on station layouts known from the Nationwide Access Register (Direct Enquires) (cf. **Section 3.6.2.2**).

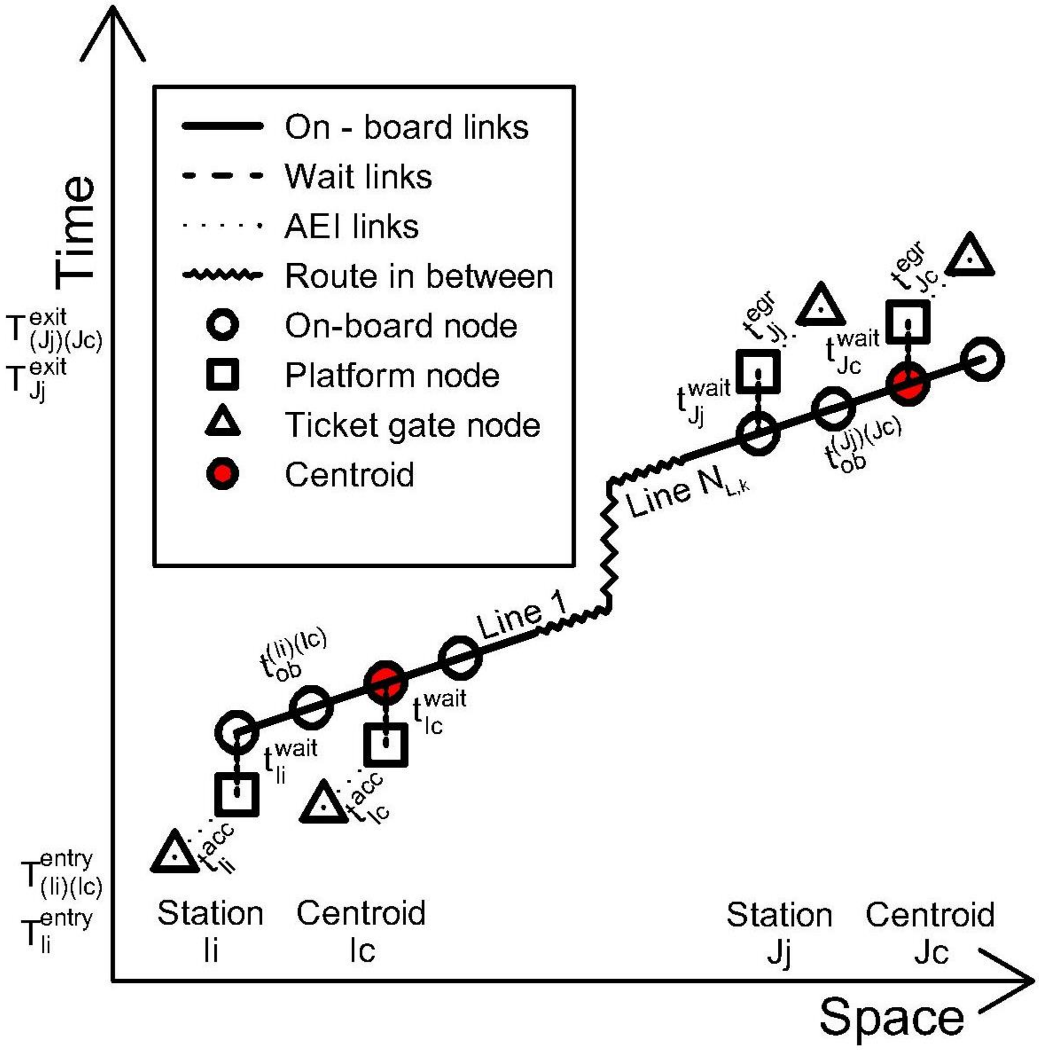


Figure 5-7 Adjustment of the Oyster entry/exit times at the origin/destination superstations, representation on a diachronic graph

5.4.3 Aggregation of the adjusted Observed Journey Times

Once the CCOJTs corresponding to each station-to-station OD pair ($CCOJT_{(Ii)(Jj)}$) are obtained, these values can be aggregated as they are adjusted to the same origin and destination superstation centroid. This aggregate dataset is called the CCOJT of the superstation-to-superstation OD pair ($CCOJT_{IJ}$):

$$CCOJT_{IJ} = \bigcup_{i=1}^{N_I} \bigcup_{j=1}^{N_J} CCOJT_{(Ii)(Jj)} \qquad (5\text{-}9)$$

where N_I and N_J is the number of stations in the origin (I) and destination (J) superstation respectively. As it follows, it is described how the finite mixture model is applied on this larger dataset with the purpose to evaluate the benefits of the superstation representation.

Section 5.4.1 stated that the centroid of a superstation can be any station, not necessarily the geometrical centroid. At this point, one may ask the question, whether the shape of the aggregated $CCOJT_{IJ}$ distribution would be different if the origin and/or the destination superstation centroid (Ic and Jc respectively) is chosen differently. This question is discussed in **Appendix E.**

5.5 Evaluation of the benefits of the superstation representation

Once the superstations are defined (cf. **Section 5.3**) and the CCOJTs for each superstation-to-superstation OD pair is calculated (cf. **Section 5.4**), the finite mixture model (cf. **Chapter 3**) is applied on this larger dataset. The purpose for this is to evaluate the benefits of the superstation representation by comparing the results of the finite mixture model applied on station-to-station and on the superstation-to-superstation dataset of the OJTs of the same case study OD pair (see **Section 5.6**).

Similarly to the methodology presented in **Section 3.3**, the finite mixture model was tested with different settings for the initial values (i.e. seeds for the random number generator) and tolerance thresholds to find the one, which gives the closest solution to the expected results (i.e. timetable, RODS). It was done so, because it is expected that among the solutions of the finite mixture model exists at least one, which reflects the actual values of the metro networks; although this may not necessarily be the global optimum. (cf. **Section 3.3.2**). Therefore, the tolerance thresholds chosen in this chapter are not always identical to the ones chosen in **Chapter 3**. Following that, the results of the finite mixture model were matched with the actual routes and compared to results of existing models (Fu, 2014) as described in **Section 3.4**.

5.6 Case studies on the London Underground

The cases in this section are the extension of the cases in **Section 3.7** (cf. **Figure 3-2** and **Table 3-2**) plus an additional case. Through the case studies presented here (see **Figure 5-8** and **Table 5-6**) the following methodologies are illustrated:

- Origin and/or destinations stations are grouped with other stations, from/to which passengers are expected to have similar route choice patterns (i.e. creating superstations, cf. **Section 5.3**).

- The OJTs of station-to-station OD pairs are adjusted to superstation centroids, so that they can be spatially aggregated (i.e. obtaining a larger dataset of CCOJTs, cf. **Section 5.4**).
- The finite mixture model (cf. **Chapter 3**) is applied on the CCOJTs of superstation-to-superstation OD pairs; and compared with the results of station-to-station OD pairs (i.e. evaluation of the superstation representation, cf. **Section 5.5**)

The previously described route choice set generation algorithm (**Chapter 4**) is applied on the extended LU inner zone network with the appropriate considerations for the OD specific attribute cut-offs (cf. **Section 5.3.3**). This extended LU inner zone network includes the LU inner zone network (cf. **Figure 4-6**), as well as some of the lines until their terminus in the LU outer zones (see **Figure 5-8**):

- **Bakerloo** line until **Elephant & Castle** (south end)
- **Central** line until **Epping** (east end)
- **Jubilee** line between **Stanmore** and **Stratford** (full length)
- **Northern** line until **Morden** (south end)
- **Victoria** line until **Brixton** (south end)

Additionally, while in the LU inner zone only the **Circle** line was considered among the common lines (cf. **Section 4.4.3**), in the extended LU inner zone also the **Metropolitan** line was included between **Wembley Park** and **Baker Street**.

In **Section 3.7** two case study OD pairs (**Case 1** and **Case 2**) were presented. The common in these two cases is that for all routes of these OD pairs the first journey leg is on the same line, therefore the origin stations can be grouped as superstations. In this chapter an additional case (**Case 3**) is presented, where both the first and the last journey leg is on the same line, therefore superstations can be created both for the origin and destination stations.

The actual difference between **Case 1** and **Case 2** – from the perspective of creating superstations – is that, while in **Case 1** there is only a short segment of the **Victoria** line before the first interchange station (**Green Park**) with 5 candidate stations for the origin superstation (**Brixton** – **Victoria**), that segment in **Case 2** is very long having 27 candidate stations on the **Central** line (**Epping** – **Chancery Lane**, including the **Hainault** loop).

In Nádudvari et al. (2015) the concept of superstations was illustrated through a different case study OD pair, which is not included in this thesis.

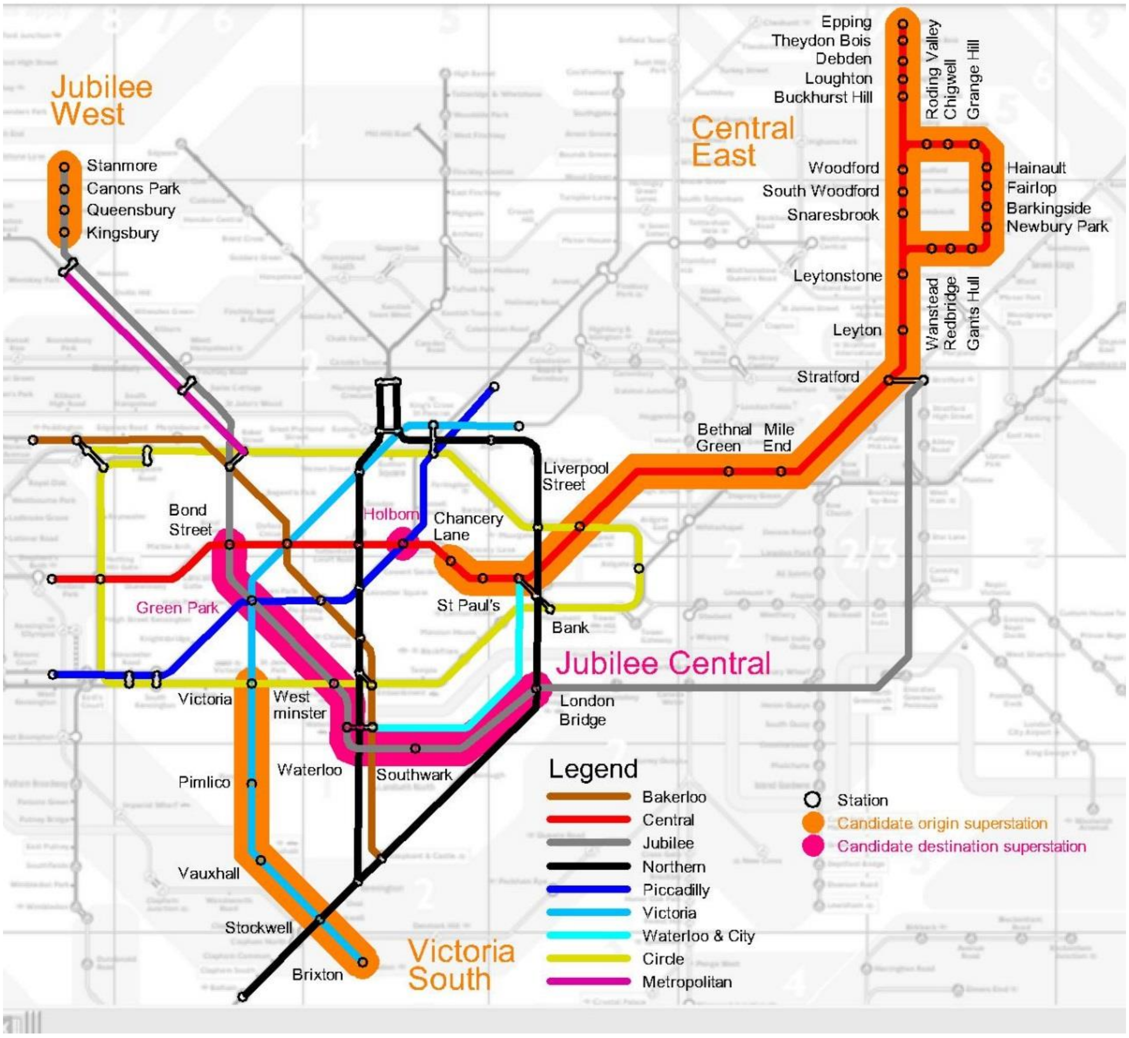


Figure 5-8 Overview of the case studies on the London Underground extended inner zone network

Table 5-6 A summary of the case studies on the LU network, superstation network representation

Case	OD pair		Route					Time	RODS		OJT
	Origin	Destination	Line 1	Interchange 1	Line 2	Interchange 2	Line 3	(min)	RC (%)	Sample	Sample
IJ	I	J	$l = 1$	$s = 1$	$l = 2$	$s = 2$	$l = 3$	$t^{SJT}_{k,IJ}$	$\omega^{RODS}_{k,IJ}$	n^{RODS}_{IJ}	n^{OJT}_{IJ}
1	Victoria South	Holborn	Vic (NB)	Oxford Circus	Cen (EB)			17.6	74.8%	1097	147
			Vic (NB)	Green Park	Pic (EB)			20.4	25.2%		
2	Central East	Green Park	Cen (WB)	Oxford Circus	Vic (SB)			21.3	81.2%	983	47
			Cen (WB)	Holborn	Pic (WB)			24.0	18.8%		
3	Jubilee West	Jubilee Central	Jub (EB)					36.3	89.0%	1196	286
			Jub (EB)	Wembley Park	Met (EB)	Finchley Road	Jub (EB)	33.3	11.0%		

Case 1: Victoria South - Holborn

Creating superstations

According to the **Definition**, the origin station, **Victoria** could be potentially grouped together with other stations as for all reasonable routes of the **Victoria** – **Holborn** station-to-station OD pair the first journey leg is on the **Victoria** line (cf. **Figure 3-3**). Therefore, the candidate stations of the origin superstation are the stations between **Brixton** and **Victoria** (cf. **Figure 5-8**). Referring to the **Definition**, the destination station (**Holborn**) cannot be grouped with other stations, as the last journey leg of the reasonable routes are on different lines (**Central** and **Piccadilly** lines).

In order to decide whether these origin stations can be grouped together – based on the **Definition** – the route choice set of each station-to-station OD pair was compared by applying the route choice set generation algorithm with the appropriate considerations for the OD specific attribute cut-offs (cf. **Section 5.3.3**) on the extended LU inner zone network (cf. **Figure 5-8**) for the OD pairs presented in **Table 5-7**.

Table 5-7 OD pairs for which route choice set was compared, **Case 1**

OD pair	Origin	Destination
1	**Brixton**	**Holborn**
2	**Stockwell**	**Holborn**
3	**Vauxhall**	**Holborn**
4	**Pimlico**	**Holborn**
5	**Victoria**	**Holborn**

Results (see **Figure 5-9**) show that the route choice set is the same from all origin stations, except from **Stockwell** station. The reason why the **Stockwell** – **Holborn** station-to-station OD pair has a different route choice set is that passengers boarding at **Stockwell** station can also take the **Northern** line[20] towards their destination and change either at **Tottenham Court Road** station to the **Central** line or at **Leicester Square** station to the

[20] The **Northern** line of the London Underground has two branches within the inner zone: via **Bank** and via **Charing Cross**. Most of the services that depart from **Stockwell** go via **Bank** and there are only few direct services that go via **Charing Cross**. In most of the cases passengers need to change at **Kennington**. Here it was assumed, that passengers choosing the **Northern** line at **Stockwell** arrive at the platform according to the departure time of the direct service, and therefore a lower value of wait time was considered.

Piccadilly line (cf. **Figure 5-8**). The generalised cost proportion of these routes are 1.12 and 1.13 respectively. Looking this more precisely, it can be further understood, that as at **Stockwell** station the northbound **Victoria** and **Northern** lines depart from adjacent platforms, passengers do not necessarily choose routes, but strategies including the option to board whichever line comes first (Nguyen and Pallottino, 1988; Spiess and Florian, 1989). Due to these differences in the route choice set, **Stockwell** station cannot be grouped with the other stations as a superstation (**Rule 2**).

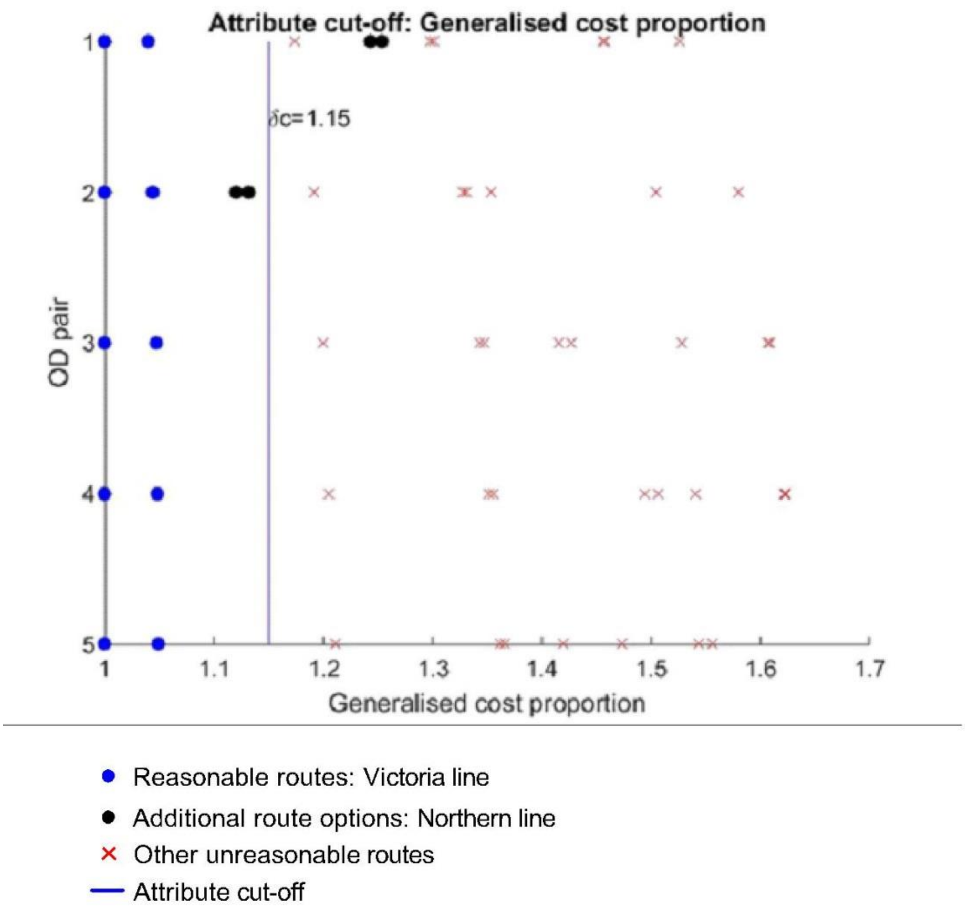


Figure 5-9 Attribute cut-off according to generalised cost proportions, **Case 1**

Among the other stations, it is clear that from **Vauxhall** and **Pimlico** stations passengers have the same route choice set as from **Victoria** station, as there are no additional attractive route options in those cases. Results showed, that passengers have the same route choice set also from **Brixton** station: even though they have the option to change to the **Northern** line at **Stockwell**, those routes are not reasonable (generalised cost proportion is 1.24 and 1.25 respectively, see. **Figure 5-9**).

In summary, the following stations could be grouped as the origin superstation: **Brixton**, **Vauxhall**, **Pimlico** and **Victoria**. This superstation is named **Victoria South**[21] (**Figure 5-10**). The fact of including **Brixton**, but not **Stockwell** station illustrates, that even though stations are not consecutive, they can be still grouped as superstations (**Rule 3**).

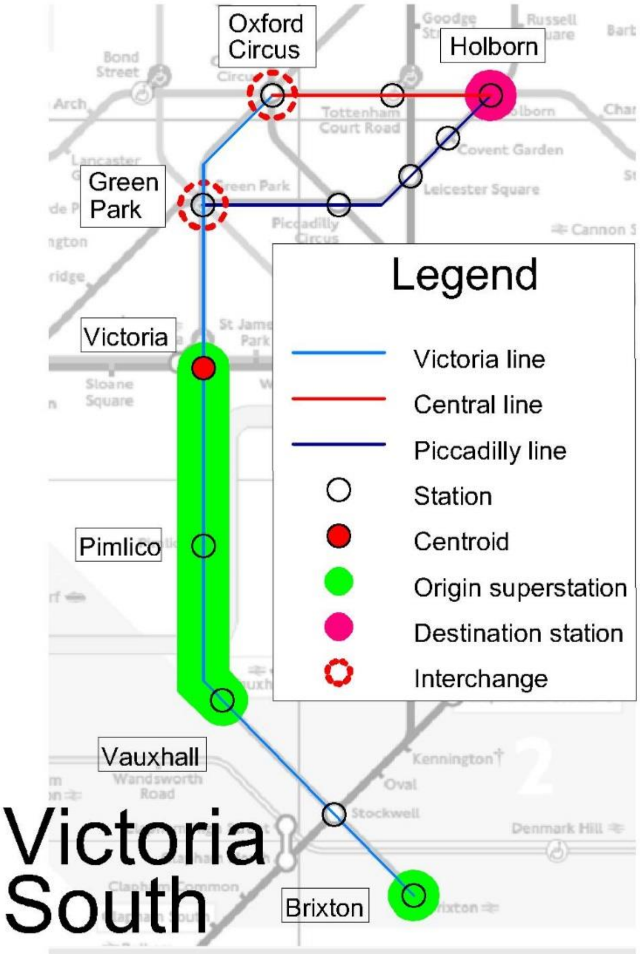


Figure 5-10 The **Victoria South**– **Holborn** superstation-to-station OD pair

Comparing the generalised cost proportion of the shortest routes ($\rho c_{k,ij}$) for the four OD pairs that have their origin at the **Victoria South** superstation (cf. **Figure 5-9**) it can be understood that they depend on the OD minimum journey time. Looking at the third shortest route: **Victoria** – **Piccadilly** (via **King's Cross**), for the OD pair with the longest minimum travel time (**Brixton – Holborn**) $\rho c_{3,ij}$ is 1.17, while for the OD pair with the shortest minimum travel time (**Victoria – Holborn**) it is 1.21. According to the earlier considerations (cf. **Chapter 4**), the option to change at **King's Cross** station is not a

[21] The **character border** in the text denotes superstation

reasonable route for these OD pairs, as staying on the **Victoria** line after **Oxford Circus** station would be a sort of turning away from the destination (Dial, 1971). Therefore the attribute cut-off ρc was set lower: 1.15 instead of 1.18 so that the route **Victoria** – **Piccadilly** (via **King's Cross**) could be excluded for all OD pairs.

Obtaining Centroid-to-Centroid adjusted Observed Journey Times

Similarly to the **Victoria** – **Holborn** station-to-station OD pair, the sample size of OJTs was small also for the OD pairs from the other stations of the origin superstation (**Figure 5-11**). Therefore a larger and better distributed data sample needs to be obtained (cf. **Section 5.4**).

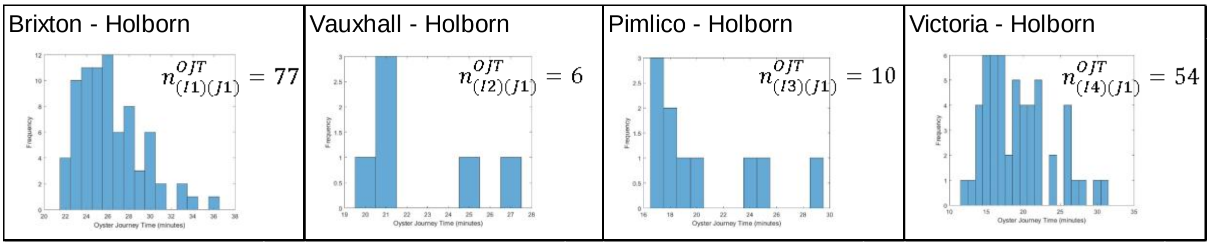


Figure 5-11 Distribution of Observed Journey Times (OJT) for station-to-station OD pairs for **Victoria South**– **Holborn**

For the origin superstation, **Victoria South**, the **Victoria** station was selected as the superstation centroid (marked with ● on **Figure 5-10** and highlighted with green in **Table 5-8**) to make the comparison more straightforward with the case study on the station-to-station OD pair (cf. **Section 5.4.1**). The destination is a single station (**Holborn**), therefore it is the centroid itself.

Table 5-8 Adjustment of OJTs according to on-board and access time for **Victoria South**– **Holborn**, superstation centroids are highlighted with green.

Origin	On-board time			Access time		
Ii	$t^{ob}_{(Ii-1)(Ii)}$ [min]	$t^{ob}_{(I1)(Ii)}$ [min]	$t^{ob}_{(Ii)(\mathrm{Ic})}$ [min]	t^{acc}_{Ii} [min]	$\Delta t^{acc}_{(Ii)(Ic)}$ [min]	
Brixton		0	7	1.9	0.5	1
Vauxhall	4	4	3	1.4	1.0	1
Pimlico	1	5	2	1.7	0.7	1
Victoria	2	7	0	2.4	0.0	0

Following this, the OJTs of the station-to-station OD pairs $(OJT_{(Ii)(Jj)})$ were adjusted to superstation centroid according to equation (5-8) (cf. **Section 5.4.2**). This way the CCOJTs corresponding to each station-to-station OD pair $(CCOJT_{(Ii)(Jj)})$ are obtained. **Table 5-8** shows the necessary adjustments according to the on-board $(t^{ob}_{(Ii)(Ic)})$ and access $(\Delta t^{acc}_{(Ii)(Ic)})$ time at the origin superstation. As the destination is a single station (**Holborn**), there is no need to do adjustments according to on-board $(t^{ob}_{(Jj)(Jc)})$ and egress $(\Delta t^{egr}_{(Jj)(Jc)})$ times at the destination. As all services on the **Victoria** line terminate at **Brixton**, the frequency and hence the wait time is the same at all origin stations, there is no need to do adjustments according to wait time $(\Delta t^{wait}_{(Ii)(Ic)})$. All adjustments values are rounded to the nearest minute as the OJTs from Oyster data are given with that precision.

The CCOJTs corresponding to each station-to-station OD pair $(CCOJT_{(Ii)(Jj)})$ were aggregated (cf. **Section 5.4.3**), resulting in a dataset $(CCOJT_{IJ})$ with larger sample size $(n^{CCOJT}_{IJ} = 147)$ (**Figure 5-12**).

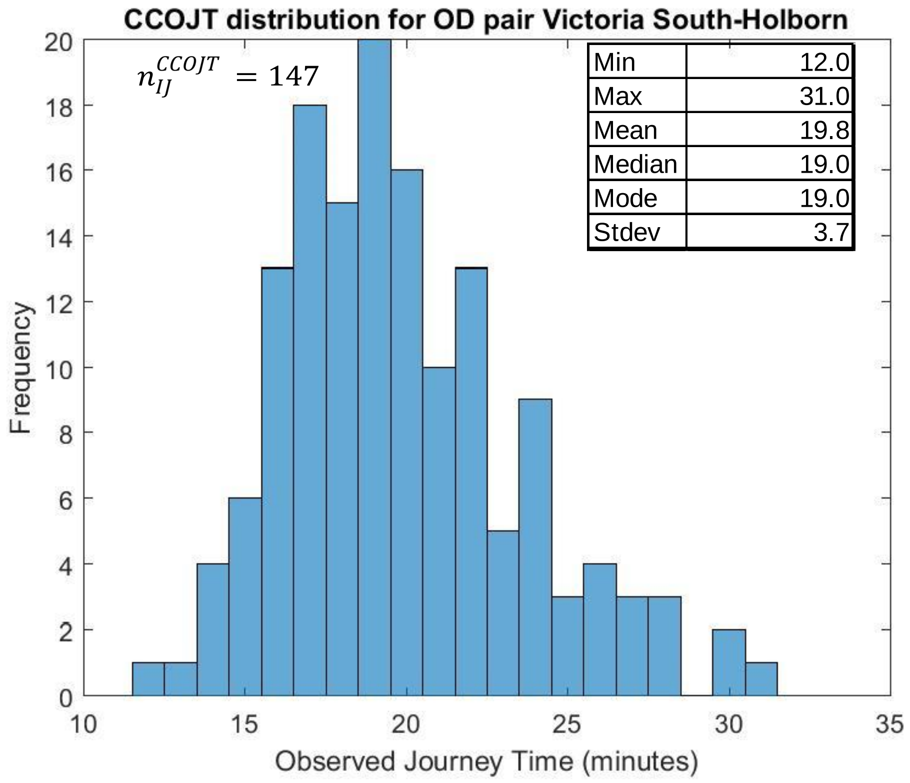


Figure 5-12 Distribution of Centroid-to-Centroid adjusted Observed Journey Times for **Victoria South**– **Holborn**

Evaluation of the benefits of the superstation representation

The finite mixture model (cf. **Chapter 3**) was applied on this larger dataset of CCOJTs. Within this dataset all entries could be considered as valid data, because the upper outer fence (cf. **Section 3.2.1**) resulted 37 minutes, while the maximum CCOJT value is 31 minutes. This valid dataset is denoted by $CCOJT^0$.

As for the superstation-to-station OD pair – similarly to the case study in **Section 3.7** – two reasonable routes were assumed, route choice was estimated as a two-component ($N_R = 2$) finite mixture distribution. Therefore, the K-means clustering algorithm was applied on the $CCOJT^0$ dataset with two clusters and with the settings described in **Section 3.3.1** to produce the initial values for the EM algorithm. Using these initial values, the EM algorithm was run with different settings for the tolerance threshold (cf. **Section 3.3.2**). The more detailed results of initial values and tolerance thresholds are reported in **Appendix F**.

From there it is understood that when the tolerance threshold is 1e-06 or greater, the EM algorithm converges to a root close to the initial value for seed 1. But when the tolerance threshold is 1e-07 or smaller, the EM algorithm converges to a root around 18.1 minutes for the mean and 66.0% for the proportion of component 1 for both seeds (see **Figure F-1** and **Figure F-2**). Similar properties could be observed for the other mixture component (labelled with $r = 2$). The log-likelihood exhibits a considerable jump between the tolerance threshold of 1e-02 and 1e-03 (**Figure F-3**).

According to RODS data, the aggregate route choice proportions for the two routes ($\omega_{k,IJ}^{RODS}$) of the **Victoria South** – **Holborn** superstation-to-station OD pair are 74.8% and 25.2% (see **Table 5-10**). Among the estimates, the one with seed 1 and tolerance threshold 1e-06 gives the best approximation to RODS results, therefore these settings were chosen for the finite mixture model (**Table 5-9**). Through this case study it resulted that this tolerance threshold is smaller than the one in case of station-to-station OD pairs (i.e. 1e-05, cf. **Section 3.7**).

Table 5-9 Finite mixture model results; with Seed = 1, Tolerance threshold = 1e-06 for Victoria South – Holborn,

OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Label	Mixture model		
r []	$\mu_{r,IJ}^{MIX}$ [min]	$\sigma_{r,IJ}^{MIX}$ [min]	$\omega_{r,IJ}^{MIX}$ [%]
1	18.2	2.4	70.8%
2	23.7	3.4	29.2%

Following this, the results of the finite mixture model were matched with the actual LU routes (cf. **Section 3.4.1**). As the centroid of the origin superstation coincides with the origin station of **Case 1** in **Chapter 3** (Victoria) and the destination station is a single station (Holborn), the Scheduled Journey Time of the actual LU routes ($t_{k,IJ}^{SJT}$) are the same as the results in **Table 3-5**. It is expected that the mixture component with the lower mean ($r = 1$) corresponds to the route with the shorter journey time ($k = 1$). Similarly the component with the higher mean ($r = 2$) to the route with the longer journey time ($k = 2$).

Table 5-10 compares the mixture results for the Victoria South – Holborn superstation-to-station OD pair with the Victoria – Holborn station-to-station OD pair, together with the results in Fu (2014) and corresponding values of the actual LU routes. **Figure 5-13** presents the probability density functions of the mixture distribution fit on the CCOJT dataset as well as of the mixture components matched with the actual LU routes.

Based on these results, the following was observed: While for the Victoria – Holborn station-to-station OD pair the proportion of mixture component 1 ($\omega_{r,(Ii)(Jj)}^{MIX}$) exhibited a significant jump from 79.8% to 33.5% between tolerance thresholds 1e-05 and 1e-06 (cf. **Figure 3-6**), the same jump for the Victoria South – Holborn superstation-to-station OD pair was much less ($\omega_{r,IJ}^{MIX}$) (from 70.8% to 67.9%) and it occurred between tolerance thresholds 1e-06 and 1e-07 (**Figure F-2**). This explains, that using a larger and better distributed dataset of superstation-to-superstation OD pairs gives more stable route choice results, which stays closer to the initial value (i.e. expected route choice results from RODS) even for smaller tolerance thresholds.

Furthermore, while the Victoria – Holborn station-to-station OD pair the log-likelihood had the jump between the tolerance threshold of 1e-05 and 1e-06 (cf. **Figure 3-7**), the same type of jump occurred between the tolerance threshold of 1e-02 and 1e-03 for the **Victoria South** – **Holborn** superstation-to-station OD pair (**Figure F-3**). From this it can be understood, that the estimates for the superstation-to-superstation OD pairs are more reliable also at a greater threshold.

Table 5-10 Matching mixture model results with the actual London Underground routes for **Victoria South– Holborn**

Purple: Mixture results, superstation OD pairs, **Red**: Mixture results, station OD pairs**, Yellow**: Fu (2014)**, Green**: actual LU routes

Mixture Label	Journey Time (min)				Route Choice (%)				Route Label	Route Matched		
	Mixture			Timetable	Mixture			RODS		Line 1	Interchange 1	Line 2
	Proposed		Fu(2014)		Proposed		Fu(2014)					
	Superstation	Station			Superstation	Station						
r	$\mu^{MIX}_{r,IJ}$	$\mu^{MIX}_{r,(Ii)(Jj)}$		$t^{SJT}_{k,IJ}$	$\omega^{MIX}_{r,IJ}$	$\omega^{MIX}_{r,(Ii)(Jj)}$		$\omega^{RODS}_{k,IJ}$	k	$l = 1$	$s = 1$	$l = 1$
1	18.2	17.6	16.6	17.6	70.8%	79.8%	75.4%	74.8%	1	Vic	Oxford Circus	Cen
2	23.7	26.1	22.2	20.4	29.2%	20.2%	24.6%	25.2%	2	Vic	Green Park	Pic

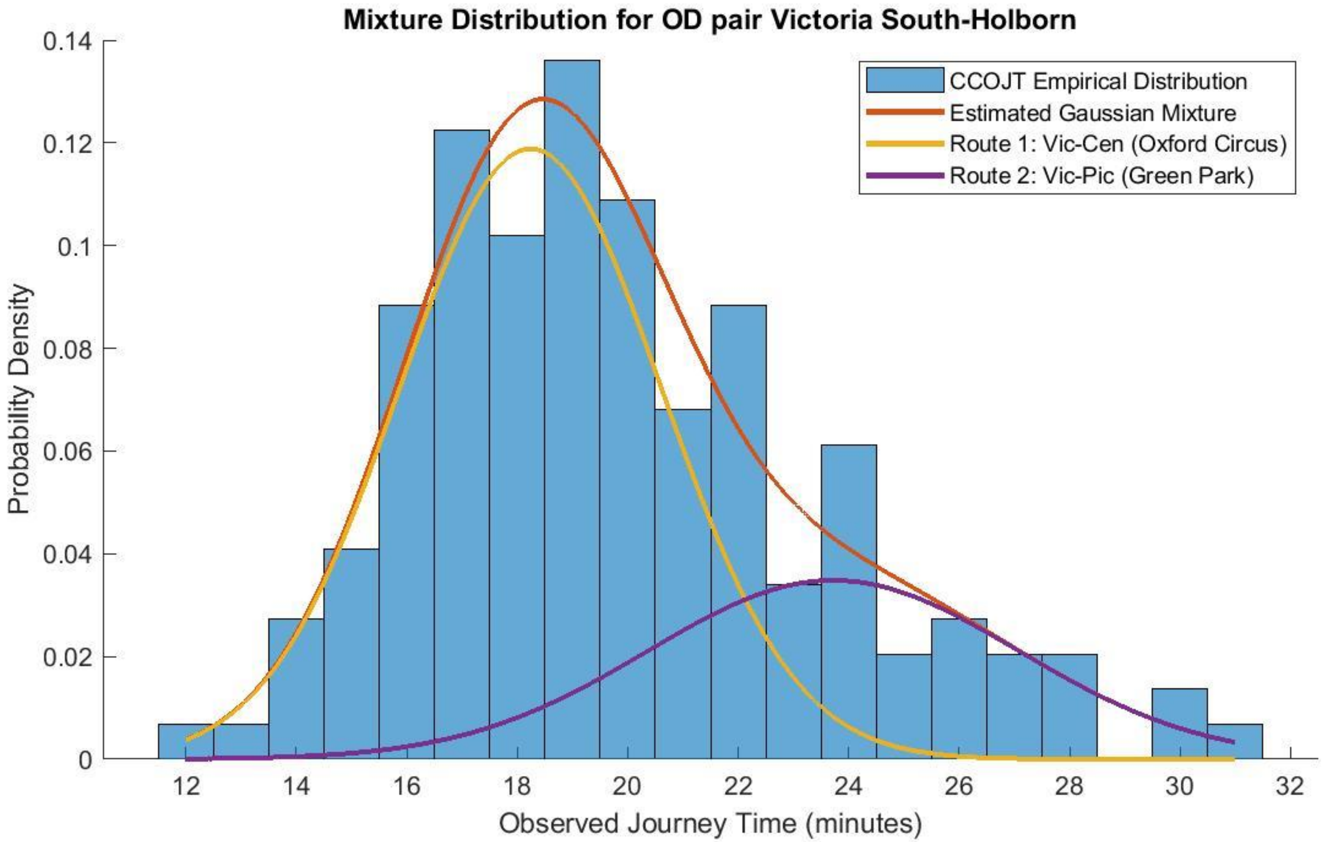


Figure 5-13 Estimated (Gaussian) journey time distribution of the routes for **Victoria South** – **Holborn**, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Case 2 Central East – Green Park

Creating superstations

According to the **Definition**, the origin station, **Liverpool Street** could be potentially grouped together with other stations, as for all reasonable routes of the **Liverpool Street** – **Green Park** station-to-station OD pair the first journey leg is on the **Central** line (cf. **Figure 3-9**). Therefore, the candidate stations of the origin superstation are the stations between **Epping** and **Chancery Lane**, including the stations on the **Hainault** loop (cf. **Figure 5-8**). Referring to the **Definition**, the destination station (**Green Park**) cannot be grouped with other stations, as the last journey leg of the reasonable routes are on different lines (**Central**, **Piccadilly** and **Jubilee** lines).

Table 5-11 OD pairs for which route choice set was compared, **Case 2**

OD pair	Origin	Destination
1	**Epping**	**Green Park**
2	**Theydon Bois**	**Green Park**
3	**Debden**	**Green Park**
4	**Loughton**	**Green Park**
5	**Buckhurst Hill**	**Green Park**
6	**Woodford**	**Green Park**
7	**South Woodford**	**Green Park**
8	**Snaresbrook**	**Green Park**
9	**Leytonstone**	**Green Park**
10	**Leyton**	**Green Park**
11	**Stratford**	**Green Park**
12	**Mile End**	**Green Park**
13	**Bethnal Green**	**Green Park**
14	**Liverpool Street**	**Green Park**
15	**Bank**	**Green Park**
16	**St Paul's**	**Green Park**
17	**Chancery Lane**	**Green Park**

In order to decide whether these origin stations can be grouped together – based on the **Definition** – the route choice set of each station-to-station OD pair was compared by applying the route choice set generation algorithm with the appropriate considerations for the OD specific attribute cut-offs (cf. **Section 5.3.3**) on the extended LU inner zone network (cf. **Figure 5-8**) for the all the OD pairs, whose origin station is on the **Epping** branch[22] of the **Central** line (between **Epping** and **Chancery Lane**) and whose destination station is **Green Park** (**Table 5-11**).

Results (see **Figure 5-14**) show that the route choice set from **Stratford** (OD 11) and **Bank** (OD 15) stations are quite different from the patterns of other stations. This is because from these stations other attractive lines are available apart from the **Central** line. From **Stratford** station passengers can also take the **Jubilee** line to **Green Park**. Although this route has slightly longer travel time, than the route **Central** – **Victoria** via **Oxford Circus** (32.7 and 29.7 minutes respectively), it has the advantage that it is a direct service. Furthermore, as **Stratford** station is the terminus of the **Jubilee** line, trains are not crowded at boarding and hence passengers may be able to get a seat. In accordance with this, the route choice set generation algorithm gave this option as the shortest route (in terms of generalised costs) for the **Stratford** – **Green Park** OD pair. From **Bank** station, many lines are available, among which taking the **Northern** line to **London Bridge**, then changing to the **Jubilee** line (cf. **Figure 5-8**) is shown to be a reasonable route. Due to these differences in the route choice set, **Stratford** and **Bank** stations cannot be grouped together with the other stations as a superstation (**Rule 2**).

Among the other stations, the route choice set is the same from the stations before **Stratford** (OD 1-10). From these stations, apart from the routes via the **Central** line, also the route option to change to the Jubilee line at **Stratford** is reasonable. It is shown to be the second shortest route (see **Figure 5-14**).

Similarly, also the stations after **Stratford**, excluding **Bank** (OD 12, 13, 14, 16 and 17) have the same route choice set among each other: the routes via the **Central** line. From **Liverpool Street** (OD 14), apart from the **Central** line, also the **Circle** line is available and the **Circle** – **Victoria** via **King's Cross** route is shown to be the fourth shortest route (1.10 times the generalised cost of the shortest route, see **Figure 5-14**). However, as it was discussed earlier (cf. **Section 4.7.4.2**), a lower cut-off (δc) value was suggested for

[22] In order to make the network model simpler, the **Central** line was represented only with the stations on the **Epping** branch. It was assumed that form the stations on the **Hainault** loop, passengers are expected to have the same route choice set.

the **Liverpool Street** – **Green Park** OD pair; and hence only the first two shortest routes (both having their first journey leg on the **Central** line) were considered reasonable. This way, the other stations with their origin on the **Central** line (**Mile End**, **Bethnal Green**, **St Paul's** and **Chancery Lane**) could be grouped together with the **Liverpool Street** station

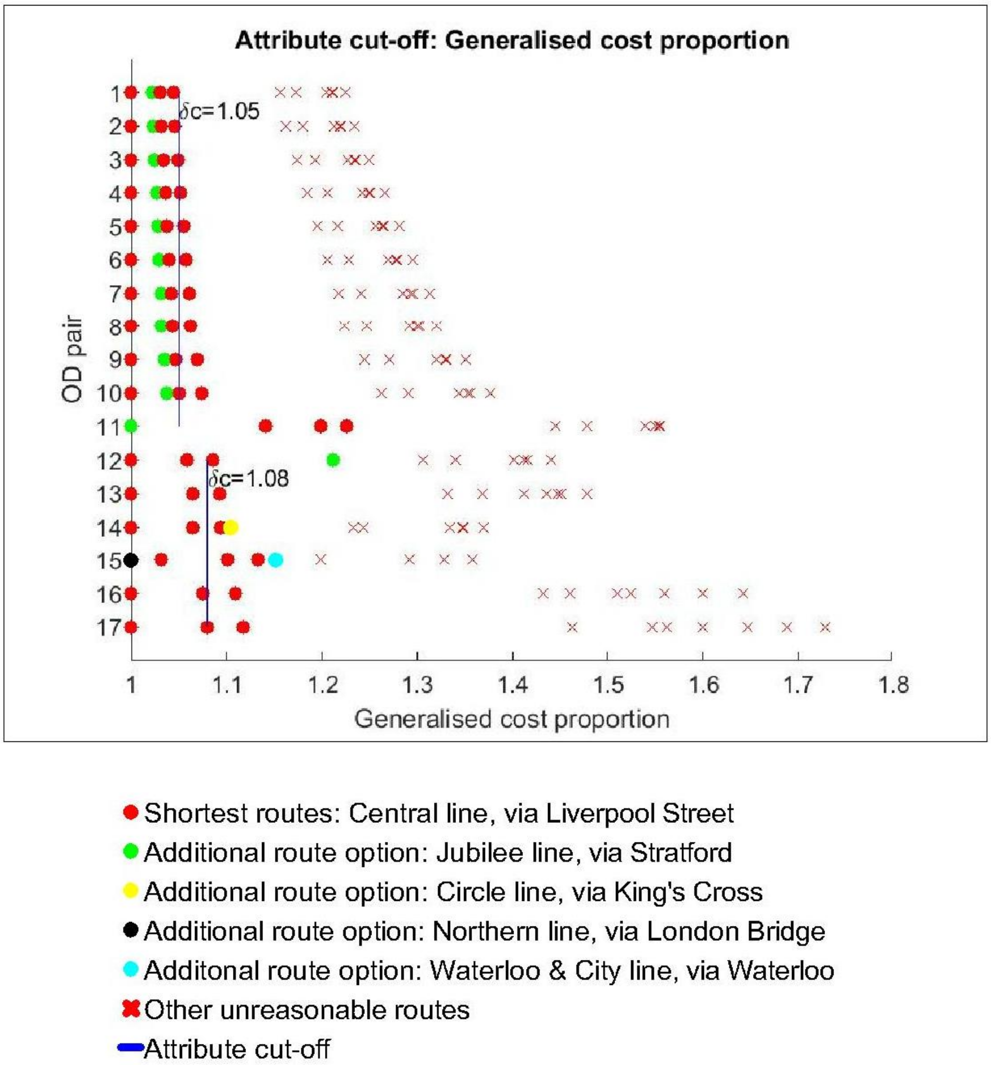


Figure 5-14 Attribute cut-off according to generalised cost proportions, **Case 2**

In summary, two superstations could be created by grouping the stations on the east end of the **Central** line (**Figure 5-15**):

- All the stations between **Epping** and **Leyton**, including the **Hainault** loop: 20 stations in total. This superstation is named **Central East Outer**.

- Stations between **Mile End** and **Chancery Lane**, excluding **Bank**: 4 stations in total. This superstation is named **Central East**.

s

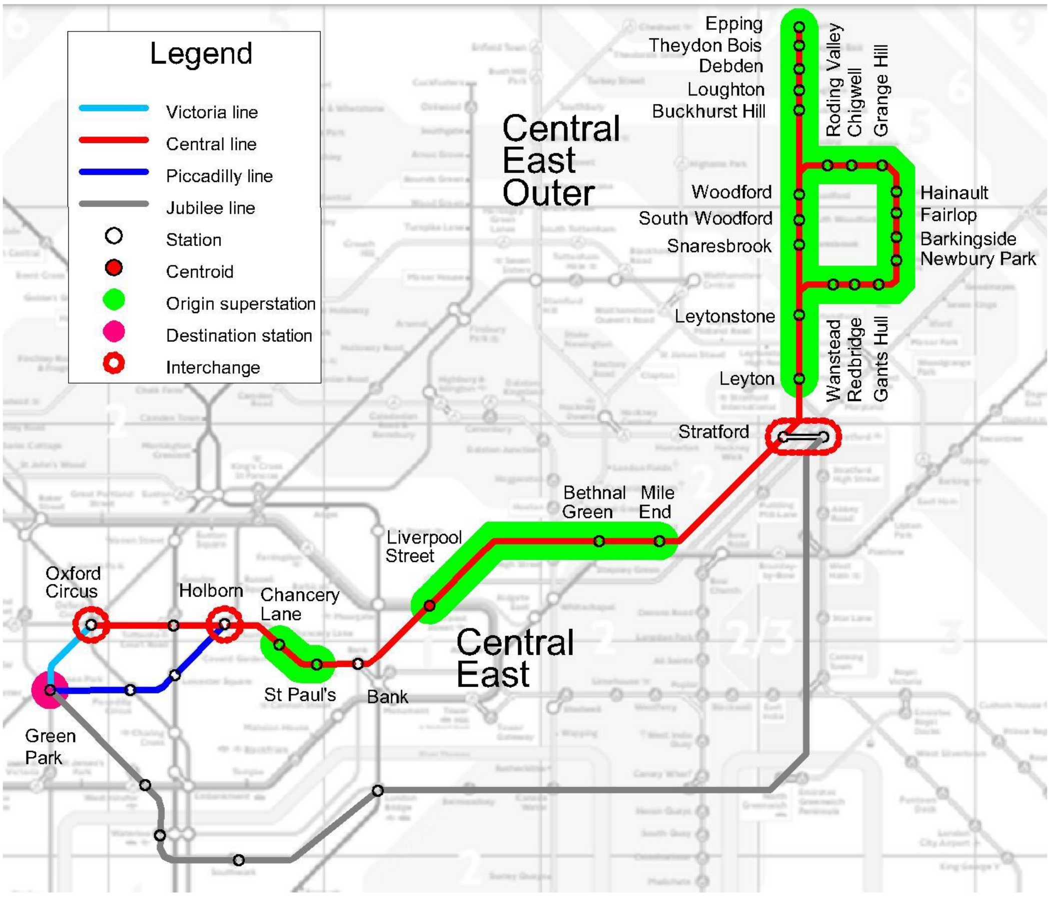


Figure 5-15 The **Central East** – **Green Park** superstation-to-station OD pair

Within the **Central East Outer** superstation, there are multiple **Central** line services (i.e. **Epping** branch, **Hainault** loop and the short runs on it). This would result that the service frequencies and hence the wait time are not equal at different origin stations. Based on **Rule 4**, these stations can be still grouped as superstations, just the appropriate adjustment needs to be made according to the difference in wait time.

The fact of excluding **Bank** station from the **Central East** superstation, but including the stations before and after, illustrates, that even though stations are not consecutive, they can be still grouped as superstations (**Rule 3**).

Regarding the attribute cut-off values (ρc), the following needs to be considered: For the OD pairs originating at the **Central East Outer** superstation, for all routes the first journey leg is on the **Central** line and the different route options are to change to the connecting lines that bring to the destination (**Green Park**). For these OD pairs the main attribute that influences ρc is the OD minimum journey time: Regarding the OD pairs originating at the **Central East** superstation, at some stations (e.g. **Liverpool Street**) there are other available lines for the first journey leg. Therefore for these OD pairs the main attribute that influences ρc is the number of available lines at the origin and destination station. As it was elucidated earlier (cf. **Section 3.7**), the **Central** – **Jubilee** (via **Bond Street**) route would be a sort of turning away from the destination (Dial, 1971), therefore the attribute cut-offs were set to exclude that route from the set of reasonable routes. This way ρc resulted 1.05 for the OD pairs that originate at the **Central East Outer** superstation and 1.08 for those, which originate at the **Central East** superstation (**Figure 5-14**). This is much lower than the value obtained in **Chapter 4** (1.18).

In the remaining part of this case study, the adjustment of Observed Journey Times (OJT) and the application of the finite mixture model on the Centroid-to-Centroid adjusted OJTs (CCOJT) is elaborated only for the **Central East** – **Green Park** superstation-to-station OD pair. Following the previous discussion on the exclusion of the **Central** – **Jubilee** (via **Bond Street**) route from the reasonable route choice set, two reasonable routes are considered for this OD pair and hence route choice is estimated with a two-component finite mixture model.

Obtaining Centroid-to-Centroid adjusted Observed Journey Times

Similarly to the **Liverpool Street** – **Green Park** station-to-station OD pair, the sample size of OJTs was small also for the OD pairs from the other stations of the origin superstation (**Figure 5-16**). Among them, no OJT record was available for the **Chancery Lane** – **Green Park** OD pair, and only 1 observation for the **St. Paul’s** – **Green Park** OD pair. Obviously, in those cases there is less sense to talk about route choice probabilities. Therefore a larger and better distributed dataset needs to be obtained (cf. **Section 5.4**).

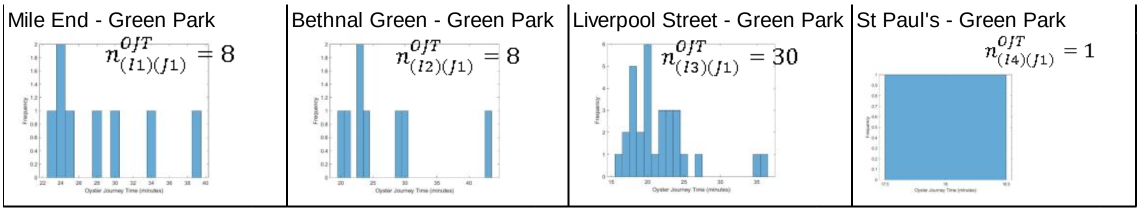


Figure 5-16 Distribution of Observed Journey Times (OJT) for station-to-station OD pairs for **Central East**– **Green Park**

For the origin superstation, **Central East**, **Liverpool Street** station was selected as the superstation centroid (marked with ● on **Figure 5-15** and highlighted with green in **Table 5-12**) to make the comparison more straightforward with the case study on the station-to-station OD pair (cf. **Section 5.4.1**). The destination is a single station (**Green Park**), therefore it is the centroid itself.

Table 5-12 Adjustment of OJTs according to on-board and access time for **Central East**– **Green Park**, superstation centroids are highlighted with green.

Origin	On-board time			Access time		
Ii	$t^{ob}_{(Ii-1)(Ii)}$ [min]	$t^{ob}_{(I1)(Ii)}$ [min]	$t^{ob}_{(Ii)(\mathrm{Ic})}$ [min]	t^{acc}_{Ii} [min]	$\Delta t^{acc}_{(Ii)(Ic)}$ [min]	
Mile End		0	6	0.3	2.3	2
Bethnal Green	3	3	3	0.4	2.2	2
Liverpool Street	**3**	**6**	**0**	**2.6**	**0.0**	**0**
St Pauls	3	9	-3	1.4	1.3	1
Chancery Lane	2	11	-5	1.4	1.2	1

Following this, the OJTs of the station-to-station OD pairs ($OJT_{(Ii)(Jj)}$) were adjusted to superstation centroid according to equation (5-8) (cf. **Section 5.4.2**). This way the CCOJTs corresponding to each station-to-station OD pair ($CCOJT_{(Ii)(Jj)}$) are obtained. **Table 5-12** shows the necessary adjustments according to the on-board ($t^{ob}_{(Ii)(Ic)}$) and access ($\Delta t^{acc}_{(Ii)(Ic)}$) time at the origin superstation. As the destination (**Green Park**) is a single station, there is no need to do adjustments according to on-board ($t^{ob}_{(Jj)(Jc)}$) and egress ($\Delta t^{egr}_{(Jj)(Jc)}$) times at the destination. As all services on the Central line start before **Mile End** station, the frequency and hence the wait time is the same at all origin stations, there is no need to do adjustments according to wait time ($\Delta t^{wait}_{(Ii)(Ic)}$). All adjustments values are rounded to the nearest minute as the OJTs from Oyster data are given with that precision.

The CCOJTs corresponding to each station-to-station OD pair ($CCOJT_{(Ii)(Jj)}$) were aggregated (cf. **Section 5.4.3**), resulting in a dataset ($CCOJT_{IJ}$) with larger sample size ($n_{IJ}^{CCOJT} = 47$) (see **Figure 5-17**).

Evaluation of the benefits of the superstation representation

The finite mixture model presented in **Chapter 3** was applied on this larger dataset of CCOJTs. Within this dataset, one CCOJT value was considered as an outlier (42 minutes); because it is above the upper outer fence, which resulted 38.25 minutes (cf. **Section 3.2.1**). This valid dataset with 46 entries is denoted by $CCOJT^0$ (**Figure 5-18**).

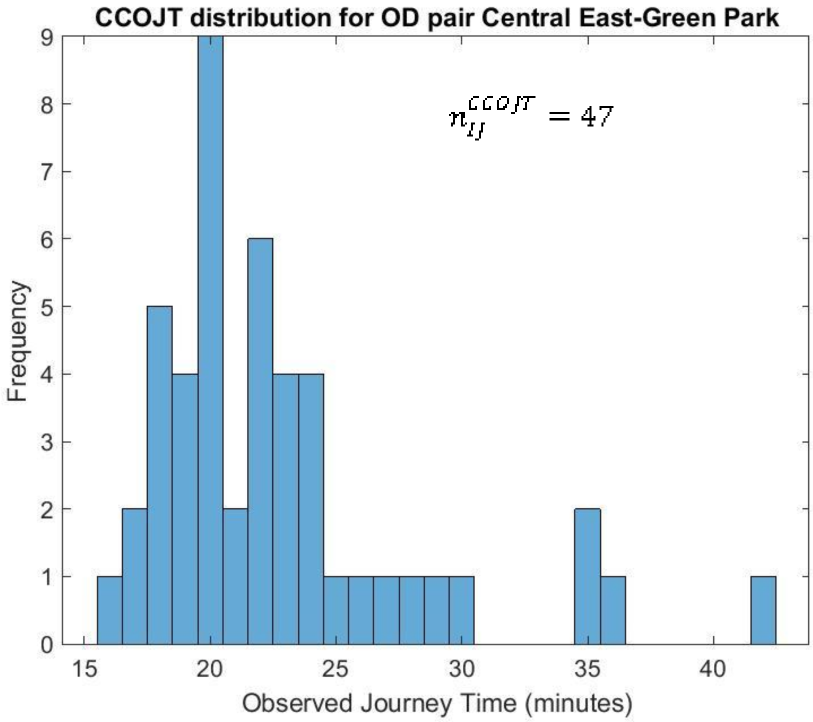
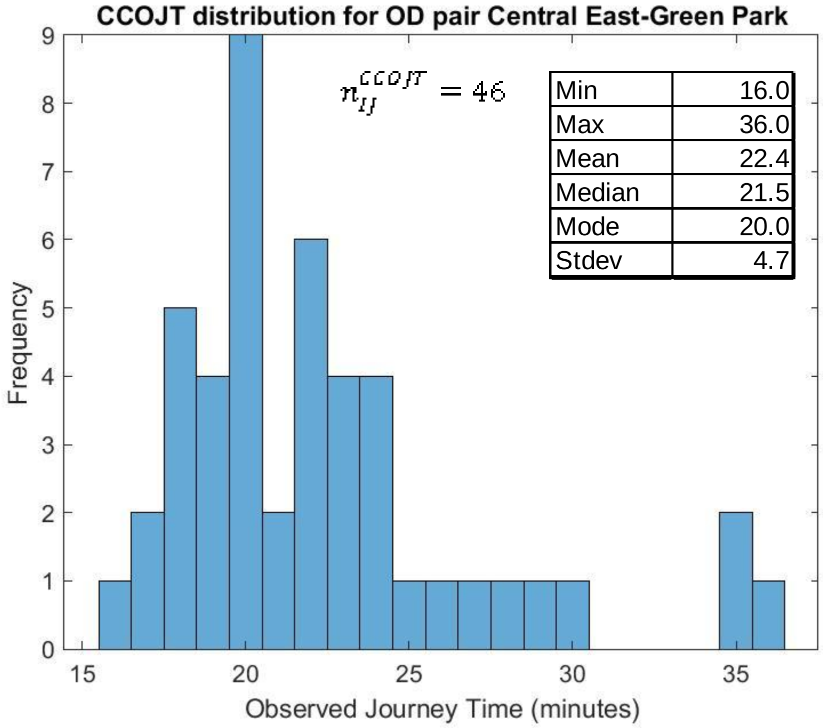


Figure 5-17 Distribution of Centroid-to-Centroid adjusted Observed Journey Times for **Central East– Green Park**



Min	16.0
Max	36.0
Mean	22.4
Median	21.5
Mode	20.0
Stdev	4.7

Figure 5-18 Valid dataset of Centroid-to-Centroid adjusted Observed Journey Times for **Central East– Green Park**

Based on the results of the route choice set generation algorithm (**Figure 5-14**), the **Central East** – **Green Park** superstation-to-station OD pair was modelled with two reasonable routes, hence route choice was estimated as a two-component ($N_R = 2$) finite mixture distribution. Therefore, the K-means clustering algorithm was applied on the OJT^0 dataset with two clusters and with the settings described in **Section 3.3.1** to produce the initial values for the EM algorithm. Using these initial values, the EM algorithm was run with different settings for the tolerance threshold (cf. **Section 3.3.2**). The more detailed results of initial values and tolerance thresholds are reported in **Appendix F**.

From there it is understood, that the EM algorithm converges to a similar value for a range of tolerance thresholds both for the mean and for the component proportion. For the mixture component labelled with $r = 1$, it starts plateauing from the tolerance threshold of 1e-07 around the value of 20.6 minutes for the mean and 80.3% for the component proportion (see **Figure F-4** and **Figure F-5**). Similar properties could be observed for the other mixture component labelled with $r = 2$. Due to these considerations, the finite mixture model was applied with the tolerance threshold of 1e-07 (**Table 5-13**). In this case, this is identical to the tolerance threshold chosen for the station-to-station OD pair (cf. **Section 3.7**).

Table 5-13 Finite mixture model results, tolerance threshold = 1e-07
for **Central East**– **Green Park**,
OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Label	Mixture model		
r []	$\mu_{r,IJ}^{MIX}$ [min]	$\sigma_{r,IJ}^{MIX}$ [min]	$\omega_{r,IJ}^{MIX}$ [%]
1	20.6	2.3	80.3%
2	29.5	4.7	19.7%

Following this, the results of the finite mixture model were matched with the actual LU routes (cf. **Section 3.4.1**). As the centroid of the origin superstation coincides with the origin station of **Case 2** in **Chapter 3** (**Liverpool Street**) and the destination station is a single station (**Green Park**), the total journey times of the actual LU routes ($t_{k,IJ}^{SJT}$) are the same as the results in **Table 3-11**. The mixture components were matched with the actual LU routes in order of their journey times.

Table 5-14 compares the mixture results for the **Central East** – **Green Park** superstation-to-station OD pair with the **Liverpool Street** – **Green Park** station-to-station OD pair together with corresponding values of the actual LU routes. Here, the comparison could not be made with the results of Fu (2014) as in his work route choice was estimated as a three-component mixture distribution, while in this chapter it was treated with two components. **Figure 5-19** presents the probability density functions of the mixture distribution fit on the CCOJT dataset as well as of the mixture components matched with the actual LU routes.

Based on these results, the following was observed: The component proportion results of the mixture model for the **Central East** – **Green Park** superstation-to-station OD pair (80.3% and 19.7%) gave a good match to the RODS (cf. **Section 3.6.3**) route choice data (81.2% and 18.8%). The same for the **Liverpool Street** – **Green Park** station-to-station with two mixture components was 93.3% and 6.7%, which is very far from the actual results.

However, the mean journey time results for the two components (20.6 and 29.5 minutes) did not show a good match to the actual LU journey times (21.3 and 24.0 minutes). A possible explanation for this could be, that the higher OJT observations (i.e. 35-36 minutes) could be also attributed to a fail-to-board event, not necessarily to the longer route.

Table 5-14 Matching mixture model results with the actual London Underground routes for **Central East– Green Park**

Purple: Mixture results, superstation OD pairs, **Red**: Mixture results, station OD pairs, **Green**: actual LU routes

Mixture Label	Journey Time (min)			Route Choice (%)			Route Label	Route Matched		
	Mixture		Timetable	Mixture		RODS		Line 1	Interchange 1	Line 2
	Proposed			Proposed						
	Superstation	Station		Superstation	Station					
r	$\mu_{r,IJ}^{MIX}$	$\mu_{r,(Ii)(Jj)}^{MIX}$	$t_{k,IJ}^{SJT}$	$\omega_{r,IJ}^{MIX}$	$\omega_{r,(Ii)(Jj)}^{MIX}$	$\omega_{k,IJ}^{RODS}$	k	$l = 1$	$s = 1$	$l = 2$
1	20.6	20.6	21.3	80.3%	93.3%	81.2%	1	Central	Oxford Circus	Victoria
2	29.5	35.5	24.0	19.7%	6.7%	18.8%	2	Central	Holborn	Piccadilly

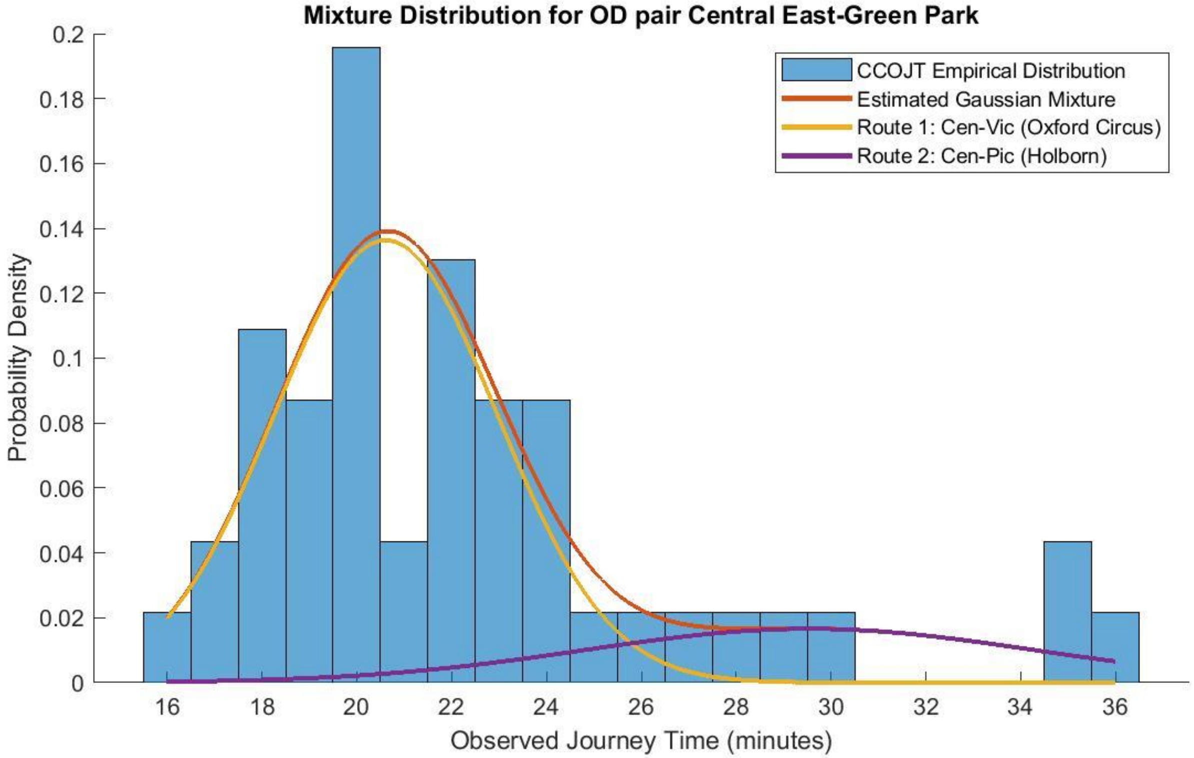


Figure 5-19 Estimated (Gaussian) journey time distribution of the routes for **Central East– Green Park**, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Case 3 Jubilee West – Jubilee Central

Creating superstations

Case 3 represents the choice problem of those passengers whose origin is on the west end of the **Jubilee** line (between **Stanmore** and **Kingsbury** stations) and their destination is on the central part of the **Jubilee** line (between **Bond Street** and **London Bridge** stations). For them one alternative is to choose a direct service (**Jubilee** line) with many intermediate stops between **Wembley Park** and **Finchley Road** stations. The other alternatives are to change to an express service (**Metropolitan** line) at **Wembley Park** station, then change back to the **Jubilee** line either at **Finchley Road** or at **Baker Street** station. Additionally, to some destinations also changing to the **Bakerloo** line at **Baker Street** station can be a convenient alternative (see **Figure 5-20**). In order to further understand which of these route options are reasonable, the route choice set generation algorithm is applied with the appropriate considerations for the OD specific attribute cut-offs (cf. **Section 5.3.3**).

According to the **Definition**, both the origin and destination stations could be potentially grouped together, as for all reasonable routes of the corresponding OD pairs the first and the last journey leg is on the **Jubilee** line. Therefore, the candidate stations of the origin superstation are the stations between **Stanmore** and **Kingsbury**; and the candidate stations of the destination superstation are the stations between **Bond Street** and **London Bridge**.

In order to decide whether these origin and destination stations can be grouped together – based on the **Definition** – the route choice set of each station-to-station OD pair (in total 24 OD pairs, see **Table 5-15**) were compared by applying the route choice set generation algorithm with the appropriate considerations for the OD specific attribute cut-offs (cf. **Section 5.3.3**) on the extended LU inner zone network (cf. **Figure 5-8**).

Results (**Figure 5-21**) show that for all OD pairs the shortest route (in terms of generalised costs) is the direct route (**Jubilee** line). The second shortest route almost for all OD pairs is the option to change to the **Metropolitan** line at **Wembley Park** station, then change back to the **Jubilee** line at **Finchley Road** station. The generalised cost of this route is 1.14-1.21 times as the generalised cost of the shortest route depending on the OD pair. This route is still considered reasonable. The option to change to the **Metropolitan** line at **Wembley Park** station, then change back to the **Jubilee** line at **Baker Street** station has a much higher generalised cost (1.28-1.41 times the generalised cost of the shortest route). This is because the interchange at **Baker Street** station requires walk time between the **Jubilee** and **Metropolitan** platforms, while at **Finchley Road** station the platforms are adjacent.

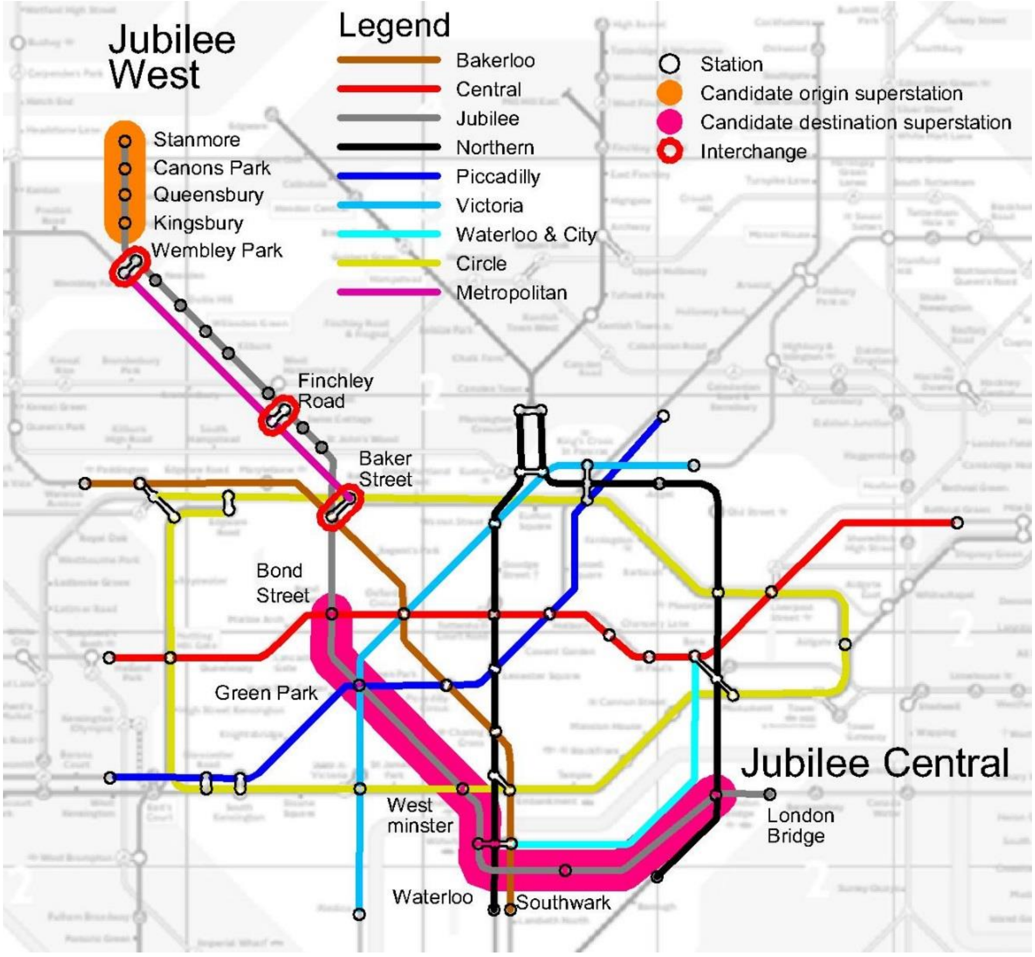


Figure 5-20 Candidate stations for origin and destination superstation, **Case 3**

Table 5-15 OD pairs for which route choice set was compared, **Case 3**

O\D	Bond Street	Green Park	Westminster	Waterloo	Southwark	London Bridge
Stanmore	1	5	9	13	17	21
Canons Park	2	6	10	14	18	22
Queensbury	3	7	11	15	19	23
Kingsbury	4	8	12	16	20	24

It can be easily understood that the route choice set is the same across the OD pairs, whose destination station is the same, but their origin station is different (e.g. OD 1-4). This can be explained with the fact that as the four origin stations (**Stanmore**, **Canons Park**, **Queensbury** and **Kingsbury** stations) are before the first interchange station; passengers having their origin at these stations have similar route choice patterns. Therefore these four stations can be grouped as the origin superstation (**Rule 1**), which is named **Jubilee West** (see **Figure 5-22**).

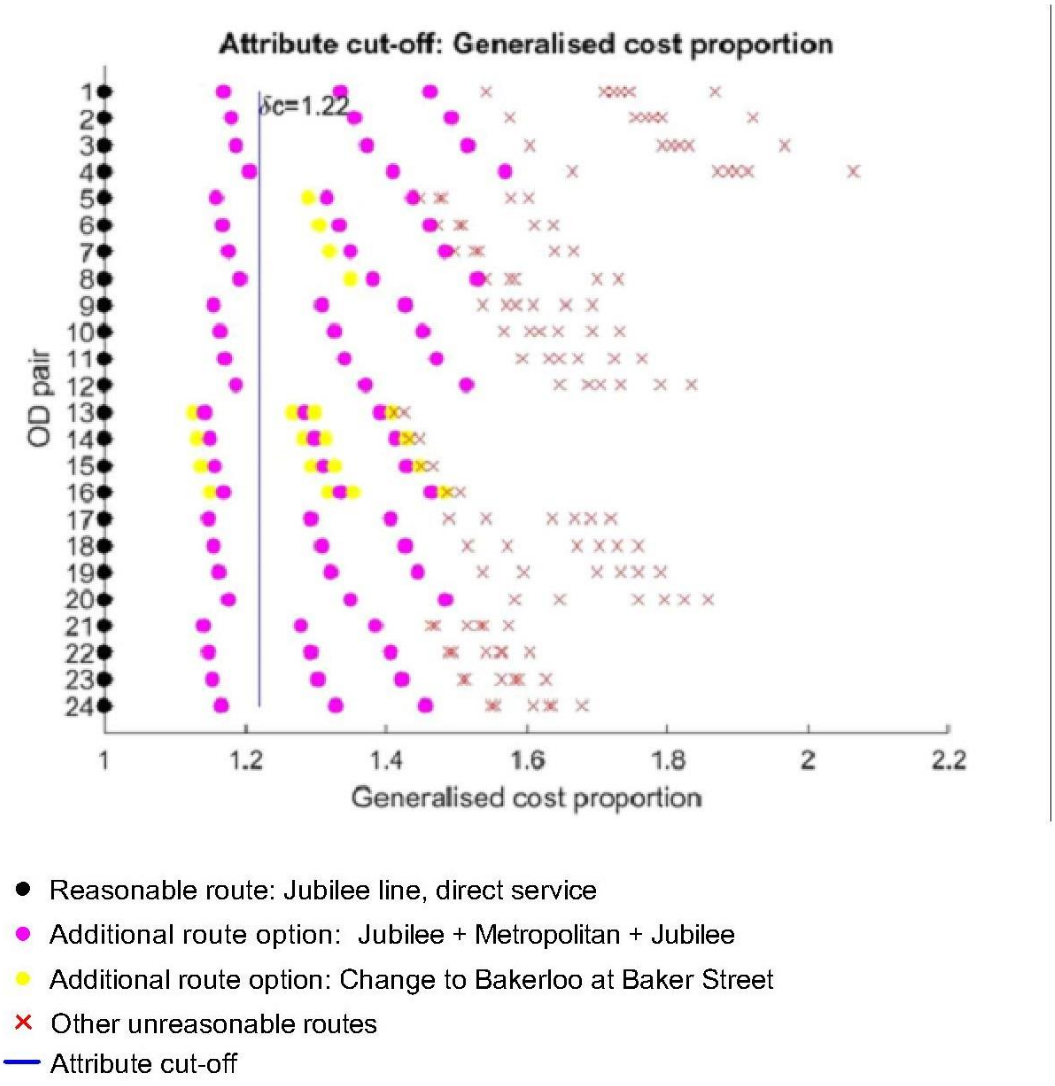


Figure 5-21 Attribute cut-off according to generalised cost proportions, **Case 3**

Regarding the destination stations, the question is more complex as there are many lines and hence route options within the LU inner zone. Results of the route choice set generation algorithm show that the route choice set to **Waterloo** station is different from the route choice set to other stations. This is because from **Baker Street** also the **Bakerloo** line is available to the **Waterloo** station. Due to this, **Waterloo** station is not included in the destination superstation (**Rule 2**).

Results show that the route choice set is the same among the OD pairs with destination at other stations (**Bond Street**, **Green Park**, **Westminster**, **Southwark** and **London Bridge**). Therefore these five stations cold be grouped together as the destination superstation (**Rule 2**), which is named **Jubilee Central** (see **Figure 5-22**). The fact of grouping these five station, without **Waterloo**, illustrates, that even though stations are not consecutive, they can be still grouped as superstations (**Rule 3**)

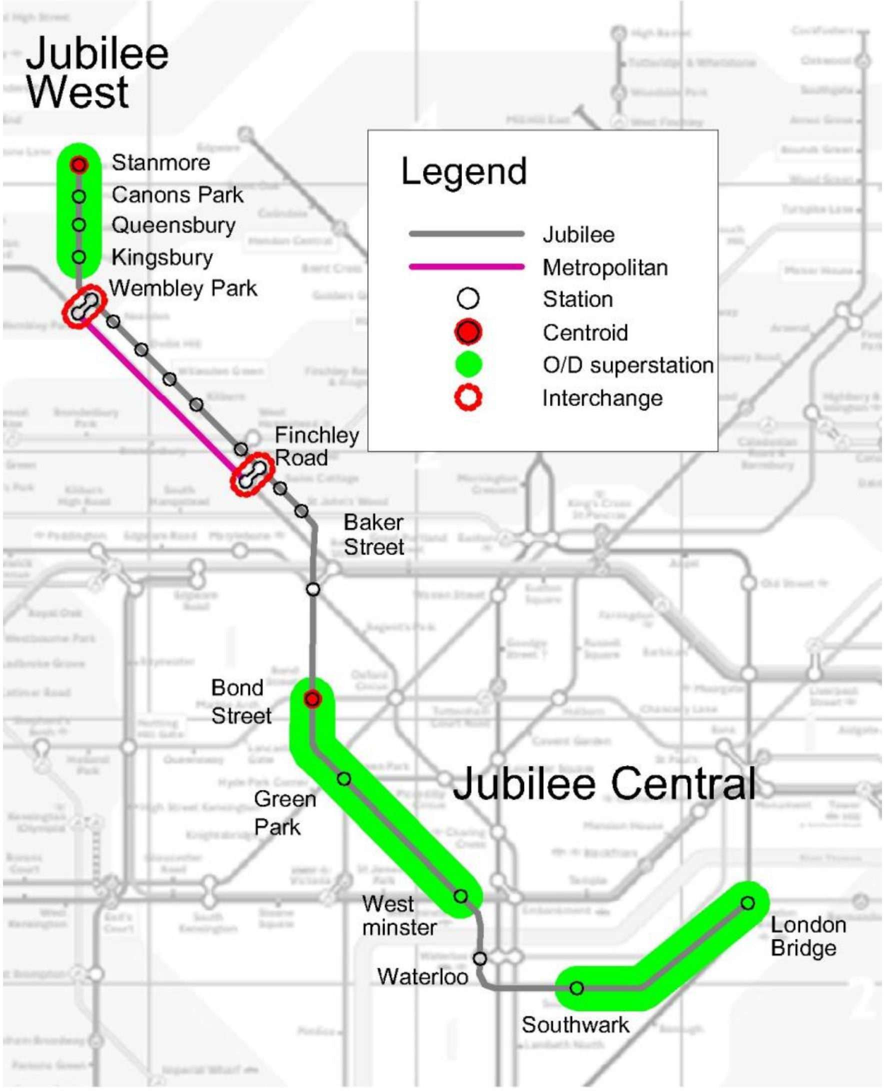


Figure 5-22 The **Jubilee West** – **Jubilee Central** superstation-to-superstation OD pair

For the OD pairs with destination at **Green Park** station, the third shortest route is the option to change to the **Bakerloo** line at **Baker Street** station then to the **Victoria** line at

Oxford Circus station. This option have the advantage that at Oxford Circus station the platforms of the Bakerloo and Victoria lines are adjacent, and the egress time from the Victoria line at Green Park station is lower than from the Jubilee line. However as the generalised cost of this this option is 1.29-1.35 times as the generalised cost of the shortest route, it was not considered as a reasonable route and hence the Green Park station could be grouped with the other four stations as a destination superstation.

In **Case 3** the main OD specific attribute that influences ρc is the presence of a direct route and an express line. As it was elucidated (cf. **Section 4.7.4.2**), when there is direct route for an OD pair, passengers consider other indirect routes reasonable only if they are much better in other attributes. In this specific case, this attribute is the presence of an express line: they can save in average 3 minutes of journey time while the interchanges are still acceptable as they happen between adjacent platforms. For this reason the indirect route could be considered reasonable, even though its generalised cost is 1.14-1.21 times the generalised cost of the shortest route. Based on these considerations, the attribute cut-off (ρc) was set as 1.22 (cf. **Figure 5-21**), which is slightly higher than the value obtained in **Chapter 4** (1.18).

Obtaining Centroid-to-Centroid adjusted Observed Journey Times

For the Jubilee West and Jubilee Central origin and destination superstations, Stanmore and Bond Street stations were selected as the superstation centroid respectively (marked with ● on **Figure 5-22** and highlighted with green in **Table 5-16** and **Table 5-17**, cf. **Section 5.4.1**).

The sample size of OJTs was small for individual OD station-to-station pairs, especially for those having their destination at Southwark (**Figure 5-23**). Therefore a larger and better distributed dataset needs to be obtained (cf. **Section 5.4**).

Table 5-16 Adjustment of OJTs according to on-board and access times at Jubilee West origin superstation, centroids are highlighted with green

Origin	**On-board time**			**Access time**		
Ii	$t^{ob}_{(Ii-1)(Ii)}$ [min]	$t^{ob}_{(I1)(Ii)}$ [min]	$t^{ob}_{(Ii)(\mathrm{Ic})}$ [min]	t^{acc}_{Ii} [min]	$\Delta t^{acc}_{(Ii)(Ic)}$ [min]	
Stanmore		**0**	**0**	**0.6**	**0.0**	**0**
Canons Park	2	2	-2	0.4	0.2	0
Queensbury	2	4	-4	0.6	0.0	0
Kingsbury	3	7	-7	0.5	0.1	0

Table 5-17 Adjustment of OJTs according to on-board and egress times at **Jubilee Central** destination superstation, centroids are highlighted with green

Destination	On-board time			Egress time		
Ii	$t^{ob}_{(Jj-1)(Jj)}$ [min]	$t^{ob}_{(J1)(Jj)}$ [min]	$t^{ob}_{(Jj)(\mathrm{Jc})}$ [min]	t^{egr}_{Jj} [min]	$\Delta t^{egr}_{(Jj)(Jc)}$ [min]	
Bond Street		**0**	**0**	**3.2**	**0.0**	**0**
Green Park	2	2	-2	3.5	-0.3	0
Westminster	2	4	-4	2.7	0.5	1
Southwark	2	6	-6	3.1	0.1	0
London Bridge	2	8	-8	1.6	1.6	2

Following this, the OJTs of the station-to-station OD pairs ($OJT_{(Ii)(Jj)}$) were adjusted to superstation centroid according to equation (5-8) (cf. **Section 5.4.2**). This way the CCOJTs corresponding to each station-to-station OD pair ($CCOJT_{(Ii)(Jj)}$) are obtained. **Table 5-16** presents the necessary adjustments according to the on-board ($t^{ob}_{(Ii)(Ic)}$) and access ($\Delta t^{acc}_{(Ii)(Ic)}$) time at the origin superstation (**Jubilee West**). **Table 5-17** presents the necessary adjustments according to the on-board ($t^{ob}_{(Jj)(Jc)}$) and egress ($\Delta t^{egr}_{(Jj)(Jc)}$) times at the destination superstation (**Jubilee Central**. As the Jubilee line has the same frequency and hence wait time across all the stations of the origin and destination superstation, there is no need to do adjustments according to wait time ($\Delta t^{wait}_{(Ii)(Ic)}$). All adjustments values are rounded to the nearest minute, as the OJTs from Oyster data are given with that precision.

The CCOJTs corresponding to each station-to-station OD pair ($CCOJT_{(Ii)(Jj)}$) were aggregated (cf. **Section 5.4.3**), resulting in a dataset ($CCOJT_{IJ}$) with larger sample size ($n^{CCOJT}_{IJ} = 286$) (**Figure 5-24**).

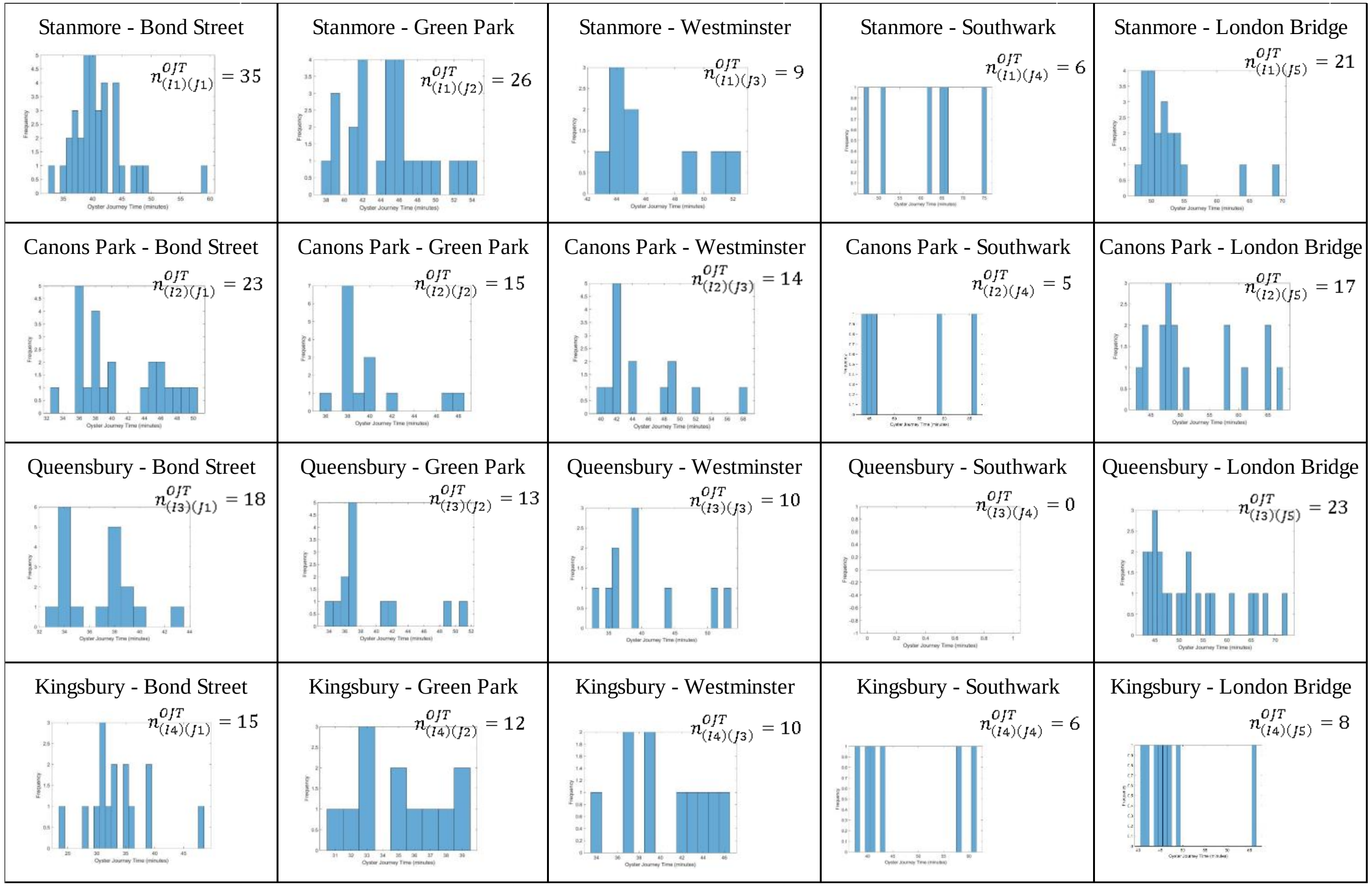


Figure 5-23 Distribution of Observed Journey Times (OJT) for station-to-station OD pairs for **Jubilee West**– **Jubilee Central**

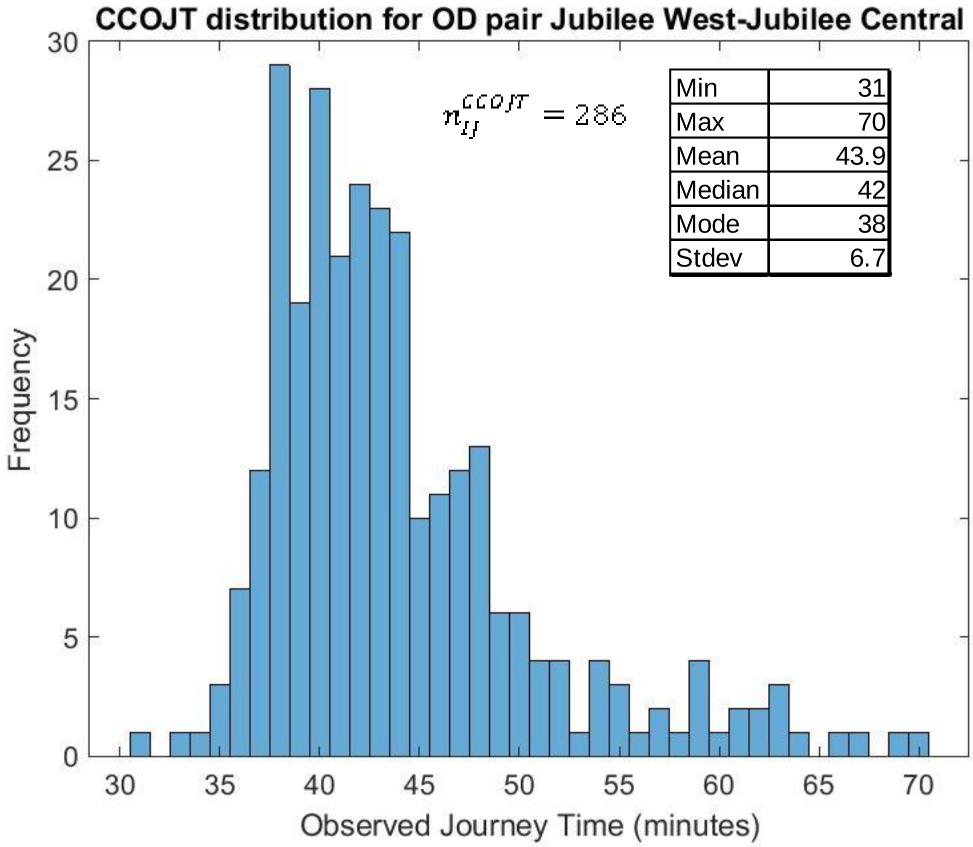


Figure 5-24 Distribution of Centroid-to-Centroid adjusted OJTs for Jubilee West– Jubilee Central

Evaluation of the superstation representation

The finite mixture model presented in **Chapter 3** was applied on this larger dataset of CCOJTs. Within this dataset all entries could be considered as valid data, because the upper outer fence (cf. **Section 3.2.1**) resulted 71 minutes, while the maximum CCOJT value is 70 minutes. This valid dataset is denoted by $CCOJT^0$.

Based on the results of the route choice set generation algorithm (**Figure 5-21**), the Jubilee West – Jubilee Central superstation-to-superstation OD pair was modelled with two reasonable routes, hence route choice was estimated as a two-component ($N_R = 2$) finite mixture distribution. Therefore, the K-means clustering algorithm was applied on the OJT^0 dataset with two clusters and with the settings described in **Section 3.3.1** to produce the initial values for the EM algorithm. Using these initial values, the EM algorithm was run with different settings for the tolerance threshold (cf. **Section 3.3.2**). The more detailed results of initial values and tolerance thresholds are reported in **Appendix F**.

From there it is understood, that although the EM algorithm converges to slightly different values for the two seeds when the tolerance threshold is set greater; at a smaller tolerance threshold, they converge to a similar value for both seeds: 41.5 minutes of mean journey time and 78.3% of proportion for component 1 at the tolerance threshold of 1e-08 (**Figure F-7** and **Figure F-8**). Similar properties could be observed for the other mixture component (labelled with $r = 2$). The log-likelihood exhibits a jump between the tolerance threshold of 1e-02 and 1e-03 (**Figure F-9**). Due to these considerations, the finite mixture model was applied with the tolerance threshold of 1e-08 (**Table 5-18**).

Table 5-18 Finite mixture model results; with Seed = 1, Tolerance threshold = 1e-08 for **Jubilee West**– **Jubilee Central**

Label	Mixture model		
r []	$\mu_{r,IJ}^{MIX}$ [min]	$\sigma_{r,IJ}^{MIX}$ [min]	$\omega_{r,IJ}^{MIX}$ [%]
1	41.5	3.6	78.5%
2	52.9	7.8	21.5%

Following this, the results of the finite mixture model were matched with the actual LU routes (cf. **Section 3.4.1**). The total journey times ($t_{k,IJ}^{SJT}$) of the actual LU routes between the superstation centroids: **Stanmore** and **Bond Street** were calculated based on equation (3-13). The results are presented in **Table 5-19**. The wait time at the origin stations and at the second interchange station (**Finchley Road** station) was considered according to equation (3-15). However at the first interchange station (**Wembley Park**), a different consideration was made: As the interchange happens between adjacent platforms and the station is above ground, passengers travelling on the **Jubilee** line may decide according to the following strategy: If they see that a **Metropolitan** service is approaching, they choose to interchange hoping that they will save time with the express service, otherwise, they stay on the **Jubilee** line as they are not likely to save time. For this reason, it is assumed that passengers on the route with the express service have shorter wait time than specified in equation (3-15): 1 minute. Considering the indirect route with shorter wait time makes it possible to model the two routes with a greater difference in their journey time (3 minutes). The mixture components were matched with the actual LU routes in order of their journey times.

Table 5-20 compares the mixture results for the Jubilee West – Jubilee Central superstation-to-superstation OD pair with the corresponding values of the actual LU routes. In this case the OJT sample size of the station-to-station OD pairs was so small (e.g. 35 OJTs for **Stanmore** – **Bond Street**) that the EM algorithm could not converge as it created an ill-conditioned covariance at iteration 1. Therefore it could not serve for comparison. As Fu (2014) did not examine this OD pair in his work, also that could not be used for validation. **Figure 5-25** presents the probability density functions of the mixture distribution fit on the CCOJT dataset as well as of the mixture components matched with the actual LU routes.

Based on these results, the following was observed: Matching the mixture components with the actual LU routes in order of their journey times it turned out, that the mixture component with lower mean (41.5 minutes) was matched with the indirect route (**Jubilee** – **Metropolitan** – **Jubilee** via **Wembley Park** and **Finchley Road**, 33.3 minutes). However that mixture component has a higher proportion (78.5%), while RODS data (cf. **Section 3.6.3**) shows, that only 11% of the passengers have chosen the indirect route. Another crucial point is that, the mixture component with the higher mean (52.9 minutes) it is much higher than the travel time of the routes understood from timetables (**Section 3.6.2**).

One explanation for these counterintuitive results is, that the higher OJT observations (i.e. 50-70 minutes) could be also attributed to a fail-to-board event on the **Metropolitan** line at **Wembley Park** station. Apart from that – due to the longer distance between the origin and destination – many other events may occur that affect the reliability of travel times, such as service delays or longer walk times at crowded stations. Furthermore, as the RODS data have been collected over several years, it may not reflect the same time period as the journey times understood from timetable.

Looking at the route choice problem from the practical prospective; even though the results in **Table 5-20** show that on average passengers could save 3 minutes of travel time, by taking the indirect route; in this specific case, travel time can be saved, only if they can manage to change back at **Finchley Road** to the **Jubilee** train ahead of the one which they got off[23]. As this question would require schedule based approach, it is beyond the scope of this thesis.

[23] This question is more complex as some of the **Jubilee** trains start from **Wembley Park**, **Willesden Green** or **West Hampstead**.

Table 5-19 Journey time of actual London Underground routes between superstation centroids for **Jubilee West**– **Jubilee Central**

Route						Journey Time [min]										
k	$l=1$	$s=1$	$l=2$	$s=2$	$l=3$	$t_{1,k}^{acc}$	$t_{1,k}^{wait}$	$t_{1,k}^{ob}$	$t_{1,k}^{ic}$	$t_{2,k}^{wait}$	$t_{2,k}^{ob}$	$t_{1,k}^{ic}$	$t_{2,k}^{wait}$	$t_{2,k}^{ob}$	$t_{2,k}^{egr}$	t_{k}^{SJT}
1	**Jub**					0.6	1.5	31.0							3.2	**36.3**
2	**Jub**	**Wembley Park**	**Met**	**Finchley Road**	**Jub**	0.6	1.5	11.0	0.0	1.0	7.0	0.0	1.0	8.0	3.2	**33.3**

Table 5-20 Matching mixture model results with the actual London Underground routes for **Jubilee West**– **Jubilee Central**

Purple: Mixture results, superstation OD pairs, **Green**: actual LU routes

Mixture Label	Journey Time (min)		Route Choice (%)		Route Label	Route Matched				
	Mixture	Timetable	Mixture	RODS		Line 1	Interchange 1	Line 2	Interchange 2	Line 3
r	$\mu_{r,IJ}^{MIX}$	$t_{k,IJ}^{SJT}$	$\omega_{r,IJ}^{MIX}$	$\omega_{k,IJ}^{RODS}$	k	$l=1$	$s=1$	$l=2$	$s=2$	$l=3$
1	41.5	33.3	78.5%	11.0%	2	Jub	Wembley Park	Met	Finchley Road	Jub
2	52.9	36.3	21.5%	89.0%	1	Jub				

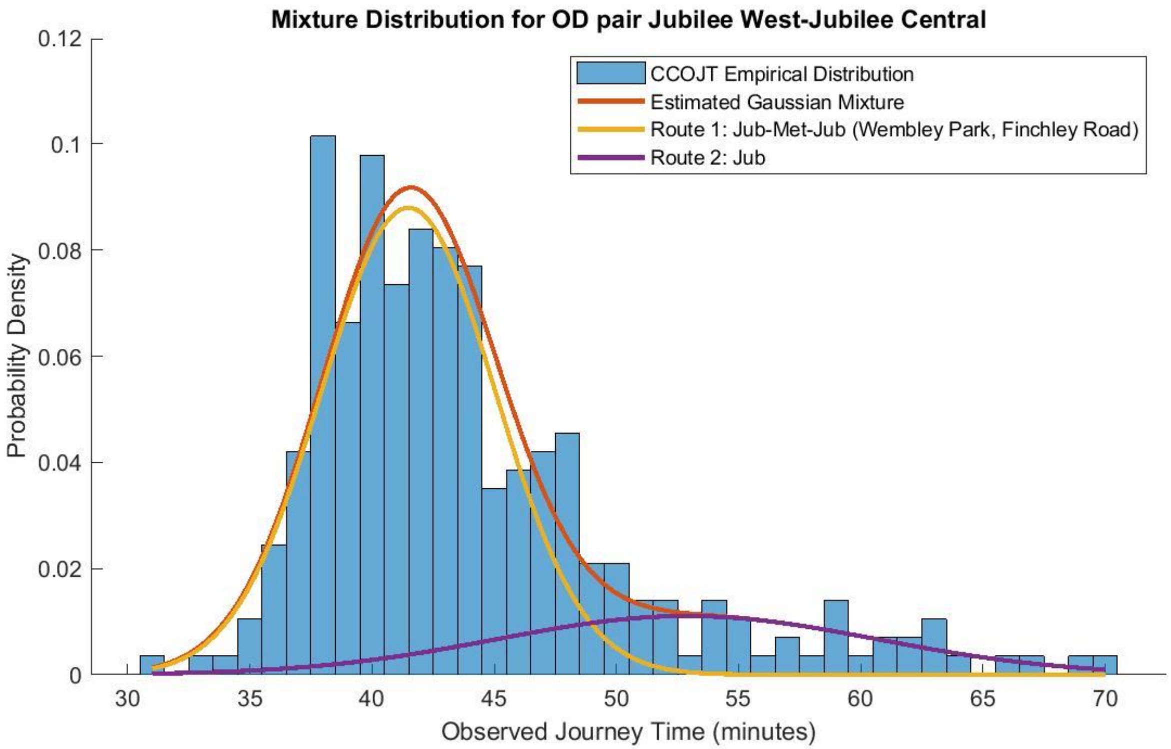


Figure 5-25 Estimated (Gaussian) journey time distribution of the routes for **Jubilee West** – **Jubilee Central**, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

5.7 Discussion and summary

5.7.1 Creating superstations

The main objective in this chapter was to overcome the issue of data availability raised in **Section 3.8.2**: As in many cases only a small OJT sample is available for station-to-station OD pairs, the finite mixture model applied on that dataset is unable to give reliable results. For this, it was proposed to group those stations from/to which passengers are expected to have similar route choice patterns. This group of stations are called superstations in this thesis.

Literature review in this section has shown that, although there have been existing studies where the concept of working with groups of stations was examined, they proposed this idea for the objective of getting a better understanding on the OD demand matrix of a public transport network. In order to comply this objective the vast majority of these studies dealt with grouping stations in the proximity of activity locations or transport hubs. The novelty in this thesis is to introduce the concept of working with groups of stations for a different objective: to overcome the data availability issues for route choice estimation. As the objective is different, the methodology of existing studies could not be applied straightforward, but new rules were set for station grouping.

Creating origin and destination superstations is equivalent to group the OD pairs with similar route choice patterns. For this, the route choice set generation model (cf. **Section 5.3.3**) was applied to find the OD pairs with the same route choice set. The novelty of the route choice set generation algorithm in this chapter with respect to the one in **Chapter 4** is that instead of applying a single cut-off value (i.e. generalised cost proportion of 1.18), it was discussed how the cut off values vary in function of certain OD specific attributes. Through the case studies (cf. **Section 5.6**) it was understood that a lower attribute cut-off is expected for OD pairs with longer minimum journey time as well as for those which have more available directions at the origin and/or destination station (**Case 2**). On the other hand, a higher attribute cut-off is expected when the choice of passengers is between a direct route and an express line (**Case 3**). Based on this, the number of reasonable routes was also identified, which is equivalent to the number of mixture components.

In these case studies 4-5 stations could be grouped as origin or destination superstations, which are not necessarily consecutive stations. **Case 2** exhibited a specific example of grouping 20 stations as the origin superstation (i.e. the **Central East Outer** superstation on the **Central** line, before the first interchange station: **Stratford**). In **Case 3**, both for the origin and destination stations could be grouped as superstations: 4 and 5 stations respectively. This gave the benefit to aggregate the adjusted OJTs of 4*5=20 station-to-station OD pairs.

Until now the superstation representation was examined only for certain OD pairs. In order to extend this concept for network problems it would be necessary to find all OD pairs, for which the superstation representation is applicable. As the LU network is very complex, this could not be done just by looking at the map, but an appropriate algorithm is required (see **Section 8.2.2**).

5.7.2 Obtaining Centroid-to-Centroid adjusted Observed Journey Times

Once the OD pairs with similar route choice patterns were grouped, their OJTs were adjusted to superstation centroids and then aggregated. This way a larger and better distributed sample of CCOJTs were obtained. Depending on the case study OD pair, the superstation representation could increase the number of available observations in different manner.

In **Case 1**, 54 OJT records were available for the **Victoria** – **Holborn** station-to-station OD pair. Grouping the **Victoria** station with 3 other stations, this sample size could be increased to 147 CCOJT records (2.72 times) for the **Victoria South** – **Holborn** superstation-to-superstation OD pair. The great benefit in this case that also the **Brixton** station could be included in the superstation, from which a larger number of OJT records (77) were available.

In **Case 2**, 30 OJT records were available for the **Liverpool Street** – **Green Park** station-to-station OD pair. Here, station grouping could increase this sample size only to 47 CCOJT records (1.57 times) for the **Central East** – **Green Park** superstation-to-superstation OD pair as there were not too many OJT records available from other stations.

In **Case 3**, the great advantage was, that both the origin and destination stations could be grouped as superstations; and hence while only up to 35 OJT records were available for station-to-station OD pairs, the sample size of CCOJT records could be increased to 286 for the **Jubilee West** – **Jubilee Central** superstation-to-superstation OD pair.

The obvious advantage of the spatial aggregation of OJT records was to demonstrate that a larger and better distributed dataset can be achieved also when only the open data is available for the proposed research.

5.7.3 Application of the finite mixture model and evaluation of the superstation representation

The data availability issues with the finite mixture model (cf. **Section 3.8.2**) in case of very few or not well distributed OJT data sample meant that the EM algorithm could not converge as it created an ill-conditioned covariance. This occurred for some of the station-to-station OD pairs of **Case 3**. Obtaining a larger sample of CCOJTs for the superstation-to-superstation OD pairs could overcome this issue as the EM algorithm could converge for all three case study OD pairs.

Another issue with the finite mixture model was, that it may converge to more possible solutions depending on the initial value (seed of the random number generator) and tolerance threshold; and often these values are quite far from each other. For example, when looking at the **Victoria** – **Holborn** station-to-station OD pair in **Case 1**, when the tolerance threshold was set to 1e-05, the proportion of component 1 (ω_1^{MIX}) resulted 79.8% for seed 1 and 35.0% for seed 2. The same ($\omega_{1,(Ii)(Jj)}^{MIX}$) for the **Victoria South** – **Holborn** superstation-to-station OD pair resulted 75.8% and 63.8% respectively. This shows that the larger sample of CCOJTs gave a larger and better distributed dataset, and applying the finite mixture model on that dataset could give more stable route choice results.

Comparing the mixture results with the actual LU routes it was understood that with the superstation representation the estimated mixture component proportions are closer to the expected (RODS) route choice proportions of the LU routes, than it was for the station-to-station OD pairs. For example, in **Case 2**, the mixture model gave the results of 93.3% and 6.7% for the proportion of the two components for the **Liverpool Street** – **Green Park** station-to-station OD pair, while the same for the **Central East** – **Green Park** superstation-to-station OD pair was 80.3% and 19.7% respectively. Comparing this to the RODS results (81.2% and 18.8%) showed that with the superstation representation the route choice estimates were closer to the expected values.

However, in all cases the mean journey times of mixture component 2 are much higher than the total journey time of the corresponding route (e.g. 52.9 minutes vs 36.3 minutes for **Case 3**). A possible explanation for this could be, that the higher OJT observations could be also attributed to a fail-to-board event, not necessarily to the longer route.

As fail-to-board delays often occur in the LU network during the AM peak, and as it has a considerable impact on the finite mixture model results; this question will be further analysed in **Chapter 6** and **Chapter 7**.

Chapter 6
Consideration of fail-to-board delays at the origin station

6.1 Introduction

In **Chapter 3**, following the concept in Fu (2014) the connection was established between the Observed Journey Time (OJT) of a passenger and his/her chosen route; and hence the finite mixture model was formulated to estimate route choice from smart card data. The model was initially applied for station-to-station OD pairs, and then – to overcome the issue of data availability – extended for superstation-to-superstation OD pairs in **Chapter 5**.

In those model specifications, capacity constraints have not yet been explicitly considered. However, it is obvious that capacity constraints do influence both the journey time (i.e. strict capacity constraints: fail-to-board delays) and route choice preferences (i.e. soft capacity constraints: discomfort due to congestion). In the context of the London Underground (LU), this issue is quite relevant as certain stations and lines are extremely congested in the AM peak (Schmöcker et al., 2008).

This chapter addresses the issue how capacity constraints can be included in the process of OJT adjustment focusing on strict capacity constraints: fail-to-board delay. In **Section 5.4**, the methodology for OJT adjustment to superstation centroids was presented. There, it was not yet taken into consideration that at different origin stations passengers may experience different fail-to-board delays. Therefore aggregating OJTs that contain different fail-to-board delays could bring bias into the Centroid-to-Centroid adjusted OJT (CCOJT) dataset and to the estimated route choice results.

To address this issue the rest of the chapter is structured as follows: Firstly, in **Section 6.2** the modelling approaches for capacity constraints are reviewed; then in **Section 6.3** the distribution of fail-to-board delays at a platform of a metro station is estimated with the selected method. The methodology for the adjustment of OJTs according to fail-to-board delays is presented in **Section 6.4**.

Following this, the proposed method is applied on the case study OD pairs. **Section 6.5** describes the data sources for understanding fail-to-board delays, and the case studies are presented in **Section 6.6**. **Section 0** concludes the chapter with the evaluation of the applied methodologies, obtained results and with the issues raised for further research.

6.2 Literature review on considering capacity constraints in route choice estimation

As elucidated in **Section 2.2**, route choice estimation is a sub-model in Transit Assignment Models (TAM) (**Figure 2-1**). In case passenger flows are below the capacity of transit links (i.e. TAMs without capacity constraints), route choice does not depend on the passenger flow; and hence it can be solved as distinct OD problems. However, when passenger flow is near or above the capacity of transit links (i.e. TAMs with capacity constraints), route choice is a function of passenger flows. As also passenger flows are resulting from the route choice of different OD pairs, it cannot be solved any more as distinct OD problems, but it should be treated as a network problem (Gentile and Noekel, 2016). This section reviews TAMs starting from the basic approaches, then focusing on how capacity constraints and the dynamics of the passenger flow are considered in different modelling approaches.

In the recent decades, the problem of transit assignment has been widely explored, especially regarding the consideration of capacity constraints (Fu et al., 2012). The two pillars of this vast amount of literature are the two main modelling approaches: One approach is the frequency-based (Nguyen and Pallottino, 1988; Spiess and Florian, 1989), where each line segment is represented as a link in the network model; and the frequency of the lines can be interpreted as a type of link cost. The other approach is the schedule-based approach (Tong, 1986; Hickman and Bernstein, 1997; Tong and Wong, 1999; Nuzzolo et al., 2001), where, each vehicle-run is represented as a link in the network model according to the schedule (space time graph).

In order to include capacity constraints within the frequency-based approach, Cea and Fernández (1993); Wu (1994); Cominetti and Correa (2001); Cepeda et al. (2006) worked with the concept of effective frequencies; while Kurauchi et al. (2003) introduced the concept of fail-to-board probabilities. Within the schedule-based approach the consideration of capacity constraints is more straightforward, because passengers failing to board can be simply assigned to the next vehicle-run link (Nuzzolo et al., 2012), however this requires a more detailed representation of the network.

To close this gap between the frequency and schedule-based TAMs, (Schmöcker et al., 2008) proposed the quasi-dynamic frequency-based transit assignment models, where the whole modelling period (e.g. AM peak) is divided to smaller time intervals (e.g. 15 minutes) and passengers failing to board in one time period are assigned to the next one. Meschini et al. (2007) used the frequency-based approach in the context of dynamic traffic assignment and applied it on multimodal (road and public transport) networks. Teklu (2007) developed a stochastic process approach to include strict capacity constraints with day-to-day dynamics in the model

Apart from the above discussed strict capacity constraints, when passengers fail to board, they also experience discomfort, even when they are able to board, but they travel on crowded trains without getting a seat (Schmöcker et al., 2011). This discomfort can be expressed by the crowding multiplier, which was evaluated for various metro networks around the world (Whelan and Crockett, 2009; Hörcher et al., 2017; Tirachini et al., 2017). These studies found that overcrowded trains can cause significant increase in the generalised costs of routes: up to 1.7 times for sitting and up to 2.2 times for standing passengers.

In this chapter the quasi-dynamic frequency-based approach (Schmöcker et al., 2008) is selected to estimate boarding and fail-to-board flows. Although the schedule-based approach (Nuzzolo et al., 2001) could give explicit estimates of fail-to-board delays as they model it at the level of individual trains; the quasi-dynamic frequency-based approach fits better the previously set objectives of route choice estimation (cf. **Section 2.4**) and is consistent with the available dataset:

In **Section 2.4** it was established that the one of the main objectives of this thesis is to explore, at what extent route choice can be modelled with smart card data at the OD level, in a frequency based context, without the need of going down to the level of individual passengers and trains. In line with this objective, the finite mixture model (Fu, 2014) was chosen for route choice estimation, which uses only the OJT distribution of passengers, but not their individual smart card records. Therefore, following this logical stream would imply, that if route choice is estimated with the finite mixture model in the frequency-based context, also fail-to-board delays should be modelled accordingly.

The data on line loads and station flows (cf. **Section 6.5.1**) is given for each 15 minute time interval, but the actual flow on each individual train is unknown. Therefore applying the quasi-dynamic frequency-based approach with the same time interval duration as the available data, would be its most straightforward application.

6.3 Inferring distribution of fail-to-board delays at platforms

In this chapter – in addition to what was presented in the previous chapters – the following notation is used:

Variable identifiers

t	Index of (a 15 minute) time interval

Variables

As the problem of fail-to-board delays is formulated for one station platform, the variable identifiers according to lines and stations are omitted in this chapter.

$q_t^{run\_in}$	On-board flow from the previous station in time interval t (passengers/15 minutes)
q_t^{dwell}	Dwell flow in time interval t (passengers/15 minutes)
q_t^{alight}	Alighting flow in time interval t (passengers/15 minutes)
q_t^{egr}	Egress flow in time interval t (passengers/15 minutes)
q_t^{acc}	Access flow in time interval t (passengers/15 minutes)
q_t^{wait}	Wait flow in time interval t (passengers/15 minutes)
q_t^{board}	Boarding flow in time interval t (passengers/15 minutes)
q_t^{fail}	Fail-to-board flow in time interval t (passengers/15 minutes)
$q_t^{run\_out}$	On-board flow to the next station in time interval t (passengers/15 minutes)
TID	**T**ime **I**nterval **D**uration (15 minutes)
κ	Capacity of trains (passengers/train)
p_t^{fail}	Fail-to-board probability in time interval t
t_t^{fail}	Average fail-to-board delay in time interval t (minutes)
δ^{fail}	Random variable of fail-to-board delay

τ	A possible outcome of fail-to-board delay, rounded to integer minutes $\tau = \{1, 2, \ldots, TID\}$
T	Total number of (15 minute) t time intervals within the AM peak
$\varphi^{\tau}_{(Ii)(Ic)}$	Index of a record within the sorted $OJT_{(Ii)(Jj)}$ sample, according to the proportion of the distribution of fail-to-board delays τ.
$OJT^{\tau}_{(Ii)(Jj)}$	Subset of $OJT_{(Ii)(Jj)}$ according to the proportion of the distribution of fail-to-board delays τ
$OJT^{\tau,fail}_{(Ii)(Jj)}$	$OJT^{\tau}_{(Ii)(Jj)}$ dataset adjusted with the fail-to-board delay of τ minutes
$OJT^{fail}_{(Ii)(Jj)}$	All $OJT^{\tau,fail}_{(Ii)(Jj)}$ subsets aggregated and sorted
$CCOJT^{fail}_{IJ}$	$OJT^{fail}_{(Ii)(Jj)}$ adjusted to superstation centroid and aggregated
$CCOJT^{fail,0}_{IJ}$	Valid dataset within $CCOJT^{fail}_{IJ}$
ψ^{τ}_{t}	Dummy variable,ψ^{τ}_{t}=1 if $t^{fail}_{t} = \tau$, otherwise ψ^{τ}_{t}=0

In this section, firstly the representation of a station platform is explained (see **Section 6.3.1**), which is followed by the application of the quasi-dynamic frequency based approach (cf. **Section 6.2**) to calculate boarding and fail-to-board passenger flow at each platform (see **Section 6.3.2**). From this, the distribution of fail-to-board delays in a time period can be inferred (see **Section 6.3.3**).

6.3.1 Representation of a station platform

In **Section 4.4.1**, the representation of a metro network with nodes and links was described (cf. **Figure 4-3**). That network representation was used for calculating the generalised costs of routes and for pathfinding in models without capacity constraints. Within that context a station platform could be represented with two nodes (i.e. on-board and platform); and all the relevant time components (i.e. on-board and wait) could be defined as the links connecting these nodes. However, when working with passenger flows and considering strict capacity constraints (i.e. fail-to-board events), the definition of additional nodes are required (Schmöcker et al., 2008) (**Figure 6-1**):

- *Arrive node*: Represents the arrival of the train, here passengers decide to alight or remain on the train
- *Platform node*: Represents passengers waiting for trains
- *Attempt node*: Represents passengers attempting to board the train, the outcome of this event can be either boarding or fail to board
- *Depart node*: Represents the departure of the train, here passengers already on train join with the newly boarded passengers

These nodes are connected with the following links, each of them having their corresponding link flows:

Run-in link ($q_t^{run\_in}$) Represents passengers on-board the train, from the depart node of the previous station to the arrive node of the current station

Dwell link (q_t^{dwell}) Represents passengers staying on train, from the arrive node to the depart node

Alight link (q_t^{alight}) Represents alighting passengers, from the arrive node to the platform node

Egress link (q_t^{egr}) Represents passengers egressing the platform, including also those who interchange to other lines

Access link (q_t^{acc}) Represents passengers accessing the platform, including also those who interchange from other lines

Wait link (q_t^{wait}) Represents passengers waiting at the platform, from platform node to attempt node

Board link (q_t^{board}) Represents boarding passengers, from attempt node to depart node

Fail link (q_t^{fail}) Represents fail-to-board passengers, from attempt node back to the platform node

Run-out link	($q_t^{run\_out}$)	Represents passengers on-board the train from the depart node of the current station to the arrive node of the next station

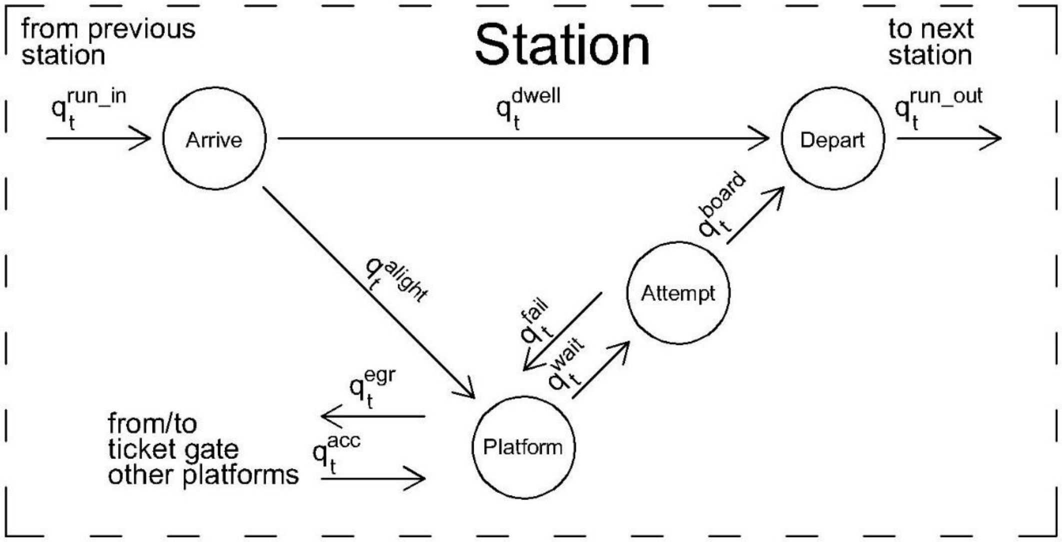


Figure 6-1 Representation of a station platform with one metro line, based on (Schmöcker et al., 2008)

The reason why here the *arrive* and *depart* nodes are distinguished is that for the calculation of fail-to-board flows it is necessary to know the passengers who stay on-board (i.e. *dwell* flow, q_t^{dwell}) after the others alighted (see **Section 6.3.2**) and this can be represented as a link between the *arrive* and *depart* node. This would not be possible with the earlier network representation (**Figure 4-3**), where a single on-board node was considered. For similar purposes, instead of having a single *platform* node, an additional attempt node was considered to distinguish the boarding (q_t^{board}) and fail-to-board (q_t^{fail}) flows.

6.3.2 Boarding and fail-to-board flows

In the quasi-dynamic frequency-based context, capacity problems occur in time interval t, when (**Figure 6-1**):

$$q_t^{acc} + q_t^{dwell} > \frac{\kappa \cdot f}{4} \tag{6-1}$$

where the access flow (q_t^{acc}) and the line capacity ($\kappa \cdot f$) can be understood from the data sources described in **Section 6.5**. It is important to note that as the capacity of trains (κ) is given in the unit of passengers/train and the frequency of trains (f) in trains/hour, it needs to be divided by 4 so that they could be compared to the flows which are given in the unit of passengers/15 minutes.

For the dwell flow (q_t^{dwell}), the following considerations can be made (cf. **Figure 6-1**): As capacity constraints are less likely to affect alighting passengers, it can be assumed that:

$$q_t^{alight} = q_t^{egr} \tag{6-2}$$

Furthermore, at the arrive node:

$$q_t^{dwell} = q_t^{run\_in} - q_t^{alight} \tag{6-3}$$

where the egress flow (q_t^{egr}) and the on-board flow from the previous station ($q_t^{run\_in}$) can be known from the data sources described in **Section 6.5**.

At stations, where capacity problems occur at least in one time interval t (see formula (6-1)), it is necessary to account for boarding and fail-to-board passenger flows. In this thesis the focus is on the AM peak (7:00-10:00), where in the first time interval ($t = 1$), no capacity problems occur, therefore the number of passengers waiting to board is identical to the access flow:

$$q_1^{wait} = q_1^{acc} \tag{6-4}$$

Following the concept in Schmöcker et al. (2008) (cf. **Figure 6-1**), boarding and fail-to-board flows can be calculated as:

$$q_t^{board} = min\left(q_t^{wait}, \frac{\kappa \cdot f}{4} - q_t^{dwell}\right) \tag{6-5}$$

and

$$q_t^{fail} = q_t^{wait} - q_t^{board} \tag{6-6}$$

Passengers who fail to board in time interval t are assigned to the waiting passengers of the next time interval ($t + 1$):

$$q_{t+1}^{wait} = q_{t+1}^{acc} + q_t^{fail} \tag{6-7}$$

Once the flows were calculated for the first time interval, equations (6-2), (6-3), (6-5), (6-6) and (6-7) are repeated for the next time intervals $(t = 2.3, \dots)$ until the last time interval $(t = T)$.

6.3.3 Distribution of fail-to-board delays within a time period

Knowing the boarding and fail-to-board flows in each time interval t, the next step is to infer fail-to-board delays. Based on the literature review in **Section 6.2**, it seemed more straightforward to follow the concept of fail-to-board probabilities (Kurauchi et al., 2003; Schmöcker et al., 2008). In this setting, fail-to-board probability in each time interval t can be interpreted as:

$$p_t^{fail} = \frac{q_t^{fail}}{q_t^{wait}} \tag{6-8}$$

And from that, fail-to-board delays in each time interval t can be inferred as:

$$t_t^{fail} = TID \cdot p_t^{fail} \tag{6-9}$$

Looking into equation (6-9), the connection between fail-to-board probabilities and delays can be understood deeper. In case no passengers fail to board in time interval t $(p_t^{fail} = 0)$, the average fail-to-board delay is zero $(t_t^{fail} = 0)$. On the contrary, if all passengers fail to board in time interval t $(p_t^{fail} = 1)$ the average fail-to-board delay is equal to the duration of the time interval ($t_t^{fail} = TID$, i.e. 15 minutes). For any p_t^{fail} value between 0 and 1 will give the average fail-to-board delay for time interval t proportional to the fail-to-board probability.

With equation (6-9), it is assumed, that no passengers delay more than the duration of time interval (TID). This can be realistic, as the 15 minute duration of a time interval can be considered large enough to exclude the possibility that a passenger would fail-to-board even in the next time interval. This is in line with Schmöcker et al. (2008), who explained the reason for choosing the duration of the time interval 15 minutes to introduce fairly large time intervals for which flows can be assumed relatively constant.

Looking at all T time intervals within the AM peak, the distribution of fail-to-board delays can be understood. Let τ denote a possible outcome of fail-to-board delay, rounded to integer minutes $\tau = \{1, 2, \dots, TID\}$. Furthermore, let:

$$\psi_t^\tau = \begin{Bmatrix} 1, & if\ t_t^{fail} = \tau \\ 0, & otherwise \end{Bmatrix} \tag{6-10}$$

The probability that in the AM peak the fail-to-board delay (t^{fail}) takes up a certain τ value, can be calculated as:

$$Pr(\delta^{fail} = \tau) = \frac{\sum_{t=1}^{T} q_t^{acc} \cdot \psi_t^\tau}{\sum_{t=1}^{T} q_t^{acc}} \tag{6-11}$$

And the cumulative probability for each τ value is:

$$Pr(\delta^{fail} \leq \tau) = \sum_{1}^{\tau} Pr(\delta^{fail} = \tau) \tag{6-12}$$

These probabilities serve as the basis for OJT adjustment (see **Section 6.4**).

6.4 Adjustment of OJTs according to fail-to-board delays

The key assumption for OJT adjustment is that the fail-to-board delays within the OJT records follow the same distribution as the fail-to-board delays understood from RODS data (cf. equation (6-12)). Making this assumption and looking at one station-to-station OD pair $(Ii)(Jj)$, let $\varphi_{(Ii)(Ic)}^\tau$ denote the index of a record within the sorted $OJT_{(Ii)(Jj)}$ sample for which is true, that:

$$\frac{\varphi_{(Ii)(Ic)}^\tau}{n_{(Ii)(Ic)}^{OJT}} = Pr(\delta^{fail} \leq \tau) \tag{6-13}$$

Where $n_{(Ii)(Ic)}^{OJT}$ denotes the sample size of the $OJT_{(Ii)(Jj)}$ dataset. From this $\varphi_{(Ii)(Ic)}^\tau$ can be calculated, and – being the index of an OJT record – is rounded to the nearest integer. Based on this, for each τ outcome of fail-to-board delay, $OJT_{(Ii)(Jj)}^\tau$ can be defined, which is a subset within the $OJT_{(Ii)(Jj)}$ dataset, where the upper limit is the $\varphi_{(Ii)(Ic)}^\tau$ index and the lower limit corresponds to the index of the upper limit of the previous subset plus 1. This way for each record q in the $OJT_{(Ii)(Jj)}^\tau$ subset the adjusted OJT can be calculated as:

$$OJT_{q,(Ii)(Jj)}^{\tau,fail} = OJT_{q,(Ii)(Jj)}^\tau - \tau \tag{6-14}$$

Having obtained the adjusted $OJT_{(Ii)(Jj)}^{\tau,fail}$ values for each subset, they can be aggregated as:

$$OJT_{(Ii)(Jj)}^{fail} = \bigcup_{\tau=1}^{TID} OJT_{(Ii)(Jj)}^{\tau,fail} \tag{6-15}$$

Once $OJT_{(Ii)(Jj)}^{fail}$ was calculated for each station-to-station OD pair, these can be adjusted to superstation centroid and hence aggregated as described in **Section 5.4**, this way obtaining $CCOJT_{IJ}^{fail}$. Following this, the finite mixture model is applied on the $CCOJT_{IJ}^{fail}$ distribution as described in **Section 5.5**. These methods are presented through the case studies in **Section 6.6**.

6.5 Data sources for understanding crowding on board and at platforms

In order to understand fail-to-board delays, it is necessary to obtain further information on the crowding levels on board and at platforms, which is a function of passenger flows (see **Section 6.5.1**) and of line capacities (see **Section 6.5.2**).

6.5.1 Passenger flows

Passenger flow on each line segment (on-board flow) and station passageway (AEI flow) is available from the TfL open data website[24]. It is produced as an output of RODS data (cf. **Section 3.6.3**), which was reconciled to passenger counts.

Knowing these deficiencies of manual surveys, the ideal would be to fully rely on automated data sources in passenger flow modelling: using exclusively smart card data both for the OD demand and for route choice and solving it as a transit assignment model for the whole network (Hörcher et al., 2017). However, this would require to run the model on the whole LU and the connecting rail network (cf. **Section 4.3**), which is beyond the scope of this thesis.

Another option would be to understand crowding from load-weigh data of platforms and trains. In the context of the case study network not all LU lines are equipped with train load weighing systems and none of the platforms have sensors for passenger count. Therefore, it would not be feasible to apply this type of data in the LU.

Knowing the complexity of the problem and the unavailability of load-weigh data in the LU, at this stage of research, it is considered justifiable to use RODS data to gain initial

[24] https://tfl.gov.uk/info-for/open-data-users/

information on crowding. As a further research, this model can be extended for a larger network and can go toward the exclusive use of smart card data.

6.5.2 Line capacities

The capacity of an LU line segment can be calculated as $\kappa \cdot f$. In this setting, the total capacity of the trains on each LU line (κ) can be understood from the rolling stock information available from the TfL website[25]; and the frequency of each LU line (f) is known from timetables (**Section 3.6.2.1**).

6.6 Case studies on the London Underground

This section continues the analysis on the superstation-to-superstation OD pairs presented in **Section 5.6** (cf. **Figure 5-8** and **Table 5-6**) introducing the problem of strict capacity constraints (fail-to-board delay). Firstly, it examines for all the three cases, whether the fail-to-board delay occurs at the origin or at the interchange station. As for **Case 1** and **Case 2**, it occurs at the origin station, the corresponding OJT adjustment is discussed in this section. For **Case 3** – as it occurs at the interchange station – the rest of the case study is discussed in **Section 7.7** (see **Table 6-1**).

Among the first two cases **Case 2** is special as there are three stations where fail-to-board delays occur, at each station with different intensity. Further from the city centre (**Mile End**), the fail-to-board delays are less severe as trains are not completely full. One station before the LU inner zone (**Bethnal Green**), the situation is the most critical, as the trains are already full, there are not many alighting passengers, but there are much more who are willing to board. The **Liverpool Street** station exhibits another type of problem as it is a destination for those who travel to the City of London, but at the same time being also a rail terminus, there are still many passengers who interchange here from other rail services. Therefore, the challenge in this case is to do the OJT adjustment in a way to consider the different intensities of fail-to-board delays at different stations.

[25] https://tfl.gov.uk/corporate/about-tfl/what-we-do/london-underground/rolling-stock

Table 6-1 A summary of the case studies on the LU network, adjustment according to fail-to-board delays

Case	Superstation OD pair		Fail-to-board event			
	Origin *I*	Destination *J*	Type	Line	Station	Max Delay
1	**Victoria South**	**Holborn**	Origin	**Victoria (NB)**	**Victoria**	6
2	**Central East**	**Green Park**	Origin	**Central (WB)**	**Mile End**	7
					Bethnal Green	15
					Liverpool Street	10
3	**Jubilee West**	**Jubilee Central**	Interchange	**Metropolitan (EB)**	**Wembley Park**	7

Case 1: Victoria South - Holborn

Fail-to-board delays at station platforms

Looking at the **Victoria South** – **Holborn** superstation-to-station OD pair (cf. **Figure 5-10**), it is checked whether capacity problems occur at any stations of the origin superstation (**Brixton**, **Vauxhall**, **Pimlico** and **Victoria** stations) as well as at the interchange stations of each reasonable route (**Oxford Circus** and **Green Park** stations). The line capacities for each LU line of the case study OD pair (i.e. **Victoria**, **Central** and **Piccadilly** lines) are calculated from the data described in **Section 6.5.2** (see **Table 6-2**) and compared to the link flows (cf. **Section 6.5.1**) according to formula (6-1).

Table 6-2 Line capacities for **Case 1**, source:
https://tfl.gov.uk/travel-information/timetables/
https://tfl.gov.uk/corporate/about-tfl/what-we-do/london-underground/rolling-stock

Line	**Station**	**Frequency** f [trains/hour]	**Train Capacity** κ [pax/trains]	**Line Capacity** $\frac{\kappa \cdot f}{4}$ [pax/15 min]
Victoria	**Victoria**	35	864	7560
Central	**Oxford Circus**	26	892	5798
Piccadilly	**Green Park**	24	684	4104

Figure 6-2 presents the dwell (q_t^{dwell}) and access (q_t^{acc}) link flows in the most congested time interval of the AM peak (08:30-08:45) for the stations of the **Victoria South** origin superstation. Among these stations, capacity problems occur only at **Victoria** station ($q_t^{dwell} + q_t^{acc} = 1.11 \cdot \frac{\kappa \cdot f}{4}$). At **Vauxhall** and **Pimlico** stations the flow is very near the line capacity (0.93 and 0.97 times respectively), however it was assumed, that all passengers are willing to board until they find available space, therefore there is no need to estimate fail-to-board delays at these stations. At **Brixton** station, being the line terminus, clearly no capacity problems occur ($q_t^{dwell} + q_t^{acc} = 0.39 \cdot \frac{\kappa \cdot f}{4}$). Similarly, no capacity problems occur for boarding at interchange stations (**Oxford Circus** and **Green Park**), where the flow is much below the capacity (0.50 and 0.45 times respectively). This can be easily understood, as these stations are top destinations where most

passengers alight therefore is sufficient space for newly boarding passengers. As the capacity problems occur at the origin station, it is necessary to do OJT adjustments according to fail-to-board delays (cf. **Section 6.4**).

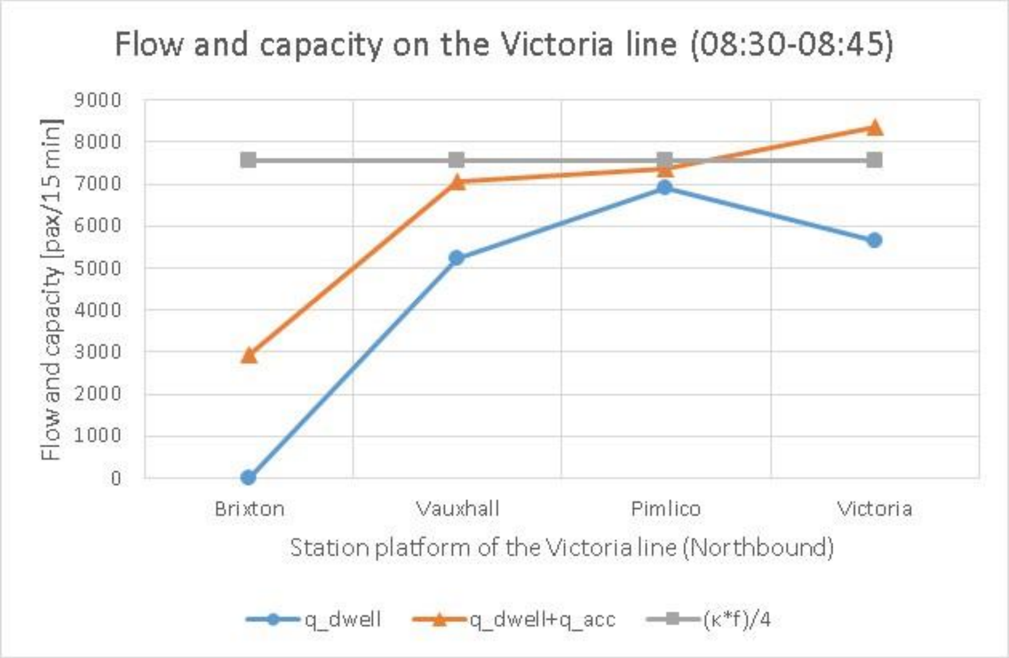


Figure 6-2 Passenger flow and line capacity on the northbound Victoria line, at the stations of the Victoria South origin supertation, peak of peak (08:30-08:45)

Focusing on the station with capacity problems (i.e. Victoria station, Victoria line northbound), q_t^{dwell}, q_t^{board} and q_t^{fail} flows are calculated for each time interval t using equations (6-2)-(6-7). Presenting these flows on **Figure 6-3**, it was understood, that capacity problems occur between 8:15 and 9:15. In this figure the border line between the column of q_t^{board} (grey) and q_t^{fail} (orange) corresponds to the line capacity ($\frac{\kappa \cdot f}{4}$= 7560 passengers/15 minutes).

From these flows p_t^{fail} and t_t^{fail} were calculated in the congested time intervals (8:15-9:15) according to equations (6-8) and (6-9) and presented the results in **Table 6-3**. The value of t_t^{fail} is rounded to integer minutes, because they serve for the adjustment of OJTs, which are also given with the same precision. Looking at the time intervals with capacity problems (8:15-9:15), there is an average of 3 minutes delay at the beginning (8:15-8:30) and at the end (9:00-9:15) of the period of congestion; and an even higher average delay of 6 minutes in between (8:30-9:00) (highlighted with yellow). These results were also presented on a histogram to describe the distribution of fail-to-board delays in the AM peak (**Figure 6-4**).

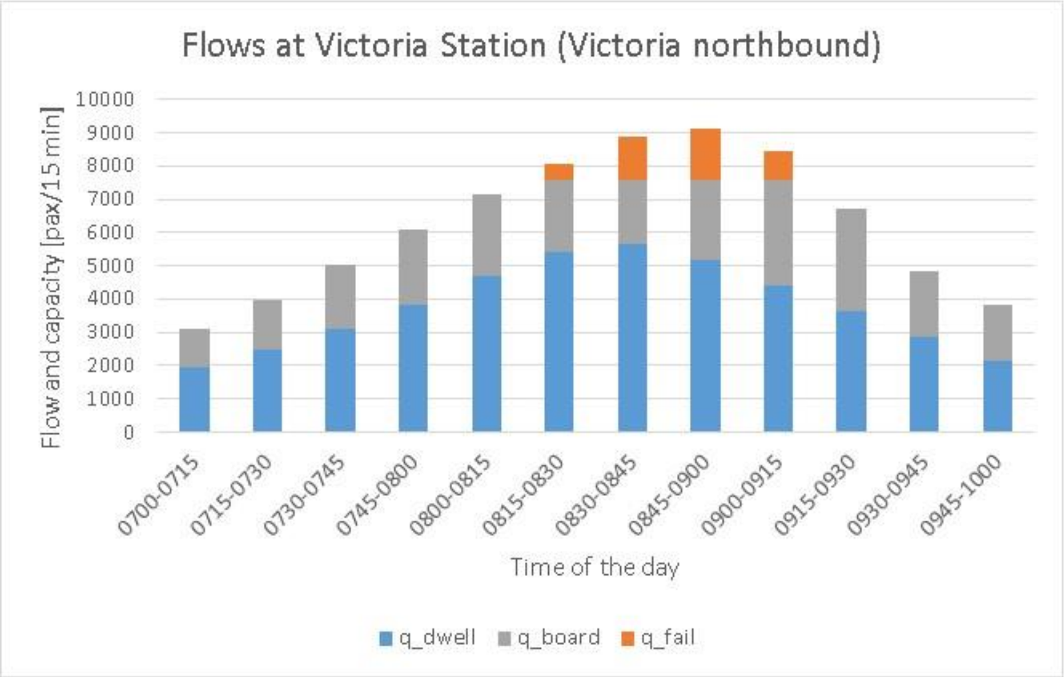


Figure 6-3 Boarding and fail-to-board flows at Victoria station (Victoria line northbound) in the AM peak (7:00-10:00)

Table 6-3 Average fail to board delays at Victoria station (Victoria line northbound) in the congested time intervals of the AM peak (8:15-9:15)

Variable	**Value in time interval *t***					
	8:00-815	8:15-8:30	8:30-8:45	8:45-9:00	9:00-9:15	9:15-9:30
q_t^{dwell}	4691	5438	5637	5183	4421	3630
q_t^{acc}	2477	2628	2732	2628	2453	2227
q_t^{wait}	2477	2628	3238	3943	4019	3107
q_t^{board}	2477	2122	1923	2377	3139	3107
q_t^{fail}	0	506	1315	1566	880	0
p_t^{fail}	0.00	0.19	0.41	0.40	0.22	0.00
t_t^{fail}	0	3	6	6	3	0

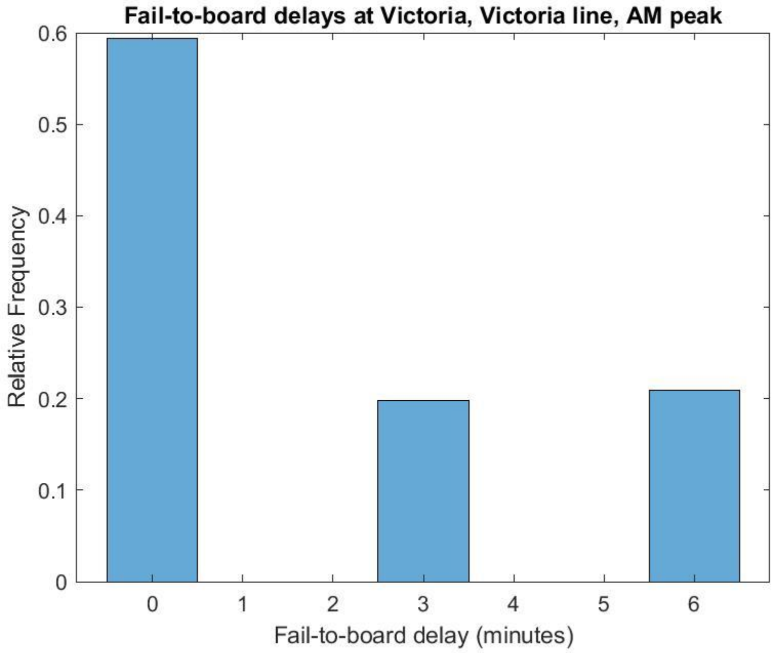


Figure 6-4 Distribution of fail-to-board delays at Victoria station (Victoria line northbound) in the AM peak (7:00-10:00)

Adjustment of OJTs according to fail-to-board delays

Following the methodology in **Section 6.4**, based on the fail-to-board delay distribution (**Figure 6-4**), the $\varphi^{\tau}_{(Ii)(Ic)}$ indices (equation (6-13)) and hence the $OJT^{\tau}_{(Ii)(Jj)}$ subsets as well as their adjustment, $OJT^{\tau,fail}_{(Ii)(Jj)}$ (equation (6-14)) were calculated for each outcome of fail-to-board delay value τ for the station, where capacity problems occur (Victoria station, northbound Victoria line, see **Table 6-4** and **Figure 6-5**).

Table 6-4 Subsets of the OJT dataset according to fail-to-board delays at Victoria station (Victoria line northbound) in the AM peak (7:00-10:00)

τ	$Pr(\delta^{fail} \leq \tau)$	$\varphi^{\tau}_{(Ii)(Ic)}$	$OJT^{\tau,fail}_{(Ii)(Jj)}$	
			min	max
0	0.59	32	12	20
3	0.79	43	20	22
6	1.00	54	22	31

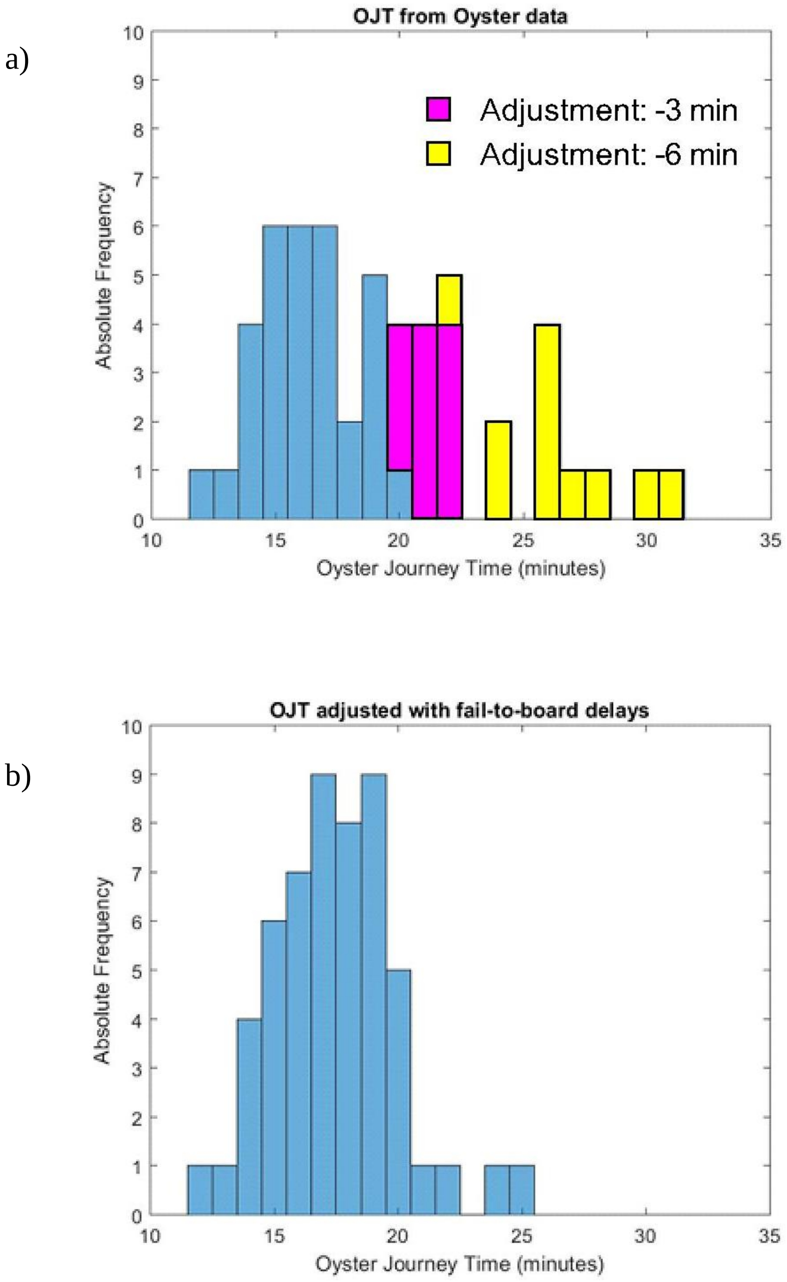


Figure 6-5 Adjustment of OJTs according to fail-to board delays at **Victoria** station a) Original OJTs from Oyster data and proposed adjustments, b) Adjusted OJTs

As no capacity problems occur at other stations of the origin superstation (**Brixton**, **Vauxhall** and **Pimlico**) (cf. **Figure 6-2**), their OJTs remain unchanged. Following this, these OJTs and the OJTs of other stations were adjusted to the superstation centroid (**Victoria** station) and then aggregated spatially (cf. **Section 5.4**). This way $CCOJT_{IJ}^{fail}$ was obtained (**Figure 6-6**).

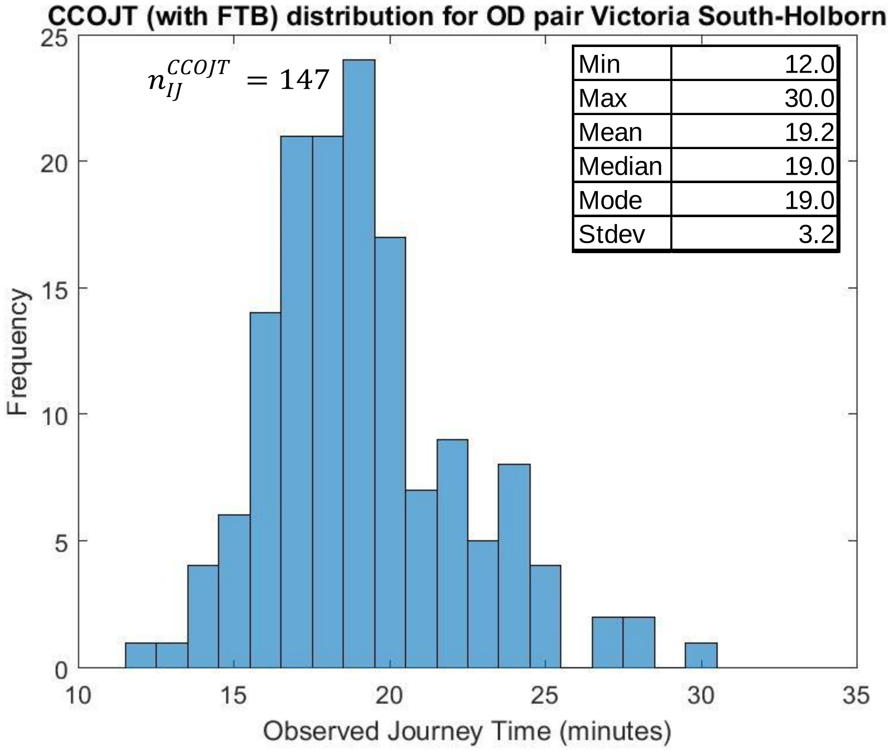


Figure 6-6 Distribution of Centroid-to-Centroid adjusted OJTs considering fail-to-board delays for **Victoria South– Holborn**

Evaluation of the OJT adjustment according to fail-to-board delays

The finite mixture model presented in **Chapter 3** was applied on CCOJT dataset adjusted according to fail-to-board delays. A more detailed description of the settings and of the results are presented in **Appendix G**. Based on that, the chosen settings for the finite mixture model are:

- Seed = 1
- Tolerance threshold = 1e-06

The results with these settings are presented in **Table 6-5**. Following this, the results of the finite mixture model were matched with the actual LU routes (cf. **Section 3.4.1**). **Table 6-6** compares the mixture results for the **Victoria South** – **Holborn** superstation-to-station OD pair with the two types of OJT adjustments:

- Only to superstation centroids (**Chapter 5**)
- To superstation centroid and according to fail-to-board delays (**Chapter 6**)

These are further compared with the earlier results presented in **Table 5-10**. **Figure 6-7** presents the probability density functions of the mixture distribution fit on the CCOJT dataset adjusted according to fail-to-board delays as well as of the mixture components matched with the actual LU routes.

Table 6-5 Finite mixture model results; with Seed = 1, Tolerance threshold = 1e-06 for **Victoria South** – **Holborn**

OJTs adjusted to superstation centroid and according to fail-to-board delays

Label	Mixture model		
r []	$\mu_{r,IJ}^{MIX}$ [min]	$\sigma_{r,IJ}^{MIX}$ [min]	$\omega_{r,IJ}^{MIX}$ [%]
1	17.9	2.0	74.8%
2	22.9	3.0	25.2%

Based on these results, the following was observed: The finite mixture model applied on the CCOJTs of the **Victoria South** – **Holborn** superstation-to-station OD pair adjusted according to fail-to-board delays gave closer results for the mean ($\mu_{r,IJ}^{MIX}$) and for the component proportion ($\omega_{r,IJ}^{MIX}$) to the actual LU values ($t_{k,(Ic)(Jc)}$ and $\omega_{k,IJ}^{RODS}$) than the results of **Chapter 5** (cf. **Table 6-6**). This is because the exceedingly high OJTs attributed to the fail-to-board delays at **Victoria** station were replaced with lower values (cf. **Section 6.4).**

However, it was understood, that the proportion of mixture component 1 ($\omega_{1,IJ}^{MIX}$) exhibited a bigger jump between tolerance thresholds 1e-06 and 1e-07, when it was applied on the $CCOJT_{IJ}^{fail}$ distribution (from 74.8% to 55.9%, cf. **Figure G-2**), than the results of **Chapter 5** (from 70.8% to 67.9%, **Figure F-2**), but it was not as big as the results of **Chapter 3** (from 79.8% to 33.5%, cf. **Figure 3-6**). Therefore, even though the adjustments according to fail-to-board delays can give closer results to the actual LU values for certain settings of the seed and tolerance threshold; it can adversely affect the convergence of the model.

Table 6-6 Matching mixture model results with the actual London Underground routes for Victoria South– Holborn

Blue: Mixture results, adjustment: superstation centroid and fail-to-board delays **Purple**: Mixture results, , adjustment: superstation centroid only, **Red**: Mixture results, station OD pairs**, Yellow**: Fu (2014)**, Green**: actual LU routes

Mixture Label	Journey Time (min)					Route Choice (%)					Route Label	Route Matched		
	Mixture				Timetable	Mixture				RODS		Line 1	Interchange 1	Line 2
	Proposed			Fu		Proposed			Fu					
	FTB	SS	Station			FTB	SS	Station						
r	$\mu^{MIX}_{r,IJ}$		$\mu^{MIX}_{r,(Ii)(Jj)}$		$t^{SJT}_{k,IJ}$	$\omega^{MIX}_{r,IJ}$		$\omega^{MIX}_{r,(Ii)(Jj)}$		$\omega^{RODS}_{k,IJ}$	k	$l = 1$	$s = 1$	$l = 1$
1	17.9	18.2	17.6	16.6	17.6	74.8%	70.8%	79.8%	75.4%	74.8%	1	Vic	Oxford Circus	Cen
2	22.9	23.7	26.1	22.2	20.4	25.2%	29.2%	20.2%	24.6%	25.2%	2	Vic	Green Park	Pic

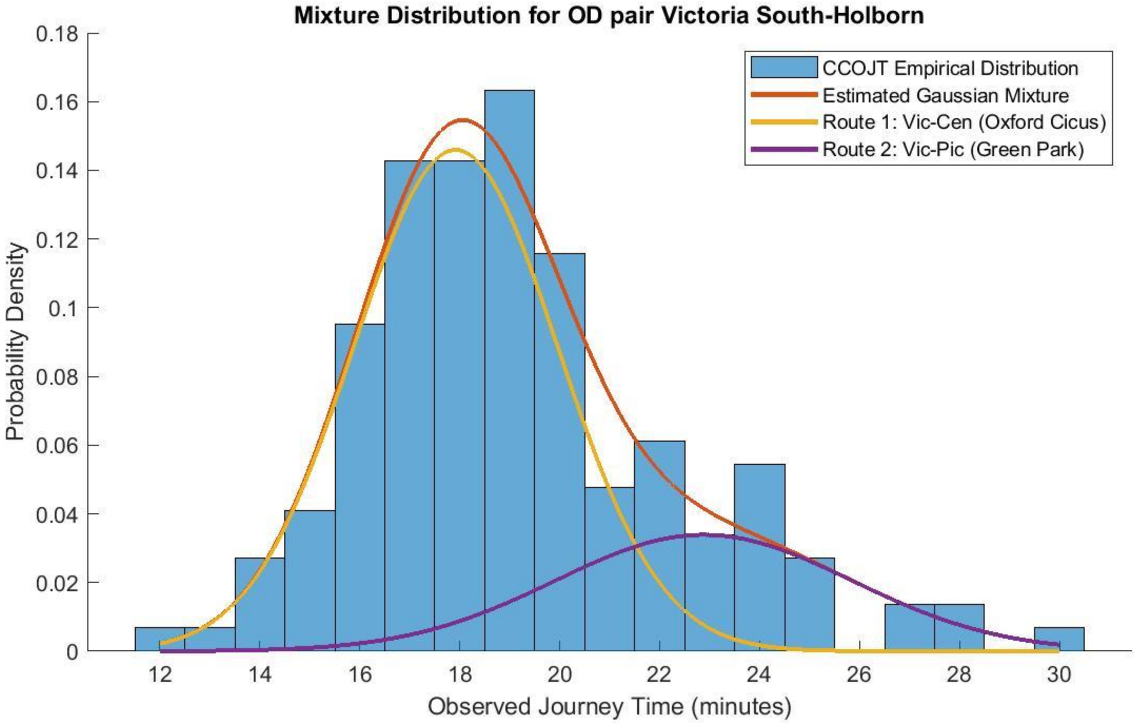


Figure 6-7 Estimated (Gaussian) journey time distribution of the routes for **Victoria South** – **Holborn**, OJTs adjusted to superstation centroid and according to fail-to-board delays

Case 2 Central East – Green Park

Fail-to-board delays at station platforms

Looking at the **Central East** – **Green Park** superstation-to-station OD pair (cf. **Figure 5-15**), it is checked whether capacity problems occur at any stations of the origin superstation (**Mile End**, **Bethnal Green**, **Liverpool Street**, **St Paul's** and **Chancery Lane** stations) as well as at the interchange stations of each reasonable route (**Oxford Circus** and **Holborn** stations). The line capacities for each LU line of the case study OD pair (**Central**, **Victoria**, and **Piccadilly** lines) are calculated from the data described in **Section 6.5.2** (see **Table 6-7**). These capacities were compared to the link flows (cf. **Section 6.5.1**) according to formula (6-1).

Table 6-7 Line capacities for **Case 2,** source:
https://tfl.gov.uk/travel-information/timetables/
https://tfl.gov.uk/corporate/about-tfl/what-we-do/london-underground/rolling-stock

Line	**Station**	**Frequency** f [trains/hour]	**Train Capacity** κ [pax/trains]	**Line Capacity** $\frac{\kappa \cdot f}{4}$ [pax/15 min]
Central	**Liverpool Street**	32	892	7136
Victoria	**Oxford Circus**	30	864	6480
Piccadilly	**Holborn**	20	684	3420

In accordance with Schmöcker et al. (2008), it was understood, that westbound **Central** line is extremely crowded between **Mile End** and **Liverpool Street** stations. This was also confirmed by an initial analysis to run the model with the actual capacity of trains (7136 passengers/15 minutes); and results showed that in many time intervals it occurs that trains already arrive at **Bethnal Green** station full and nobody is able to board. Under these circumstances of extreme crowding, it is expected that passengers choose to board the trains above their nominal capacity, experiencing this way greater discomfort.

To quantify the relationship between the actual and nominal capacity is not a straightforward task due to the lack of the relevant information. Whelan and Crockett (2009) investigated the relationship between the load factor and crowding multiplier in London and South East England. According to their results, when the trains are at their nominal capacity (load factor: 100%), the crowding multiplier for standing passengers is 1.50: while it is 1.67, when the load factor is 120%. These results could be confronted with the increase in the generalised cost of the route due to the delay, which occurs when passengers choose not to board the trains at their nominal capacity. Based on these considerations, it was assumed that in the LU passengers choose the board up to the 10% more than the nominal capacity of the trains, to avoid further delays. Therefore the actual capacity of the **Central** line was considered 7850 passengers/15 minutes.

Figure 6-14 presents the dwell (q_t^{dwell}) and access (q_t^{acc}) link flows in the most congested time interval of the AM peak (08:30-08:45) for the stations of the **Central East** origin superstation. Among these stations, capacity problems occur at **Mile End**, **Bethnal Green** and **Liverpool Street** stations, where the flow ($q_t^{dwell} + q_t^{acc}$) is 1.07, 1.13 and 1.11 times the actual capacity respectively. As for the interchange stations (**Oxford Circus** and **Holborn**), even though they are crowded, the flow ($q_t^{dwell} + q_t^{acc}$) still remains below the line capacity (0.89 and 0.92 times respectively). As the capacity problems occur at the origin stations, it is necessary to do OJT adjustments according to fail-to-board delays (cf. **Section 6.4**).

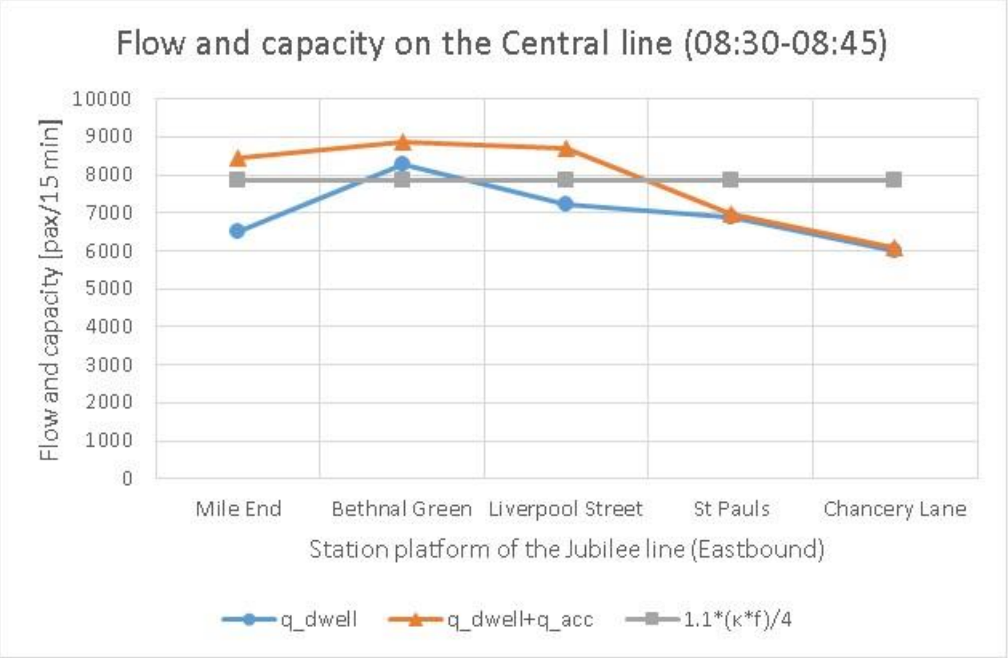


Figure 6-8 Passenger flow and line capacity on the westbound **Central** line, at the stations of the **Central East** origin supertation, peak of peak (08:30-08:45)

Focusing on the stations with capacity problems (i.e. **Central** line westbound platform at **Mile End**, **Bethnal Green** and **Liverpool Street** stations), q_t^{dwell}, q_t^{board} and q_t^{fail} flows are calculated for each time interval t using equations (6-2)-(6-7). Presenting these flows on **Figure 6-9** it was understood that capacity problems occur between 8:15 and 9:15. In this figure the border line between the column of q_t^{board} (grey) and q_t^{fail} (orange) corresponds to the actual line capacity ($1.1 \cdot \frac{(\kappa \cdot f)}{4}$ = 7850 passengers/15 minutes).

From these flows p_t^{fail} and t_t^{fail} were calculated in the congested time intervals (8:15-9:15) according to equations (6-8) and (6-9) and the results are presented in **Table 6-8**. The value of t_t^{fail} is rounded to integer minutes, because they serve for the adjustment of OJTs, which are also given with the same precision. These results were also presented on a histogram to describe the distribution of fail-to-board delays in the AM peak (**Figure 6-10**).

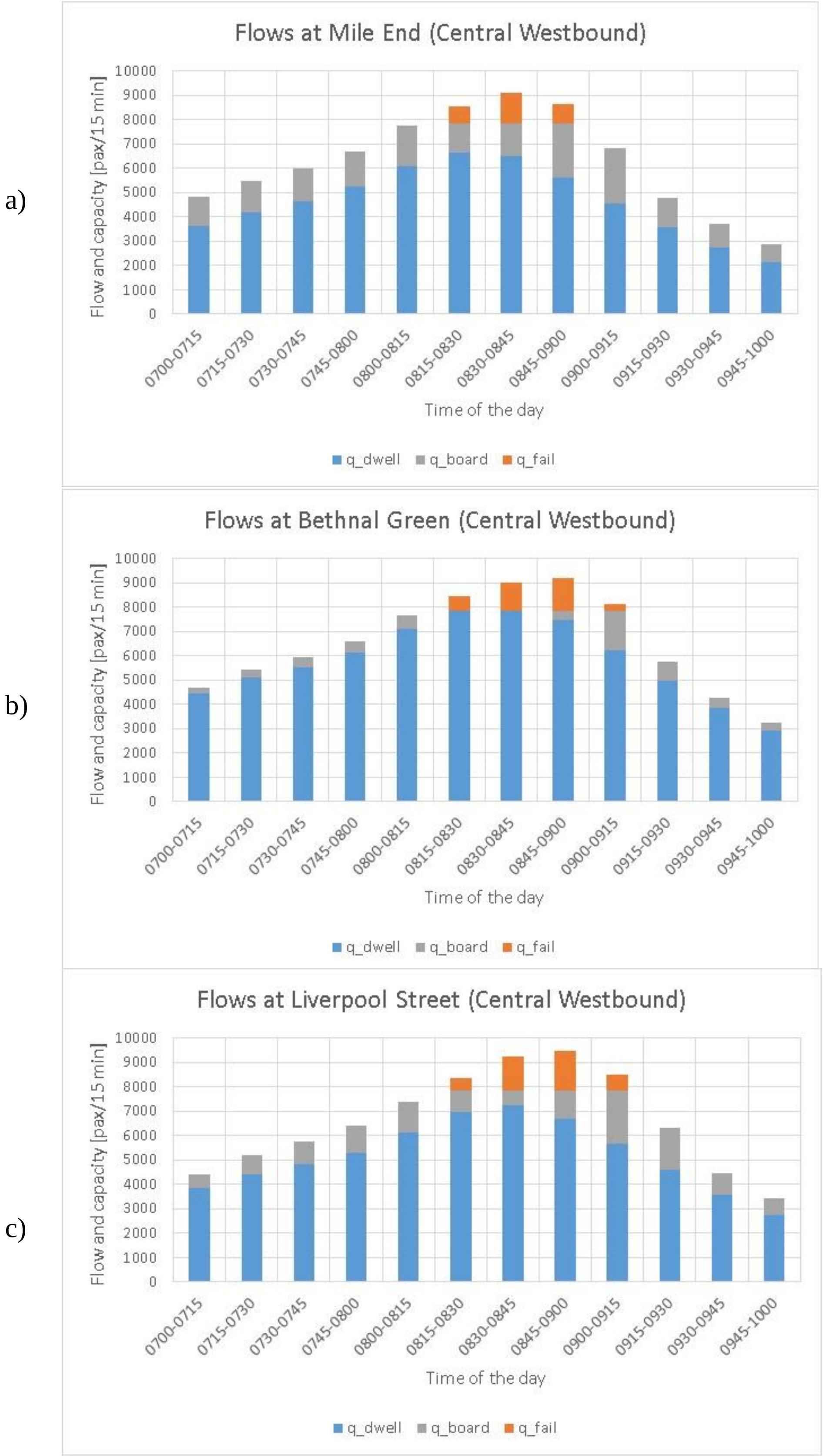


Figure 6-9 Boarding and fail-to-board flows at **Mile End** (a), **Bethnal Green** (b) and **Liverpool Street** (c) stations (**Central** line westbound) in the AM peak (7:00-10:00)

Table 6-8 Average fail to board delays at **Mile End** (a), **Bethnal Green** (b) and **Liverpool Street** (c) stations (**Central** line westbound) in the congested time intervals of the AM peak (8:15-9:15)

a)

Variable	**Value in time interval *t***					
	8:00-8:15	8:15-8:30	8:30-8:45	8:45-9:00	9:00-9:15	9:15-9:30
q_t^{dwell}	6084	6646	6493	5609	4537	3565
q_t^{acc}	1668	1896	1932	1768	1520	1244
q_t^{wait}	1668	1896	2624	3036	2315	1244
q_t^{board}	1668	1204	1357	2241	2315	1244
q_t^{fail}	0	692	1268	795	0	0
p_t^{fail}	0.00	0.37	0.48	0.26	0.00	0.00
t_t^{fail}	0	5	7	4	0	0

b)

Variable	**Value in time interval *t***					
	8:00-8:15	8:15-8:30	8:30-8:45	8:45-9:00	9:00-9:15	9:15-9:30
q_t^{dwell}	7111	7850	7850	7494	6241	4969
q_t^{acc}	569	602	569	535	530	502
q_t^{wait}	569	602	1171	1706	1880	774
q_t^{board}	569	0	0	356	1609	774
q_t^{fail}	0	602	1171	1350	272	0
p_t^{fail}	0.00	1.00	1.00	0.79	0.14	0.00
t_t^{fail}	0	15	15	12	2	0

c)

Variable	**Value in time interval *t***					
	8:00-8:15	8:15-8:30	8:30-8:45	8:45-9:00	9:00-9:15	9:15-9:30
q_t^{dwell}	6135	6982	7230	6682	5672	4597
q_t^{acc}	1245	1404	1477	1397	1226	1045
q_t^{wait}	1245	1404	2013	2791	2849	1717
q_t^{board}	1245	868	620	1168	2178	1717
q_t^{fail}	0	536	1394	1623	672	0
p_t^{fail}	0.00	0.38	0.69	0.58	0.24	0.00
t_t^{fail}	0	6	10	9	4	0

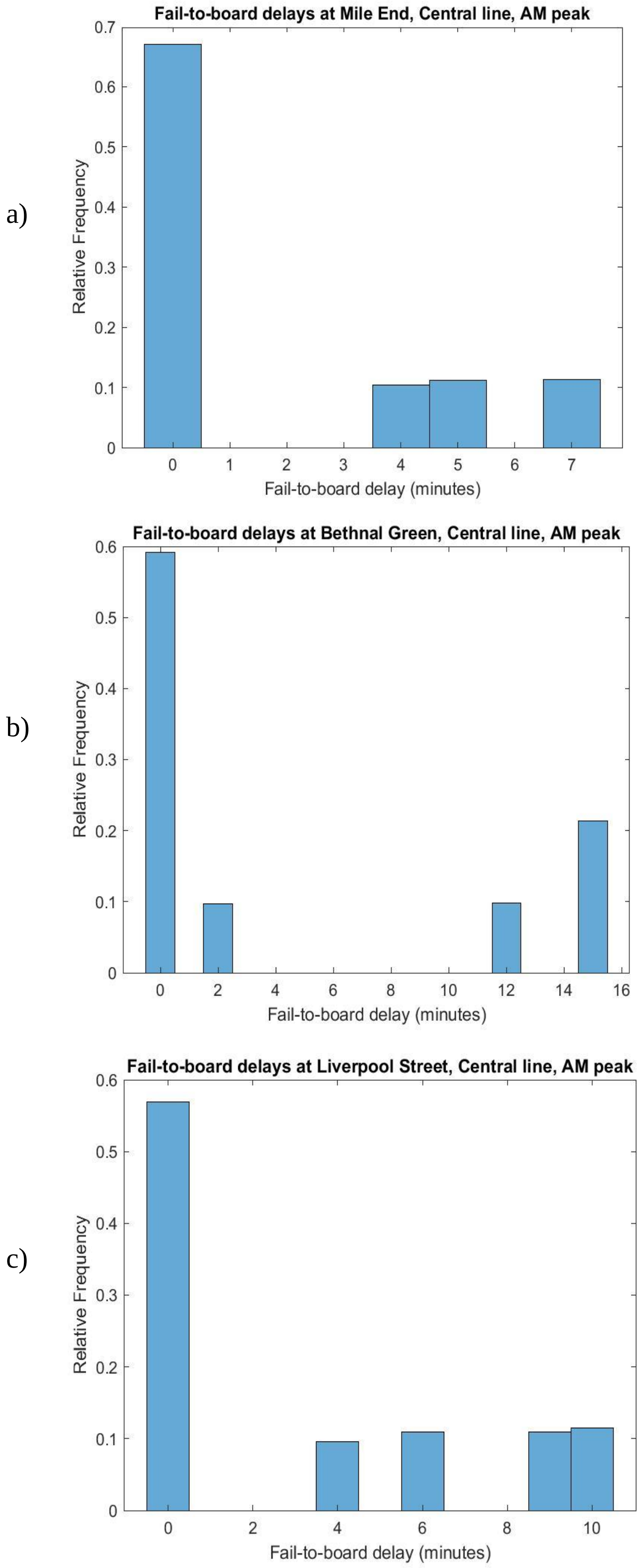


Figure 6-10 Distribution of fail-to-board delays at **Mile End** (a), **Bethnal Green** (b) and **Liverpool Street** (c) stations (**Central** line westbound) in the AM peak (7:00-10:00)

Adjustment of OJTs according to fail-to-board delays

Following the methodology in **Section 6.4**, based on the fail-to-board delay distribution (**Figure 6-10**), the $\varphi^{\tau}_{(Ii)(Ic)}$ indices (equation (6-13)) and hence the $OJT^{\tau}_{(Ii)(Jj)}$ subsets as well as their adjustment, $OJT^{\tau,fail}_{(Ii)(Jj)}$ (equation (6-14)) were calculated for each outcome of fail-to-board delay value τ for the stations where capacity problems occur (i.e. **Mile End**, **Bethnal Green** and **Liverpool Street** stations, westbound **Central** line see **Table 6-9** and **Figure 6-11**).

Table 6-9 Subsets of the OJT dataset according to fail-to-board delays at **Mile End** (a), **Bethnal Green** (b) and **Liverpool Street** (c) stations (**Central** line westbound) in the AM peak (7:00-10:00)

a)

τ	$Pr(\delta^{fail} \leq \tau)$	$\varphi^{\tau}_{(Ii)(Ic)}$	$OJT^{\tau,fail}_{(Ii)(Jj)}$	
			min	max
0	0.67	5	23	28
4	0.77	6	30	30
5	0.89	7	34	34
7	1.00	8	39	39

b)

τ	$Pr(\delta^{fail} \leq \tau)$	$\varphi^{\tau}_{(Ii)(Ic)}$	$OJT^{\tau,fail}_{(Ii)(Jj)}$	
			min	max
0	0.59	5	20	24
2	0.69	6	29	29
12	0.79	6	29	29
15	1.00	8	30	43

c)

τ	$Pr(\delta^{fail} \leq \tau)$	$\varphi^{\tau}_{(Ii)(Ic)}$	$OJT^{\tau,fail}_{(Ii)(Jj)}$	
			min	max
0	0.57	17	16	21
4	0.67	20	22	22
6	0.78	23	23	23
9	0.88	27	24	25
10	1.00	30	27	36

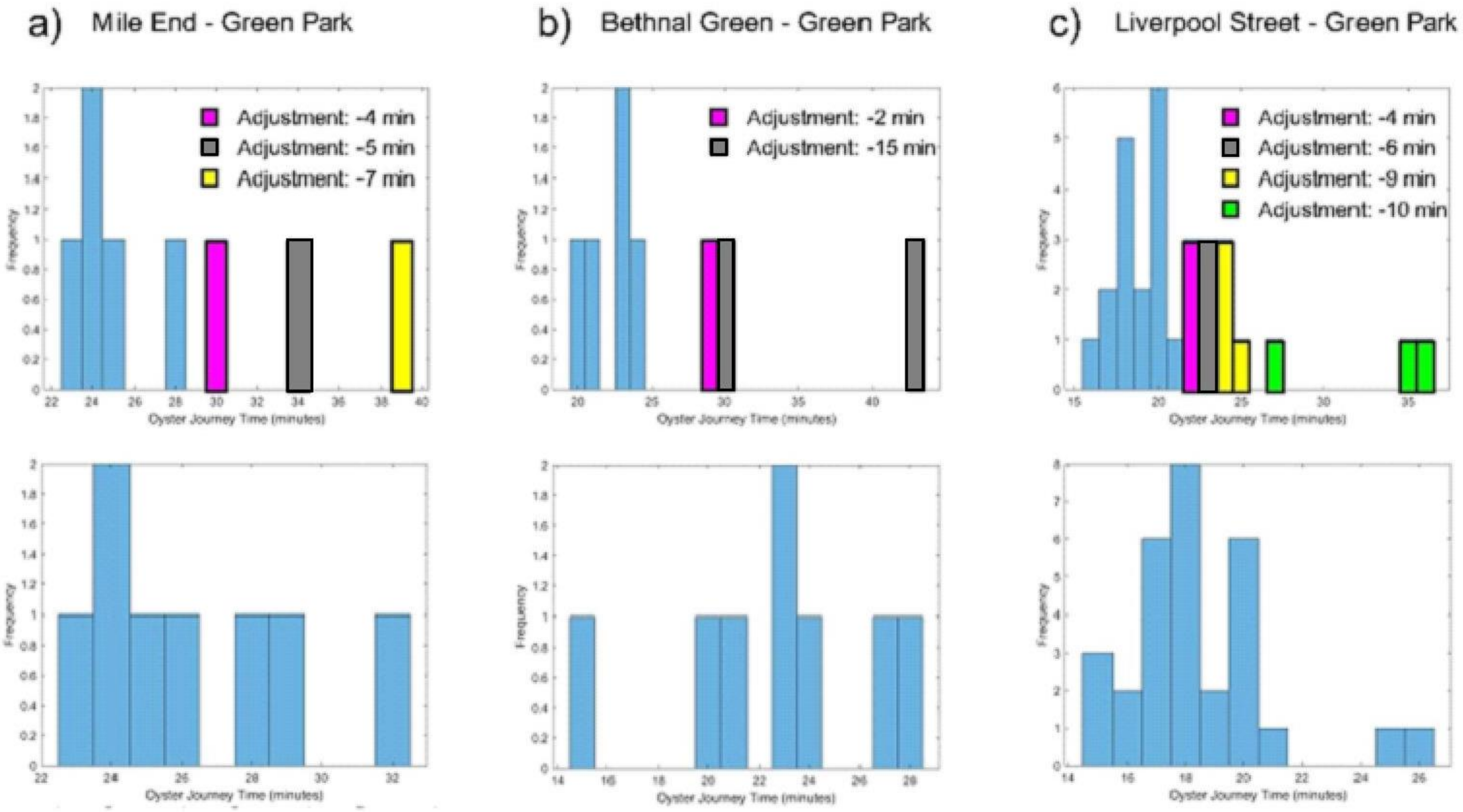
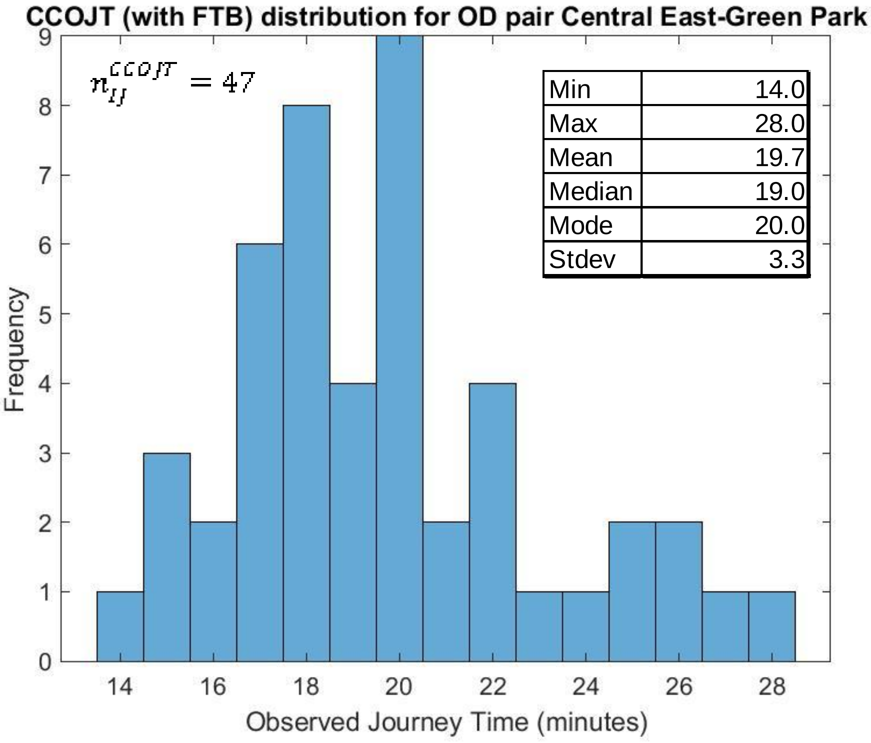


Figure 6-11 Adjustment of OJTs according to fail-to board delays at **Mile End** (a), **Bethnal Green** (b) and **Liverpool Street** (c) stations (**Central** line westbound) in the AM peak (7:00-10:00); above: Original OJTs from Oyster data and proposed adjustments, below: Adjusted OJTs

Each subset of $OJT_{(Ii)(Jj)}^{\tau,fail}$ were aggregated to obtain the adjusted $OJT_{(Ii)(Jj)}^{fail}$ dataset for station-to-station OD pairs (equation (6-15)). Following this, these OJTs were further adjusted to superstation centroid (i.e. **Liverpool Street** station) and then aggregated spatially as described in **Section 5.4**. This way the adjusted $CCOJT_{IJ}^{fail}$ was obtained for the **Central East** – **Green Park** superstation-to-station OD pair (see **Figure 6-12**).



Min	14.0
Max	28.0
Mean	19.7
Median	19.0
Mode	20.0
Stdev	3.3

Figure 6-12 Distribution of Centroid-to-Centroid adjusted OJTs considering fail-to-board delays for **Central East**– **Green Park**

Evaluation of the OJT adjustment according to fail-to-board delays

The finite mixture model presented in **Chapter 3** was applied on CCOJT dataset adjusted according to fail-to-board delays. A more detailed description of the settings and of the results are presented in **Appendix G**. Based on that, the chosen settings for the finite mixture model are:

- Seed = 1
- Tolerance threshold = 1e-07

The results with these settings are presented in **Table 6-10**. Following this, the results of the finite mixture model were matched with the actual LU routes (cf. **Section 3.4.1**). **Table 6-11** compares the mixture results for the **Central East** – **Green Park** superstation-to-station OD pair with the two types of OJT adjustments:

- Only to superstation centroids (**Chapter 5**)
- To superstation centroid and according to fail-to-board delays (**Chapter 6**)

These are further compared with the earlier results presented in **Table 5-14**. **Figure 6-13** presents the probability density functions of the mixture distribution fit on the CCOJT dataset adjusted according to fail-to-board delays as well as of the mixture components matched with the actual LU routes.

Table 6-10 Finite mixture model results; with Seed = 1, Tolerance threshold = 1e-07 for **Central East**– **Green Park**

OJTs adjusted to superstation centroid and according to fail-to-board delays

Label	Mixture model		
r []	$\mu_{r,IJ}^{MIX}$ [min]	$\sigma_{r,IJ}^{MIX}$ [min]	$\omega_{r,IJ}^{MIX}$ [%]
1	18.7	2.3	86.2%
2	25.9	1.3	13.8%

Based on these results, the following was observed: The finite mixture model applied on the CCOJTs of the **Central East** – **Green Park** superstation-to-station OD pair adjusted also according to fail-to-board delays gave closer results to the journey time of the actual LU route for the mean of component 2 ($\mu_{2,IJ}^{MIX}$); however this journey time value was quite low (18.7 minutes). Regarding the component proportion, it was understood that the results of the finite mixture model were closer to the RODS results of the actual LU routes when the OJTs were adjusted only to superstation centroid, but not according to fail-to-board delays (**Chapter 5**).

Furthermore, it was understood that when the OJTs were adjusted only to superstation centroid (**Chapter 5**), the K-means clustering algorithm gave the same initial values for all seeds, and there was a only a slight difference between the results of the finite mixture model for different tolerance thresholds (cf. **Figure F-4** and **Figure F-5**). However, with the OJTs adjusted also according to fail-to-board delays (**Chapter 6**), the finite mixture model converged to different roots for different seeds, when the tolerance threshold was set 1e-04 or greater; and a greater jump was observable between the tolerance threshold of 1e-03 and 1e-04 for seed 1 and between 1e-04 and 1e-05 for seed 2 (**Figure G-4** and **Figure G-5**).

Overall, from **Case 2** it was understood that when the OJTs were adjusted only to superstation centroid, but not according to fail-to-board delays (**Chapter 5**); could give more reliable estimates, both in terms of closeness of results to the actual LU routes and in terms of the convergence of the finite mixture model. One possible reason why the adjustment according to fail-to-board delays (**Chapter 6**) could not improve the model estimates is that the assumption of the 10% additional capacity with respect to the nominal capacity of trains was still underestimating the actual willingness of passengers to board overcrowded trains to avoid fail-to-board delays. Supposing a higher additional capacity could have improved the model estimates.

Another possible reason for not obtaining closer results with the adjustment according to fail-to-board delays is, that the sample size for Case 2 was very small (47 CCOJT records), which was still insufficient to represent well the actual journey time distribution of passengers.

Table 6-11 Matching mixture model results with the actual London Underground routes for Central East– Green Park

Blue: Mixture results, adjustment: superstation centroid and fail-to-board delays **Purple**: Mixture results, , adjustment: superstation centroid only, **Red**: Mixture results, station OD pairs, **Green**: actual LU routes

Mixture Label	Journey Time				Route Choice (%)				Route Label	Route Matched		
	Mixture			Timetable	Mixture			RODS		Line 1	Interchange 1	Line 2
	Proposed				Proposed							
	FTB	SS	Station		FTB	SS	Station					
r	$\mu^{MIX}_{r,IJ}$		$\mu^{MIX}_{r,(Ii)(Jj)}$	$t^{SJT}_{k,IJ}$	$\omega^{MIX}_{r,IJ}$		$\omega^{MIX}_{r,(Ii)(Jj)}$	$\omega^{RODS}_{k,IJ}$	k	$l = 1$	$s = 1$	$l = 1$
1	18.7	20.6	20.6	22.3	86.2%	80.3%	93.3%	81.2%	1	Cen	Oxford Circus	Vic
2	25.9	29.5	35.5	25.5	13.8%	19.7%	6.7%	18.8%	2	Cen	Holborn	Pic

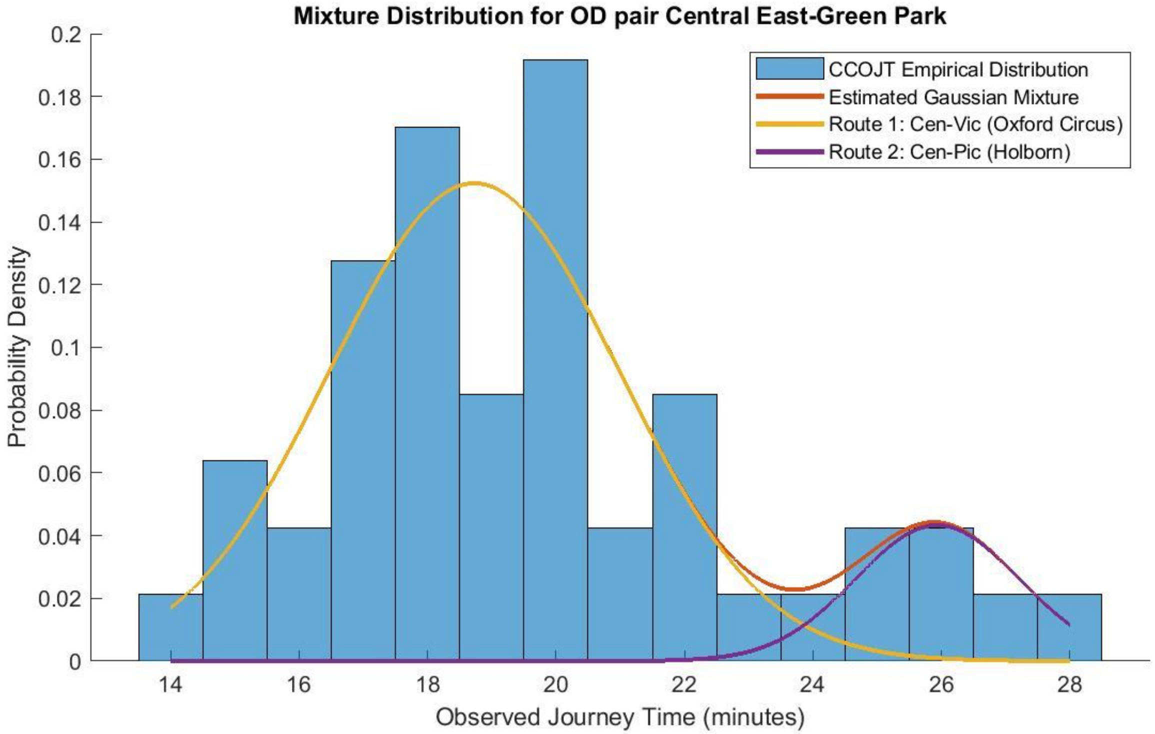


Figure 6-13 Estimated (Gaussian) journey time distribution of the routes for **Central East**– **Green Park**, OJTs adjusted to superstation centroid and according to fail-to-board delays

Case 3 Jubilee West – Jubilee Central

Fail-to-board delays at station platforms

Looking at the Jubilee West – Jubilee Central superstation-to-superstation OD pair (cf. **Figure 5-22**) it is checked whether capacity problems occur at any stations of the origin superstation (**Stanmore**, **Canons Park**, **Queensbury** and **Kingsbury** stations) as well as at the interchange stations of each reasonable route **Wembley Park** and **Finchley Road** stations). The line capacities for each LU line of the case study OD pair (**Jubilee** and **Metropolitan** lines) are calculated from the data described in **Section 6.5.2** (see **Table 6-12**) and compared to the link flows (cf. **Section 6.5.1**) according to formula (6-1).

Table 6-12 Line capacities for **Case 3** source:
https://tfl.gov.uk/travel-information/timetables/
https://tfl.gov.uk/corporate/about-tfl/what-we-do/london-underground/rolling-stock

Line	**Station**	**Frequency** f [trains/hour]	**Train capacity** κ [pax/trains]	**Line Capacity** $\frac{\kappa \cdot f}{4}$ [pax/15 min]
Jub	**Stanmore**	20	817	4085
Met	**Wembley Park**	12	1044	3132
Jub	**Finchley Road**	30	817	6128

Figure 6-14 presents the dwell (q_t^{dwell}) and access (q_t^{acc}) link flows in the most congested time interval of the AM peak (08:00-08:15) for the stations of the Jubilee West origin superstation. It can be understood that no capacity problems occur for any of these stations as they are near the line terminus. Among them that maximum passenger flow is at **Kingsbury** station, which is far below the line capacity ($q_t^{dwell} + q_t^{acc} = 0.34 \cdot \frac{\kappa \cdot f}{4}$).

Among the interchange stations of the reasonable routes, the capacity problems occur only at **Wembley Park** station for the **Metropolitan** line[26] ($q_t^{dwell} + q_t^{acc} = 1.08 \cdot \frac{\kappa \cdot f}{4}$ between 08:15 and 08:30). At **Finchley Road** station, the flows for the **Jubilee** are high, but still under the line capacity ($q_t^{dwell} + q_t^{acc} = 0.75 \cdot \frac{\kappa \cdot f}{4}$, between 08:30 and 08:45). As the capacity problems occur at the interchange station, the OJT adjustment applied for the previous cases is not relevant here, as it is not explicitly known, whether an OJT record belongs to the congested route or not. To address this issue, the methodology of **Chapter 7** is discussed first, and then the corresponding case study is presented in **Section 7.7**. Here, the case study is described until finding the fail-to-board delays at **Wembley Park** station.

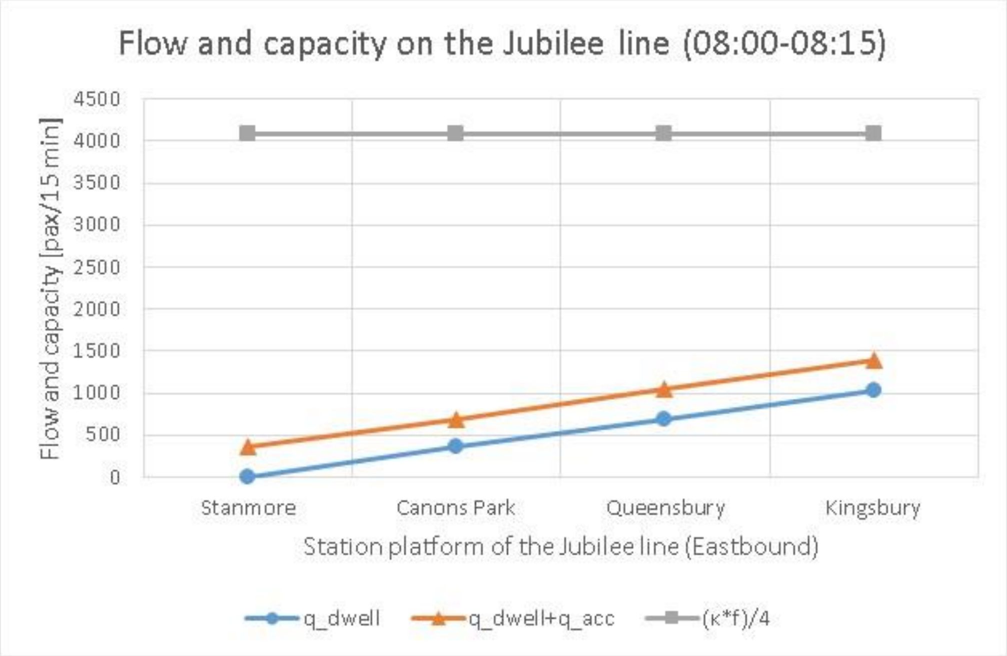


Figure 6-14 Passenger flow and line capacity on the eastbound **Jubilee** line, at the stations of the **Jubilee West** origin supertation, peak of peak (08:00-08:15)

At **Wembley Park** station, (**Metropolitan** line eastbound), q_t^{dwell}, q_t^{board} and q_t^{fail} flows are calculated for each time interval t using equations (6-2)-(6-7). Presenting these

[26] Some of the **Metropolitan** services stop at **Wembley Park** station, some of them pass through without stopping. From RODS data (cf. **Section 6.5.1**), only the "line load" before **Wembley Park** station can be known, but it is not distinguished on what type of **Metropolitan** service it is. In this thesis, it was assumed, that 2/3 of the total "line load" is on services that stop at **Wembley Park**. http://content.tfl.gov.uk/amersham-guide-dec18.pdf

flows on **Figure 6-15**, it was understood that capacity problems occur only between 8:00 and 8:45. There the border line between the column of q_t^{board} (grey) and q_t^{fail} (orange) corresponds to the line capacity ($\kappa \cdot f$ = 3132 passengers/15 minutes).

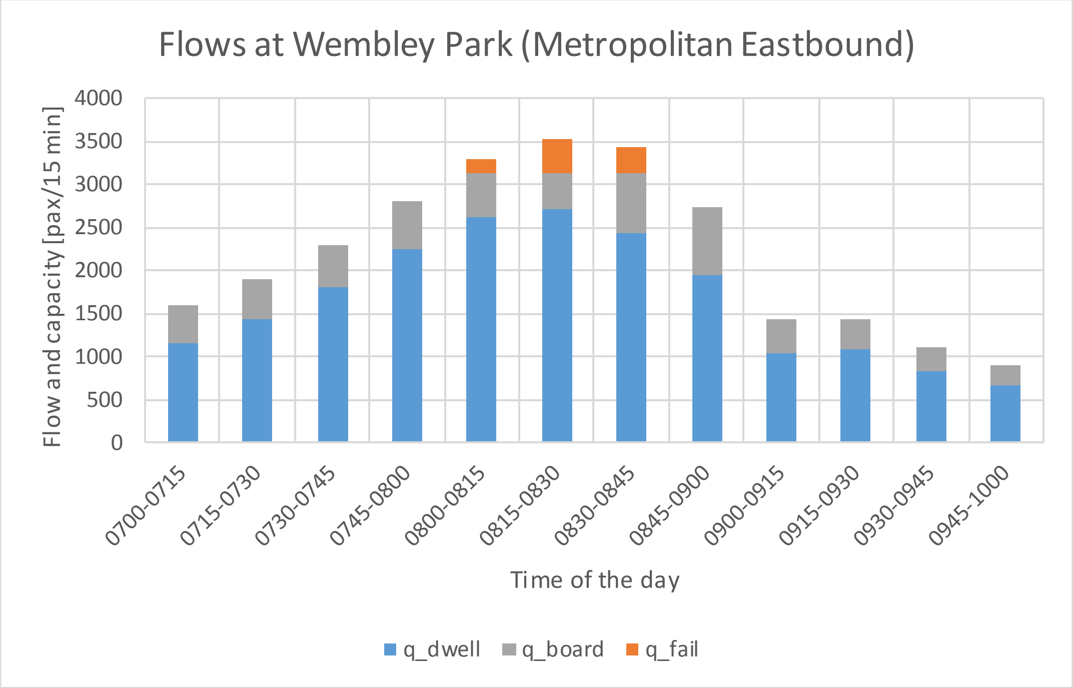


Figure 6-15 Boarding and fail-to-board flows at **Wembley Park** station (**Metropolitan** line eastbound) in the AM peak (7:00-10:00)

Table 6-13 Average fail to board delays at **Wembley Park** station (**Metropolitan** line eastbound) in the congested time intervals of the AM peak (8:00-8:45)

Variable	**Value in time interval *t***				
	8:00-8:15	8:15-8:30	8:30-8:45	8:45-9:00	8:00-8:15
q_t^{dwell}	2245	2631	2705	2426	1941
q_t^{acc}	573	655	675	605	491
q_t^{wait}	573	655	829	1007	792
q_t^{board}	573	501	427	706	792
q_t^{fail}	0	154	402	301	0
p_t^{fail}	0.00	0.24	0.48	0.30	0.00
t_t^{fail}	0	4	7	4	0

From these flows p_t^{fail} and t_t^{fail} were calculated in the congested time intervals (8:00-8:45) according to equations (6-8) and (6-9) and presented the results in **Table 6-13**. The value of t_t^{fail} is rounded to integer minutes. Looking at the time intervals with capacity problems (8:00-8:45), there is an average of 4 minutes delay at the beginning (8:00-8:15) and at the end (8:30-8:45) of the period of congestion; and an even higher average delay of 7 minutes in between (8:15-8:30) (highlighted with yellow).

6.7 Discussions

6.7.1 Applicability of the quasi-dynamic approach

In this chapter the quasi-dynamic approach was applied to infer boarding and fail-to-board flows as well as fail-to-board delays. The duration of the time intervals was set to be 15 minutes. This has been proved to be a good tool to estimate fail-to-board delays without the need of going schedule-based (calculating for individual passengers and trains). One drawback is that with this method only the average values of fail-to-board delays could be obtained for each 15 minute time intervals. To obtain a more detailed picture, whilst remaining in the quasi dynamic context, one could assume a probabilistic distribution for passenger arrivals and train headways and estimate the distribution of fail-to-board delays accordingly.

For the capacity of trains (κ) the rolling stock information from the TfL website was used, which would correspond to the nominal capacity of trains. For **Case 1** it was assumed that on the **Victoria** line at **Victoria** station trains can carry up to their nominal capacity, and passengers above that fail to board. On the other hand, for **Case 2** it was assumed that on the **Central** line between **Mile End** and **Liverpool Street** trains can carry an additional 10% of passengers above their nominal capacity as passengers are more likely to travel under greater discomfort in order to avoid further delays. The critical issue at this point is that the assumption on the relationship between the actual and nominal capacity of trains affects the results for the adjusted OJTs.

Here, results showed that for both cases the OJTs were over-adjusted as a lower additional capacity was assumed. For the **Victoria-Holborn** station-to-station OD pair (**Case 1**) it can be observed that the OJT distribution adjusted according to fail-to-board delays (cf. **Figure 6-5** b) is skewed left. In reality for the OJTs distribution of an OD pair one would expect that it is skewed right. This can be seen clearer looking at the CCOJT distribution of the **Central East** – **Green Park** superstation-to-station OD pair (**Case 2**), where the CCOJT values of 14 and 15 minutes are unrealistically small.

A possible refinement of this model could be to analyse more in details passengers' perception to discomfort (Whelan and Crockett, 2009; Li and Hensher, 2011; Hörcher et al., 2017; Tirachini et al., 2017) depending on the case study OD pair and calculate with the actual capacity of trains accordingly.

6.7.2 Benefits of adjustments according to fail-to-board delays

The crucial point in the methodology applied in this chapter is that adjusting the observed $OJT_{(Ii)(Jj)}$ dataset with the modelled τ fail-to-board delays and hence applying the finite mixture model on the adjusted and aggregated $CCOJT_{IJ}^{fail}$ distribution would be an optimistic view about the likelihood of passengers' boarding, as it would assume that fail-to-board delays had not occurred. In reality, fail-to-board delays do occur, and it is expected that they influence route choice.

For this reason, the benefits of the adjustment according to fail-to-board delays is not as evident as it was when the OJTs were adjusted only to superstation centroid (**Chapter 5**). Although the for **Case 1**, it could give closer results for the mean and for the component proportion to the actual LU routes; for **Case 2**, the estimates for the mean resulted lower and for the component proportion they were further from the RODS results of the actual routes.

A possible approach to take into consideration the different route choice behaviour in different time periods would be to apply the finite mixture model on the dataset of each time period. For example one time period could be the peak of peak (i.e. 8:15-9:15) and the other would be the normal AM peak (i.e. 7:00-8:15 and 9:15-10:00). Although this approach may give better estimates, it would conflict with the general aim of this thesis, which is to aggregate the OJT records to have more reliable estimates.

Thinking one step ahead, a different approach would be to apply the finite mixture model on the CCOJT of the whole AM peak, and update the estimated route choice probabilities with the additional information on the time interval, when the passenger accessed the crowded platform. The principles of this approach is discussed in **Chapter 7**, however there it is presented in another context.

6.7.3 Data sources on passenger flows

In this thesis RODS data reconciled to passenger counts was used to understand passenger flows on the LU lines and within the LU stations. The deficiency with this approach could be explained in the following: Relying still on manual surveys would not comply with the overall objective of this thesis to go towards automatically collected data systems for route choice estimation (**Section 2.1**). More specifically – as presented in **Section 6.5.1** – RODS data was collected over several years, different years at different stations, therefore the route choice and hence the passenger flow results do not reflect the time period of the Oyster data collection.

In theory, passenger flows could be understood from the route choice estimates for all OD pairs of the LU network and the problem could be resolved as a transit assignment model (cf. **Figure 2-1**). However, this would require to build the model for the entire LU and rail network of Greater London (cf. **Section 4.3**), which is beyond the scope of this thesis.

Chapter 7
Consideration of fail-to-board delays at the interchange station

7.1 Introduction

In **Chapter 6** the question of fail-to-board delays at the origin station was discussed. In line with the **Definition** of superstations, the case studies in this thesis focus on origin destination (OD) pairs with the property that for all reasonable routes the first journey leg is on the same line (cf. **Figure 5-10**, **Figure 5-15** and **Figure 5-22**). Therefore, in case the fail-to-board delay occurred at any station of the origin superstation (**Case 1** and **Case 2**), their Observed Journey Time (OJT) could be simply adjusted to its uncongested equivalent and hence the OJTs of different origin stations could be aggregated.

However, in case the fail-to-board delay occurs at the interchange station (**Case 3**), the question is more challenging, because it is not possible to know deterministically, whether an OJT record corresponds to the congested route or not; it can be estimated only in a probability space (cf. **Section 3.2**). Therefore, the OJT adjustment proposed in **Section 6.4** is not applicable in this case, but a different approach is required.

In this section notation is used as follows.

Variables

$t_{k,Ii}^{entry-ic}$ Journey time between the entry ticket gate of origin station Ii and the departure platform of the subsequent journey leg at the congested interchange station, on route k of OD pair IJ

$l\#,k,IJ$ Index of the journey leg on which passengers experience fail-to-board delays on route k of OD pair IJ

$s\#,k,IJ$ Index of the interchange station at which passengers experience fail-to-board delays on route k of OD pair IJ

$T_{q,k,IJ}^{plat-ic}$ Arrival time stamp of passenger q at the interchange station (departure platform of the subsequent journey leg), given that he/she chooses route k of OD pair IJ

$t_{q,k,IJ}^{\mathrm{EXP}}$ Expected Journey Time of passenger q using route k between centroids of superstations I and J, given $T_{q,Ii}^{entry}$, timetables and station layouts

$\delta_{q,r,IJ}^{\mathrm{EXP}}$ Random variable of $t_{q,k,IJ}^{\mathrm{EXP}}$ in the Bayesian framework

$\delta_{q,IJ}^{\mathrm{CCOJT}}$ Journey time observation of passenger q adjusted to the centroids of superstations I and J (CCOJT) (minutes)

$\Delta_{q,IJ}$ Elementary event that the CCOJT of passenger q is $\delta_{q,IJ}^{\mathrm{CCOJT}}$

$\Delta_{qr,IJ}$ Elementary event that the Expected Journey Time of passenger q is $\delta_{q,r,IJ}^{\mathrm{EXP}}$, given that he/she chooses route r and his/her entry time is $T_{q,Ii}^{entry}$

$\delta_{qr,IJ}$ The journey time distribution of passenger q on route r

$CCOJT_{r,IJ}^{naive}$ Sub-dataset of $CCOJT_{IJ}$ based on **naïve** inference $\zeta(q)$

$n_{r,IJ}^{naive}$ Total number of passengers in the $CCOJT_{IJ}$ dataset who were assigned to route r based **naïve** inference $\zeta(q)$

$\mu_{r,IJ}^{UPD}$ Mean journey time for route r of superstation-to-superstation OD pair IJ, with the **upd**ate according to the additional information on fail-to-board delays (minutes)

$\sigma_{r,IJ}^{UPD}$ Standard deviation of journey time for route r of superstation-to-superstation OD pair IJ, with the **upd**ate according to the additional information on fail-to-board delays (minutes)

$\omega_{r,IJ}^{UPD}$ Aggregate choice probabilities for route r of superstation-to-superstation OD pair IJ, with the **upd**ate according to the additional information on fail-to-board delays

$t_{r,IJ}^{REF}$ Reference time of route r

ψ_r^{naive} Dummy variable, $\psi_r^{naive} = 1$ if, $\zeta(q) = r$, otherwise $\psi_r^{naive} = 0$

Functions

$f\ (\delta_{qr})$ Probability density function of δ_{qr}

$\zeta(q)$ Assignment function for the naïve inference of the route choice of each passenger q, based on their posterior probabilities (CCOJT and additional information on fail-to-board delays)

The main objective in this chapter is to update the prior knowledge on route choice (i.e. results of **Chapter 5**) with the additional information on fail-to-board delays in a Bayesian framework following the concept in Fu (2014). **Figure 7-1** illustrates a Bayesian network structure how the information on fail-to-board delay (δ^{fail}) can be associated with the smart card observations and how it influences the journey time and route choice of passengers.

In this framework the observations (marked with **orange** nodes) are the entry time (T^{entry}) and the exit time (T^{exit}) of passengers[27]. From these input their OJT (δ^{OJT}) can be directly obtained (cf. **Section 3.6**), and hence their Centroid-to-Centroid adjusted OJT (CCOJT, δ^{CCOJT}) can be calculated (cf. **Section 5.4**). This obvious dependency is marked with **orange** solid arc and the inferred variables are marked as **blue** nodes.

It was illustrated through the case studies in **Section 6.6** that the fail-to-board delay (δ^{fail}, cf. **Table 6-3** and **Table 6-8**) varies within the AM peak. Therefore, it is dependent on the entry time of the passenger (T^{entry}). Furthermore, it can be easily understood, that a passenger experiences fail-to-board delay only, if he/she chooses the congested route. Therefore fail-to-board delay is also dependent on the chosen route (r)[28]. Additionally, the expected journey time of a passenger (δ^{EXP}) also depends on the other time components (i.e. on-board, wait, access egress interchange) he/she experienced along

[27] In **Section 7.1**, the purpose is to give a simplified representation of the Bayesian network structure by elucidating the dependencies among the variables. Therefore the variable identifiers are not used in this section. In **Section 7.2** this Bayesian network structure is further expanded, therefore all variables presented with the appropriate identifiers.

[28] In **Section 7.1** – as it focuses on the dependencies of the variables - routes are denoted with r, regardless whether they refer to mixture component or actual LU route. In **Section 7.3** the concept for matching these two labels is further explained.

his/her chosen route. Therefore, the expected journey time is further dependent on the chosen route (r).

The problem here is that the chosen route (r) is unobservable, it can be only learnt in a probability space (this type of dependency is marked with a **blue** dashed arc) from the journey time of passengers. In this chapter, route choice is inferred not only based on the CCOJT (δ^{CCOJT}) of passengers (cf. **Chapter 5**), but also based on their expected journey time (δ^{EXP}).

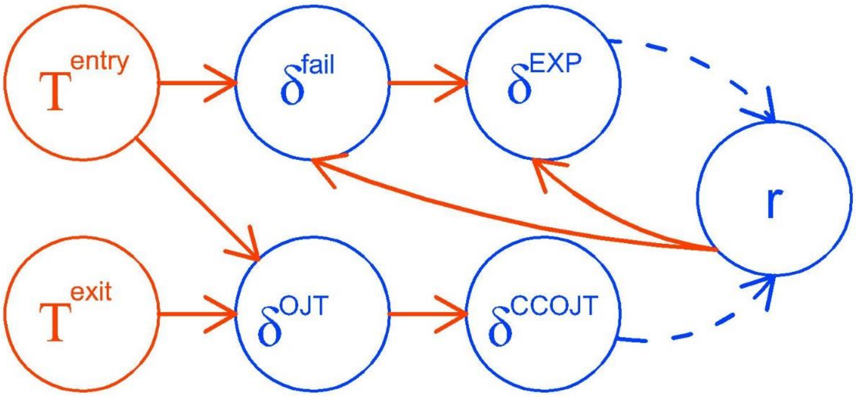


Figure 7-1 Bayesian network structure to illustrate passengers probabilistic route choices

The rest of this chapter is structured as follows. In order to have the correct input for the Bayesian updating process, in **Section 7.2** the calculation of the Expected Journey Times are presented; and in **Section 7.3** the principles for matching the mixture components with the actual routes are clarified. **Section 7.4** presents the Bayesian framework itself to obtain the updated individual route choice probabilities; and **Section 7.5** describes the methodology to infer the aggregate values corresponding to each route. In **Section 7.6** the proposed approach is compared to Fu (2014). **Section 7.7** presents the case study on the London Underground (LU). Finally, **Section 7.8** concludes the chapter with the evaluation of the obtained results and with the lessons learnt from that.

7.2 Connection between passengers' entry time and Expected Journey Time

In this section the dependencies between the entry time of a passenger (T^{entry}) and his/her expected journey time (δ^{EXP}) are further explained and presented on an expanded Bayesian network structure (**Figure 7-2**).

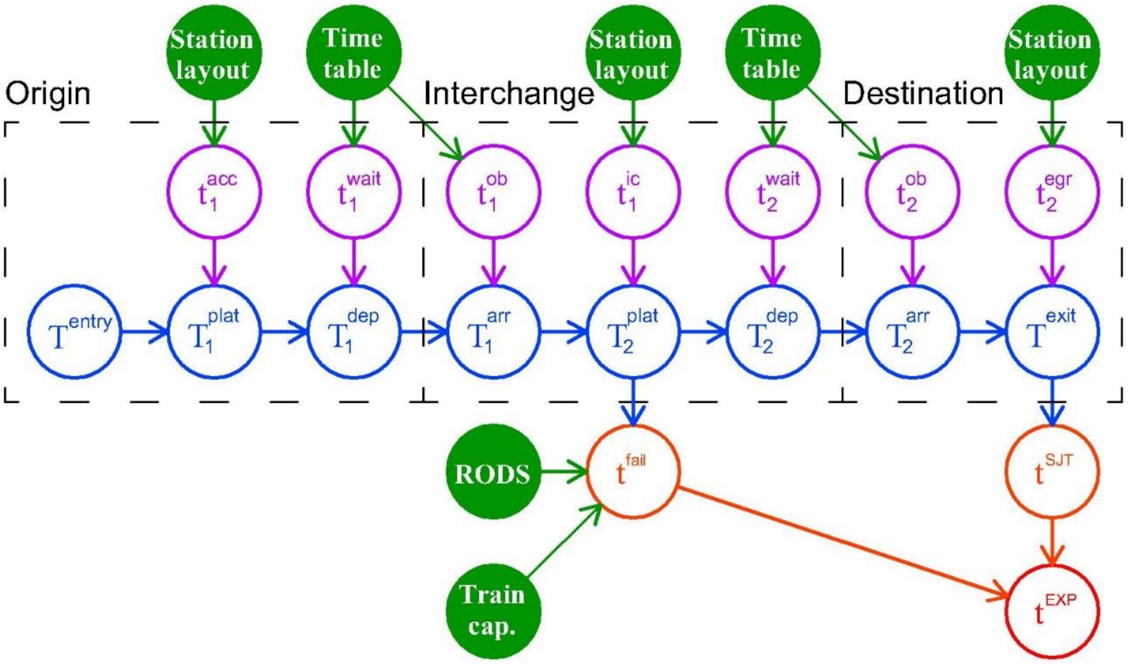


Figure 7-2 Bayesian network structure to illustrate the connection between passengers entry time and Expected Journey Time

In this Bayesian network structure the data sources, from which the journey time components can be understood (i.e. **timetable**, **station layout**, Rolling Origin Destination Survey (**RODS**) and **train capacities**, cf. **Section 3.6** and **Section 6.5**) are marked as **green** filled nodes. The journey time components themselves (i.e. **acc**ess, **wait**, **o**n-**b**oard, **i**nter**c**hange and **egr**ess) are marked as **purple** nodes, and the resulting time stamps (**entry**, **plat**form arrival, train **dep**arture, train **arr**ival and **exit**) as **blue** nodes. In this Bayesian network structure, the subtotals of journey times are the Scheduled Journey Time (**SJT**) and the **fail**-to-board delay (marked as **orange** nodes); and the final result is the **Exp**ected Journey Time (marked as a **red** node).

The entry time of passenger q at station Ii of the origin superstation ($T_{q,Ii}^{entry}$) can be directly understood from smart card data (cf. **Section 3.6.1**). In the previous chapters (**Chapter 3**, **Chapter 5** and **Chapter 6**), smart card records were used only with their OJTs ($\delta_{q,(Ii)(Jj)}^{\mathrm{OJT}}$, cf. equation (3-14)). In this chapter, in order to gain a better understanding on the associated fail-to-board delays, also their entry time is required as an input.

Having the adequate information on train timetables and station layouts (cf. **Section 3.6.2**) the journey time between station Ii (entry ticket gate) and the congested interchange station (departure platform of subsequent journey leg) on route k can be obtained as the sum of the time components (cf. **Figure 7-2**):

$$t_{k,Ii}^{entry-ic} = t_{1,k,Ii}^{acc} + \sum_{l=1}^{l\#,k,IJ} (t_{l,k,IJ}^{wait} + t_{l,k,IJ}^{ob}) + \sum_{s=1}^{s\#,k,IJ} t_{s,k,IJ}^{ic} \tag{7-1}$$

where $l\#, k, IJ$ is the index of the journey leg, on which passengers experience fail-to-board delays and $s\#, k, IJ$ is the corresponding interchange station. The journey time, $t_{k,Ii}^{entry-ic}$ is dependent on the route and it is applicable only if fail-to-board delays are experienced at the interchange station within the AM peak.

Knowing the entry time stamp of passenger q at station Ii of the origin superstation ($T_{q,Ii}^{entry}$) and the journey time from there to the congested interchange station on route k ($t_{k,Ii}^{entry-ic}$), the arrival time (at the departure platform of subsequent journey leg, $T_{q,k,IJ}^{plat-ic}$) can be calculated as:

$$T_{q,k,IJ}^{plat-ic} = T_{q,Ii}^{entry} + t_{k,Ii}^{entry-ic} \tag{7-2}$$

Recalling the concept of journey time adjustments for superstations (**Section 5.4**), it is possible understand the following: The available dataset from smart card contains the entry times of passengers at entering different stations; but the interchange station is the same for all passengers regardless their entry station. In fact, equation (7-2) can be interpreted as an adjustment of $T_{q,Ii}^{entry}$ with $t_{k,Ii}^{entry-ic}$ to the congested interchange station. Therefore, $T_{q,k,IJ}^{plat-ic}$ does not contain the index Ii and hence the values originally coming from different entry stations can be aggregated.

Once the arrival time of passenger q at the congested interchange station on route k ($T_{q,k,IJ}^{plat-ic}$) is obtained, it can be explicitly known, which time interval t this arrival time falls into, and hence the corresponding fail-to-board delay ($t_{q,k,IJ}^{fail}$) can be inferred as described in **Section 6.3**.

The Scheduled Journey Time (SJT) of each route – without considering fail-to-board delays – can be understood based on train timetables and station layouts (cf. equation (3-13)). As the SJT is calculated in the frequency based context (see **Section 7.6**); it is same for all passengers on a given route k, and hence $t_{k,IJ}^{SJT}$ does not contain the index q.

Knowing the SJT on each route k between the centroids of superstations I and J ($t_{k,IJ}^{SJT}$) and the fail-to-board delay of each passenger q on that route ($t_{q,k}^{fail}$), the superstation-to-superstation equivalent of the Expected Journey Time can be calculated as:

$$t_{q,k,IJ}^{\mathrm{EXP}} = t_{k,IJ}^{SJT} + t_{q,k,IJ}^{fail} \tag{7-3}$$

The Expected Journey Time of routes ($t_{q,k,IJ}^{\mathrm{EXP}}$) is supposed to be used as an additional condition to update the previously obtained route choice probabilities ($\omega_{r,IJ}$). At this point it is important to note that these two variables have different identifiers for the routes, r and k respectively. Analogously to the previous chapters, while r is used for mixture components, k indicates the index of actual LU routes. The reason why it is necessary to have two distinct variable identifiers is that it is explicitly unknown which mixture component r corresponds to which actual LU route k. In order to proceed with the updating methodology, and hence to use $\omega_{r,IJ}$ and $t_{q,k,IJ}^{\mathrm{EXP}}$ in the same Bayesian framework; it is necessary to make an a priori assumption on the matching of r and k. This is discussed in **Section 7.3**.

7.3 Matching mixture components with actual routes

In **Section 3.4.1** it was proposed to match the mixture components (r) with the actual LU routes (k) in the ascending order of their corresponding journey times. However, looking at the results in **Chapter 5**, especially **Case 3** (**Table 5-20**) showed that applying this approach does not always give the correct match.

In this chapter, apart from the mean journey time of mixture components ($\mu_{r,IJ}^{MIX}$), also their standard deviation ($\sigma_{r,IJ}^{MIX}$) is examined. The reason for this can be resumed in the following: Staying at the example of **Case 3**, it is expected, that the direct route has smaller standard deviation, because passengers have only one journey leg; which means no interchange time, less total wait time and expectedly, less variance of the on-board time. On the other hand, on the indirect route, having three journey legs, the variation of all these values are larger. Furthermore, due to the difference in the fail-to-board delays of different passengers q at the congested interchange station ($t_{q,k}^{fail}$), the expected journey time ($t_{q,k,IJ}^{\text{EXP}}$, cf. equation (7-3)) has an even larger standard deviation. Based on these considerations, it makes sense to match the mixture component having smaller standard deviation with the direct route and the one having larger standard deviation with the indirect route.

It is acknowledged that this rule may not hold for all OD pairs. It may occur that, if the travel time on the second or third journey leg of the indirect route is less variable than the travel time on the direct route, the overall variance is smaller for the indirect route. However, assuming that the **Jubilee** trains do not exhibit great variance on the **Wembley Park** – **Finchley Road** segment, and knowing that passengers do experience fail-to-board delays for the same segment of the **Metropolitan** line; the above consideration could be acceptable for the case study in this chapter.

To have a more advanced matching method, both the mean ($\mu_{r,IJ}^{MIX}$) and standard deviation ($\sigma_{r,IJ}^{MIX}$) of the finite mixture results need to be considered; and the corresponding mean and standard deviation of the actual LU routes needs to be modelled based on the distribution of their journey time components (i.e. $t_{1,k,Ii}^{acc}$, $t_{l,k}^{wait}$, $t_{l,k}^{ob}$, $t_{s,k}^{ic}$ and $t_{L,k}^{egr}$, (Wahaballa et al., 2017)) as well as of the and fail-to-board delays on them. This is beyond the scope of this thesis.

Having made the above described considerations to match the mixture components with the actual LU routes, all variables referring to routes are identified with index r in the Bayesian framework (see **Section 7.4**). There, the random variable of Expected Journey Time used is denoted as $\delta_{q,r,IJ}^{\text{EXP}}$.

7.4 Updating the posterior route choice probabilities

The Bayesian framework for updating the route choice probabilities understood from the finite mixture model (cf. **Chapter 5**) can be formulated as follows. As described in **Section 3.2**, $choice_{qr,IJ}$ denotes the elementary event that passenger q has chosen route r; and $\Delta_{q,IJ}$ the elementary event that the CCOJT of passenger q is $\delta_{q,IJ}^{\text{CCOJT}}$ (cf. **Section 5.4**). Furthermore, in this chapter $\Delta_{qr,IJ}$ defined, which denotes the elementary event that the Expected Journey Time of passenger q is $\delta_{q,r,IJ}^{\mathrm{EXP}}$, given that he/she chooses route r and his/her entry time is $T_{q,Ii}^{entry}$.

Having defined these elementary events in a probability space, the objective is to obtain $Pr\left(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ}\right)$, which is the probability that passenger q has chosen route r, given that his/her CCOJT is $\delta_{q,IJ}^{\mathrm{CCOJT}}$ and his/her expected journey time on route r is $\delta_{q,r,IJ}^{\mathrm{EXP}}$. Following the steps in Fu (2014), this can be expressed as:

$$Pr\left(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ}\right) = \frac{\Pr\left(\Delta_{q,IJ}|choice_{qrIJ},\Delta_{qr,IJ}\right)\cdot\Pr\left(choice_{qr,IJ}\right)}{\sum_{r\epsilon R}\Pr\left(\Delta_{q,IJ}|choice_{qr,IJ},\Delta_{qr,IJ}\right)\cdot\Pr\left(choice_{qr,IJ}\right)} \tag{7-4}$$

In this context, $\Pr\left(choice_{qr,IJ}\right)$ is the probability that passenger q has chosen route r without any knowledge on his journey time. According to formula (3-5), this was associated with the component proportion understood from the finite mixture model ($\omega_{r,IJ}$). In the Bayesian framework is the prior.

Furthermore, $\Pr\left(\Delta_{q,IJ}|choice_{qr,IJ},\Delta_{qr,IJ}\right)$ is the likelihood of observing $\delta_{q,IJ}^{\mathrm{CCOJT}}$ given that q has chosen r and the expected journey time on that route was $\delta_{q,r,IJ}^{\mathrm{EXP}}$. Following the concept in Fu (2014), this likelihood can be explained in the following way: Let $\delta_{qr,IJ}$ denote the journey time distribution of passenger q on route r. Similarly to the considerations in **Section 3.2.2**, it is assumed, that also the PDF of $\delta_{qr,IJ}$, $f\ \left(\delta_{qr,IJ}\right)$ follows a Gaussian distribution. There, the mean corresponds to the expected journey time of passenger q on route r ($\delta_{q,r,IJ}^{\mathrm{EXP}}$). Furthermore, it is assumed that the standard deviation of each passenger q on route r is the same as the standard deviation of route r estimated with the finite mixture model ($\sigma_{r,IJ}^{MIX}$, cf. **Section 3.2**). This way, the likelihood that the journey time of q would be $\delta_{q,IJ}^{\mathrm{CCOJT}}$ can be associated with the probability density of the PDF $f\ \left(\delta_{qr,IJ}\right)$ having the above described parameters:

$$\Pr\left(\Delta_q|choice_{qr,IJ},\Delta_{qr,IJ}\right) \approx f\ \left(\delta_{qr,IJ} = \delta_{q,IJ}^{\mathrm{CCOJT}}|\delta_{q,r,IJ}^{\mathrm{EXP}},\sigma_{r,IJ}^{MIX}\right) \tag{7-5}$$

Substituting formulae (3-5) and (7-5) into equation (7-4), $Pr\left(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ}\right)$ can be expressed as:

$$Pr\left(choice_{qr,IJ}\,|\Delta_{q,IJ},\Delta_{qr,IJ}\right) = \frac{\omega_{r,IJ}\cdot f\ \left(\delta_{qr}=\delta_{q,IJ}^{\text{CCOJT}}\,|\delta_{q,r,IJ}^{\text{EXP}},\sigma_{r,IJ}^{MIX}\right)}{\sum_{r\epsilon R}\omega_{r,IJ}\cdot f\ \left(\delta_{qr}=\delta_{q,IJ}^{\text{CCOJT}}\,|\delta_{q,r,IJ}^{\text{EXP}},\sigma_{r,IJ}^{MIX}\right)} \tag{7-6}$$

In equation (7-6) the additional condition, which updates the prior knowledge on route choice ($\sigma_{r,IJ}^{MIX}$) is the random variable of the expected journey time: $\delta_{q,r,IJ}^{\text{EXP}}$. As it was presented through equation (7-3), $t_{q,k,IJ}^{\text{EXP}}$ is a function of the SJT ($t_{k,IJ}^{SJT}$) and of the fail-to-board delay ($t_{q,k,IJ}^{fail}$). In this chapter the focus was on the variation of $t_{q,k,IJ}^{fail}$ depending on the entry time of passengers ($T_{q,Ii}^{entry}$) and on the route (k); and the variation of $t_{k,IJ}^{SJT}$ was not examined in depth as the frequency-based approach was followed. However, doing a more detailed analysis on the variation of $t_{k,IJ}^{SJT}$ in function of its components (3-13) could further improve the route choice updates $Pr\left(choice_{qr,IJ}\,|\Delta_{q,IJ},\Delta_{qr,IJ}\right)$ also when fail-to-board delay is not accounted for.

7.5 Inferring reference time of routes and aggregate route choice probabilities

As the result of **Section 7.4** the posterior probabilities were calculated for each passenger q on route r, conditional on their CCOJT and Expected Journey Time ($Pr\left(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ}\right)$). From the practical point of view (i.e. public transport operators who are interested in passenger flows, cf. **Section 2.1**), it is necessary to have further information on the aggregate values of mean, standard deviation and choice probabilities of each route.

To obtain these values naïve inference is used following the concept in Fu (2014). The logic in naïve inference is that if the posterior probability for route r is higher than for route r' ($Pr\left(choice_{qr}=r|\Delta_q,\Delta_{qr}\right)\geq\left(choice_{qr}=r'|\Delta_q,\Delta_{qr}\right)$), then it is more likely, that passenger q chooses route r. Going one step ahead, the inference could be drawn that the actual choice of passenger q is the route with the highest posterior probability. This way the assignment function can be defined for the naïve inference as:

$$\zeta(q)=argmax_{r\in R}\left(Pr\left(choice_{qr}=r|\Delta_q,\Delta_{qr}\right)\right) \tag{7-7}$$

The output of the assignment function is the route label r for each passenger q. With this assignment it is possible to obtain N_R sub-datasets within the $CCOJT_{IJ}$ dataset – denoted by $CCOJT_{r,IJ}^{naive}$ – where in each sub-dataset those passengers can be found who are assigned to route r with the naïve inference.

$$CCOJT_{r,IJ}^{naive} = \{\delta_{q,IJ}^{\mathrm{CCOJT}} : \zeta(q) = r\} \qquad (7\text{-}8)$$

Having obtained this, it is possible to understand the updated values of mean ($\mu_{r,IJ}^{upd}$) and standard deviation ($\sigma_{r,IJ}^{upd}$) on each route r as the mean and standard deviation of the corresponding sub-dataset $CCOJT_{r,IJ}^{naive}$. Furthermore, the total number of passengers in the $CCOJT_{IJ}$ dataset who chooses route r is equal to the size of the $CCOJT_{r,IJ}^{naive}$ sub-dataset, denoted by $n_{r,IJ}^{naive}$:

$$n_{r,IJ}^{naive} = \left|CCOJT_{r,IJ}^{naive}\right| \qquad (7\text{-}9)$$

From this, the aggregate choice probabilities of route r can be calculated as:

$$\omega_{r,IJ}^{upd} = \frac{n_{r,IJ}^{naive}}{n_{IJ}^{CCOJT}} \qquad (7\text{-}10)$$

In order to understand how the updated estimates can be associated with the actual LU routes (cf. **Section 7.3**), the estimated mean journey time ($\mu_{r,IJ}^{upd}$) is compared to the reference time of the actual LU routes ($t_{r,IJ}^{REF}$[29]):

$$t_{r,IJ}^{REF} = \frac{\sum_{q=1}^{n_{IJ}^{CCOJT}} t_{q,r,IJ}^{\mathrm{EXP}} \cdot \psi_{r,IJ}^{naive}}{n_{r,IJ}^{naive}} \qquad (7\text{-}11)$$

where $\psi_{r,IJ}^{naive}$ is a dummy variable, $\psi_{r,IJ}^{naive} = 1$ if, $\zeta(q) = r$, otherwise $\psi_{r,IJ}^{naive} = 0$. As explained earlier, depending on the difference in fail-to-board delays, $t_{q,r,IJ}^{\mathrm{EXP}}$ may vary across each passenger q on route r (7-3); therefore there is the necessity to calculate their average, which gives the reference time of route r: $t_{r,IJ}^{REF}$. In addition to the journey times, also the updated estimates of aggregate route choice ($\omega_{r,IJ}^{upd}$) are compared with $\omega_{r,IJ}^{RODS}$ (cf. **Section 3.6.3**).

[29] As the results of the finite mixture model had been previously matched with the actual LU routes, in this section for all variables in identifier r is used

7.6 Comparison with Fu (2014)

In this chapter the method in Fu (2014) was applied to update the prior knowledge on route choice ($\omega_{r,IJ}$) with the additional condition on the expected journey time of passengers ($\delta^{\mathrm{EXP}}_{q,r,IJ}$). The difference between his model and the method applied in this thesis can be resumed in the following:

- The approach to infer the Expected Journey Time ($t^{\mathrm{EXP}}_{q,k,IJ}$) (see **Section 7.6.1**)
- The assumptions made for the fail-to-board delay ($t^{fail}_{q,k,IJ}$) (see **Section 7.6.2**)

7.6.1 The approach to infer the Expected Journey Time

In Fu (2014) the input of $T^{entry}_{q,Ii}$ was applied to infer the train that passenger q could catch, and hence the arrival time at the exit station was modelled for each route k. Following this, $t^{\mathrm{EXP}}_{q,k,IJ}$ was obtained as the difference between $T^{entry}_{q,Ii}$ and the modelled exit time. It is important to note that it requires to represent each individual train (schedule-based context).

The method proposed in this thesis uses the input of $T^{entry}_{q,Ii}$ to infer fail-to-board delays of passenger q ($t^{fail}_{q,k,IJ}$) and hence $t^{\mathrm{EXP}}_{q,k,IJ}$ is calculated according to equation (7-3). In this approach trains are still represented with their frequency (frequency-based context), as $T^{entry}_{q,Ii}$ is not applied to infer a train, but to identify the time interval t, in which the passenger arrives at the congested interchange station. Having the possibility to remain in the frequency-based context means less computational time (cf. **Section 2.3.3**).

Obviously, as the smart card data processing and hence the intended application of the model is still off-line, computational time is not a relevant issue (cf. **Section 1.2**). However, in future, once the technology arrives at real-time data processing, it will be advantageous to have models that can estimate route choice at a lower computational time, especially if there is the need to apply the model for many OD pairs.

7.6.2 The assumptions made for the fail-to-board delay

Fu (2014) considered fail-to-board delays as a component of the wait time. He made the assumption that for each journey leg l of each route k, half of the passengers can board the first, half of them the second train, and calculated the reference time of routes ($t^{REF}_{r,IJ}$) accordingly . The problem with this assumption consists in the following: Firstly, the fail-to-board delay is not constant along the metro lines: the stations near the terminus are not congested at all (e.g. origin stations in **Case 3**), while stations closer to the LU inner zone can have extreme congestion (e.g. origin stations in **Case 2**). Secondly, the fail-to-board delay is not constant within the AM peak: it occurs only in the peak of peak (8:00-9:15).

In order to take into account the variation of the fail-to-board delay along the metro line and within the AM peak, in this thesis it was inferred from actual data on passenger flows and on train capacities, following the method described in **Section 6.3**, and the reference time of routes ($t_{r,IJ}^{REF}$) was calculated accordingly This could bring some improvement to the consideration of fail-to-board delays, however due to the deficiency of RODS data applied for passenger flows (cf. **Section 6.5.1**) and to the further assumptions (cf. **Section 6.6**), it still cannot give a perfect picture on it.

7.7 Case study on the London Underground

In **Section 6.6**, the passenger flows were compared to the line capacities at the origin and interchange stations of the three case studies, and hence the fail-to-board delays were calculated at the congested stations. Based on that, in it was understood, that for **Case 3**, fail-to-board delays occur along the indirect route, at the **Wembley Park** interchange station on the eastbound **Metropolitan** line (**Figure 6-15** and **Table 6-13**). In this section, this additional information on fail-to-board delays will be used to update the route choice probabilities understood from the finite mixture model.

Case 3 Jubilee West – Jubilee Central

Expected Journey Time of passengers on each route

The entry time of each passenger q ($T_{q,Ii}^{entry}$), as well as their OJT ($\delta_{q,(Ii)(jj)}^{\mathrm{OJT}}$) is known from Oyster data (cf. **Section 3.6.1**) for the **Jubilee West** – **Jubilee Central** superstation-to-superstation OD pair (**Figure 5-22**). As it was examined in **Section 5.6**, none of the 286 Oyster records were considered outliers.

From each station of the origin superstation (i.e. **Stanmore**, **Canons Park**, **Queensbury** and **Kingsbury** stations on the **Jubilee** line) the journey time was calculated to the departure platform of the congested interchange station (i.e. **Metropolitan** line eastbound at **Wembley Park** station. This is route $k = 2$) ($t_{2,Ii}^{entry-ic}$) according to equation (7-1) (see **Table 7-1**), where the journey time components were understood from timetables and station layouts (cf. **Section 3.6.2**). Following this, for each passenger q, the arrival time at the interchange station ($T_{q,2,IJ}^{plat-ic}$) was calculated according to equation (7-2).

Table 7-1 Journey time from entry ticket gate to congested station platform for **Jubilee West**– **Jubilee Central**

Journey time from entry ticket gate to congested station platform (minutes)	
$t_{2,Ii}^{entry-ic}$	
From\To	**Wembley Park**
Stanmore	13.1
Canons Park	10.9
Queensbury	9.1
Kingsbury	6.0

Knowing $T_{q,2,IJ}^{plat-ic}$, it can be explicitly understood, which 15 minute time interval t (cf. **Section 6.3**) it falls into; and hence the corresponding average fail-to-board delays on this route ($t_{q,2,IJ}^{fail}$) can be inferred for each passenger q based on **Table 6-13**. Results show that passengers arriving at **Wembley Park** between 8:15 and 8:30 experience 7 minutes of fail-to-board delay in average, while those who arrive in the time intervals 8:00-8:15 and 8:30-8:45 this delay is 4 minutes. Before and 8:00 and after 8:45 no fail-to-board delay occur.

As it was discussed earlier, the SJT between the superstation centroids ($t_{k,IJ}^{SJT}$) on the direct route (**Jubilee** line) is 36.3 minutes, while it is 33.3 minutes on the indirect route (**Jubilee** – **Metropolitan** – **Jubilee** via **Wembley Park** and **Finchley Road**). Knowing the SJT ($t_{k,IJ}^{SJT}$) and the fail-to-board delay of each passenger q on each route k ($t_{q,k,IJ}^{fail}$), the Expected Journey Time ($t_{q,k,IJ}^{\mathrm{EXP}}$) can be calculated according to equation (7-3).

The CCOJT of each passenger q ($\delta_{q,IJ}^{\mathrm{CCOJT}}$, inferred from $\delta_{q,(Ii)(jj)}^{\mathrm{OJT}}$ cf. **Section 5.4**) and their Expected Journey Time on each route k ($t_{q,k,IJ}^{\mathrm{EXP}}$) are compared on **Figure 7-3**. From there, it can be understood, that when there is no congestion on the indirect route, $t_{q,k,IJ}^{\mathrm{EXP}}$ is lower (33.3 minutes), but when there is congestion, $t_{q,k,IJ}^{\mathrm{EXP}}$ is higher (37.3 minutes and 40.3 minutes depending on the time interval) than the direct route (36.3 minutes). Similarly, for $\delta_{q,IJ}^{\mathrm{CCOJT}}$, it can be observed that they are relatively lower (less than 50 minutes) for those passengers who are supposed to arrive at **Wembley Park** before

8:00 on route 2, and they are exceedingly high (up to 70 minutes) for those passengers who are supposed to arrive there between 8:00 and 9:15.

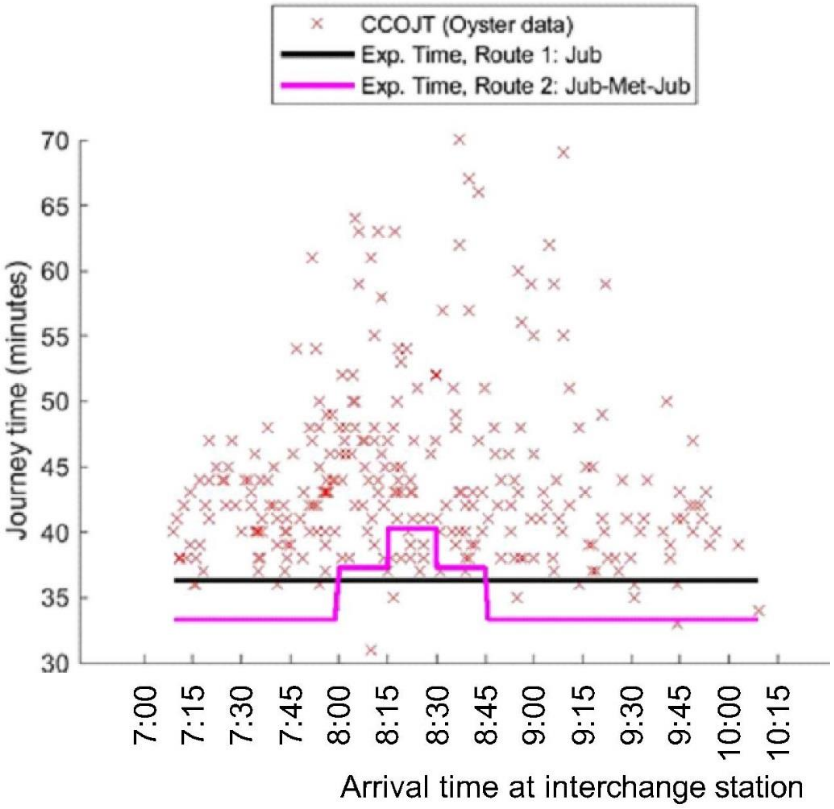


Figure 7-3 Comparison between $\boldsymbol{\delta}_{\boldsymbol{q},\boldsymbol{IJ}}^{\text{CCOJT}}$ and $\boldsymbol{t}_{\boldsymbol{q},\boldsymbol{k},\boldsymbol{IJ}}^{\text{EXP}}$ for **Jubilee West– Jubilee Central**

It is important to note that the adjusted $\delta_{q,IJ}^{\text{CCOJT}}$ observations are almost always considerably larger than the inferred $t_{q,k,IJ}^{\text{EXP}}$ values. This can be due to the lack of information on the actual values of the components of $t_{q,k,IJ}^{\text{EXP}}$: $t_{k,IJ}^{SJT}$ and $t_{q,k,IJ}^{fail}$ (cf. equation (7-3)). The lower value for $t_{k,IJ}^{SJT}$, can be due to the underestimation of on-board ($t_{l,k,IJ}^{ob}$) or wait ($t_{l,k,IJ}^{wait}$) time by using timetables as a data source (cf. **Section 3.6.2.1**) and not taking into consideration the possible service delays; as well as due the insufficient information on passengers walk speed through crowded passageways (cf. **Section 3.6.2.2**) and hence the underestimated values of access ($t_{1,k,IJ}^{acc}$) egress ($t_{L,k,IJ}^{egr}$) interchange ($t_{s,k,IJ}^{ic}$) times (cf. equation (3-13)). The lower value for $t_{q,k,IJ}^{fail}$ can be associated with the insufficient information on the relationship between the nominal and actual capacity of trains.

Furthermore, it is also unknown, what proportion of passengers are on those services of the Metropolitan line, which stops at Wembley Park (cf. **Section 6.6**).

Updating route choice probabilities with the additional condition on Expected Journey Time

The finite mixture model applied on the CCOJT distribution of the Jubilee West – Jubilee Central superstation-to-superstation OD pair gave the results of 78.5% of proportion for mixture component 1 ($\omega_{1,IJ}^{MIX}$) and 21.5% of proportion for mixture component 2 ($\omega_{2,IJ}^{MIX}$) (cf. **Table 5-20** and **Figure 5-25**). As explained earlier (cf. equation (3-5)), these proportions were associated with the priors of route choice in the Bayesian framework.

In order to apply equation (7-6) correctly to update these priors in a Bayesian framework, it is crucial to know which mixture component (r) corresponds to which actual route (k). Results show that mixture component 1 has 41.5 minutes of mean journey time ($\mu_{1,IJ}^{MIX}$) with 3.6 minutes of standard deviation ($\sigma_{1,IJ}^{MIX}$); and mixture component 2 has 52.9 minutes of mean journey time ($\mu_{2,IJ}^{MIX}$) and 7.8 minutes of standard deviation ($\sigma_{2,IJ}^{MIX}$) (cf. **Table 5-18**). Based on the considerations in **Section 7.3** it is assumed that the mixture component with smaller standard deviation ($r = 1$) corresponds to the direct route, while the component with the greater standard deviation ($r = 2$) corresponds to the indirect route.

As explained in **Section 7.4** the likelihood function is associated with the probability density function of the journey time distribution of passenger q on route r ($f\ \left(\delta_{qr}\right)$, cf. equation (7-5)). It was assumed, that it follows a Gaussian distribution with the mean of $\delta_{q,r,IJ}^{\mathrm{EXP}}$ and standard deviation of $\sigma_{r,IJ}^{MIX}$ (i.e. 3.6 minutes for route 1 and 7.8 minutes for route 2). With these parameters, the probability density was calculated at the value of $\delta_{qr,IJ} = \delta_{q,IJ}^{\mathrm{CCOJT}}$ for each passenger q on each route r.

Having obtained both the priors and the likelihood function, the route choice conditional on CCOJT and on Expected Journey Time ($Pr\left(choice_{qr,IJ}\middle|\Delta_{q,IJ},\Delta_{qr,IJ}\right)$) was calculated according to equation (7-6) for each passenger q on each route r. These results were compared to the route choice conditional on CCOJT ($\Pr\left(choice_{qr,IJ}\middle|\Delta_{q,IJ}\right)$), which was calculated according to equation (3-6) from the finite mixture model.

From **Figure 7-4** it can be understood that in general, the update according to the additional condition on the expected journey time of passengers (function of fail-to-board delays) made the posterior probabilities $Pr(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ})$ lower for the direct route (a) and higher for the indirect route (b) than the posterior probabilities conditional only on their CCOJTs ($\Pr(choice_{qr,IJ}|\Delta_{q,IJ})$). Among them the greatest difference is for those passengers whose CCOJT is around 41-45 minutes and their inferred arrival time at **Wembley Park** (if they chose route 2) is in the time interval of 8:15-8:30. Based only on their CCOJT, $\Pr(choice_{qr,IJ}|\Delta_{q,IJ})$ is quite low for route 2; however knowing that in that time interval, 7 minutes of fail-to-board delay is expected on that route, it is more likely that these longer CCOJTs correspond to the fact of experiencing fail-to-board delays and hence $Pr(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ})$ became much higher. In other words, having a CCOJT record and the additional information on the corresponding fail-to-board delay, the posterior probability that this record belongs to the congested indirect route becomes higher.

Understanding the characteristics of the actual LU routes in **Case 3** it is expected that very small CCOJT values (31-34 minutes) correspond to the indirect route in uncongested time intervals (before 8:00 or after 8:45), as without having fail-to-board delay that route has shorter journey time (cf. **Figure 7-3**). Furthermore, knowing that the direct route has a smaller standard deviation (cf. **Section 7.3**), it is more likely that the CCOJTs around the expected journey time ($t_{q,k,IJ}^{\mathrm{EXP}}$) of the direct route (35-39 minutes) correspond to the direct route. Finally, the very large CCOJTs (40-70 minutes) are likely to correspond the indirect route in congested time intervals (between 8:00 and 8:45)

The results of this case study reflected quite well what was expected based on the characteristics of the actual LU routes. For the very small CCOJT values (31-34 minutes) it gave around 80-85% for $Pr(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ})$ of route 1. For the CCOJTs around the expected journey time of route 1 (35-39 minutes), the $Pr(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ})$ for route 1 was above 85%. For CCOJTs between 40 and 48 minutes the $Pr(choice_{qr,Ij}|\Delta_{q,IJ},\Delta_{qr,IJ})$ varies significantly depending on the arrival time of the passenger at the interchange station (hence his corresponding fail-to-board delay). The actual turning point is at the CCOJTs of 44-45 minutes, as the $Pr(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ})$ is around 0.5 for both routes. For CCOJTs of 49 minutes and above, the $Pr(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ})$ is very low for route 1.

Additionally, it is important to discuss the question of the quickest possible time on each route as a physical constraint. In this specific example, from **Table 5-19**, it can be understood, the total on-board time on the direct route is 31 minutes (assuming that due to timetable constraints, trains cannot arrive earlier than the scheduled arrival time). Considering the passenger does not need to wait for the metro service and runs very fast at the entry and exit station, it is expected to have 32 minutes as the quickest possible time for route 1. Therefore, looking at the CCOJT record of 31 minutes (**Figure 7-4**), it is physically impossible, that this journey could occur on route 1. At this point, the question of quickest possible journey time was not included in the model, but was proposed as a further improvement.

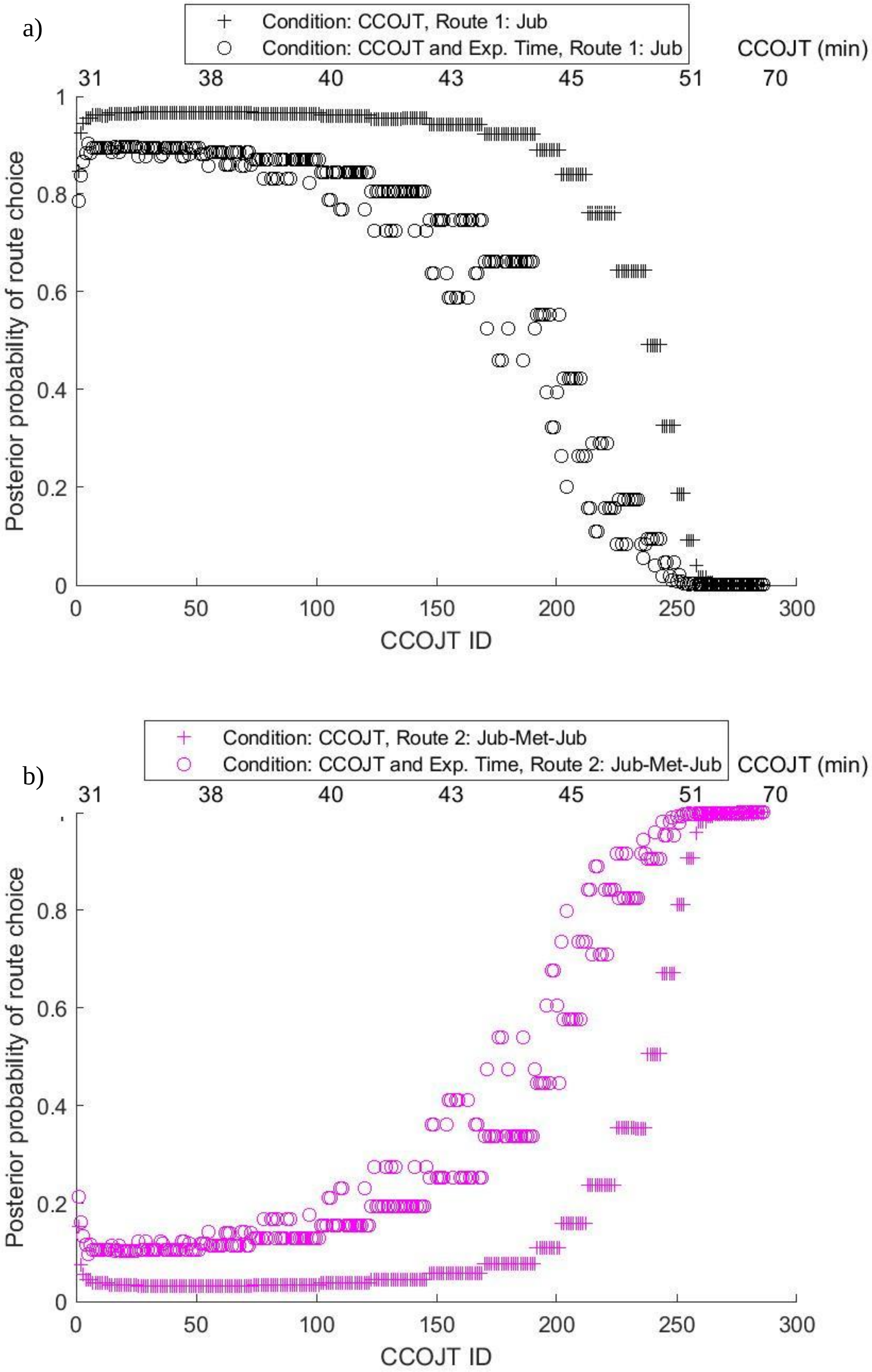


Figure 7-4 Comparison between $\mathbf{Pr}\left(\boldsymbol{choice}_{qr}|\boldsymbol{\Delta}_{q}\right)$ and $\boldsymbol{Pr}\left(\boldsymbol{choice}_{qr}|\Delta_{q},\Delta_{qr}\right)$ for Jubilee West – Jubilee Central; a) Route 1, b) Route 2

Aggregate route choice

After having obtained the updated posterior probabilities for each passenger q on route r ($Pr\left(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ}\right)$), the corresponding aggregate values for each route r ($\mu_{r,IJ}^{UPD}$, $\sigma_{r,IJ}^{UPD}$ and $\omega_{r,IJ}^{UPD}$) are inferred with naïve inference as described in **Section 7.5**. The results are resumed in **Table 7-2**.

Table 7-2 Aggregate results updated according to the additional information on fail-to-board delays for Jubilee West– Jubilee Central

Label	Updated with fail-to-board		
r []	$\mu_{r,IJ}^{UPD}$ [min]	$\sigma_{r,IJ}^{UPD}$ [min]	$\omega_{r,IJ}^{UPD}$ [%]
1	40.3	2.6	67.8%
2	51.5	6.3	32.2%

In **Table 7-3**, the estimated mean values ($\mu_{r,IJ}^{UPD}$) were compared to the reference time of routes ($t_{r,IJ}^{REF}$, cf. (7-11)) and the corresponding aggregate route choice probabilities ($\omega_{r,IJ}^{UPD}$) with the values understood from RODS data ($\omega_{r,IJ}^{RODS}$). Regarding journey times, it is possible to observe, that the updated values ($\mu_{r,IJ}^{UPD}$) are lower for both routes than the results understood form the finite mixture model ($\mu_{r,IJ}^{MIX}$) and hence it is closer to the reference time of routes ($t_{r,IJ}^{REF}$). Interestingly, $t_{r,IJ}^{REF}$ (which includes the fail-to-board delays on the congested route, cf. equation (7-3) and (7-11)) is very similar for the two routes (36.3 and 36.2 minutes respectively). Concerning route choice, the updated aggregate route choice probabilities ($\omega_{r,IJ}^{UPD}$) are lower for route 1 and higher for route 2 than the corresponding results from the finite mixture model ($\omega_{r,IJ}^{MIX}$) and hence they are further from the RODS results ($\omega_{r,IJ}^{RODS}$).

Table 7-3 Matching the updated results with the actual London Underground routes for Jubilee West– Jubilee Central

Blue: Updated results, superstation OD pairs, Purple: Mixture results, superstation OD pairs, Green: actual LU routes

Mixture label	Journey Time (min)			Route Choice (%)			Route label	Route matched				
	Updated	Mixture	Reference	Updated	Mixture	RODS		Line 1	IC 1	Line 2	IC 2	Line 3
r	$\mu_{r,IJ}^{UPD}$	$\mu_{r,IJ}^{MIX}$	$t_{r,IJ}^{REF}$	$\omega_{r,IJ}^{UPD}$	$\omega_{r,IJ}^{MIX}$	$\omega_{r,IJ}^{RODS}$	k	$l = 1$	$s = 1$	$l = 2$	$s = 2$	$l = 3$
1	40.3	41.5	36.3	67.8%	78.5%	89.0%	1	Jub				
2	51.5	52.9	36.2	32.2%	21.5%	11.0%	2	Jub	Wembley Park	Met	Finchley Road	Jub

7.8 Discussions

As it was pointed out in the case study (cf. **Figure 7-3**), the Expected Journey Time of passengers ($t_{q,k,IJ}^{\mathrm{EXP}}$) is quite low with respect to their CCOJT ($\delta_{q,IJ}^{\mathrm{CCOJT}}$). Due to this, the posterior probabilities ($Pr\left(choice_{qr,IJ}|\Delta_{q,IJ},\Delta_{qr,IJ}\right)$) were underestimated for route 1 and overestimated for route 2 (cf. equation (7-6)). A possible improvement could be to estimate $t_{l,k,IJ}^{ob}$ and $t_{l,k,IJ}^{wait}$ based on the actual departure/arrival time of trains and to model $t_{1,k,IJ}^{acc}$, $t_{L,k,IJ}^{egr}$ and $t_{s,k,IJ}^{ic}$ taking into consideration that also the walk speed depends on the crowding along the station passageways (see **Section 8.2.5**).

Additionally, even though the considerations that were made for matching the results of the finite mixture model with the actual LU routes (cf. **Section 7.3**) could work for this specific case (**Case 3**, **Jubilee West** – **Jubilee Central**); to make it applicable to all OD pairs of a metro network, it would require more advanced statistical method.

Chapter 8
Concluding remarks and further work

8.1 Conclusions

The core problem discussed in this thesis is to estimate route choice in complex metro networks from smart card data at origin destination (OD) level. Using the finite mixture model in this context and applying it for certain station-to-station OD pairs of the London Underground (LU) a few issues emerged. To address them, this thesis proposed:

- a route choice set generation model that can find automatically the set of reasonable routes for any type of OD pair of a metro network;
- the superstation representation of metro networks and the corresponding spatial aggregation of Observed Journey Times (OJT) understood from smart card data and
- A methodology to adjust the OJTs as well as the route choice estimates of the finite mixture model according to the additional information the on fail-to-board delay at the origin and interchange stations

As it follows, the lessons learnt regarding these models are discussed.

8.1.1 Route choice set generation in complex metro networks

This thesis has made a forward step in developing a model that can automatically find the set of reasonable routes for any type of OD pair of a complex metro network. This set was obtained based on the generalised costs of routes (Raveau et al., 2014). Applying the proposed model on the LU inner zone (cf. **Chapter 4**), the following was understood:

- In average a route can be considered reasonable, if its generalised cost is no more than 1.18 times the generalised cost of the shortest route.
- There are OD pairs, which have unreasonable routes with a lower generalised cost proportion (1.09-1.17).
- There are also OD pairs, which have reasonable routes with higher generalised cost proportion (1.21-1.23).

In essence, results showed that it is not possible to generate the set of reasonable routes purely based on their generalised cost as they also depend on OD specific attributes. Therefore, based on the case studies in **Chapter 5** it was further understood that a lower attribute cut-off is expected for OD pairs with longer minimum journey time as well as for those which have more available directions at the origin and/or destination station; and that a higher attribute cut-off is expected when the choice of passengers is between a direct route and an express line. Although these dependencies were pointed put and illustrated, the explicit formulation of the model was not yet made, but it is suggested for further research (see **Section 8.2.1**).

8.1.2 Introducing the concept of superstations

This thesis introduced the concept of superstations with the purpose to overcome the data availability issues of station-to-station OD pairs in the context of route choice estimation. Applying that for the case study OD pairs of the LU:

- Certain number of OD pairs could be grouped
    - 4-5 OD pairs in cases where only the origin stations could be grouped,
    - 20 OD pairs in cases where both origin and destination station could be grouped and
- the sample size of OJTs could be increased
    - 1.6-2.7 times in case only the origin stations could be grouped,
    - 8.2 times in case both origin and destination station could be grouped.

The finite mixture model applied on this larger dataset

- brought better convergence for all case study OD pairs and
- gave closer results to the actual LU routes in most cases.

Having tested the superstation representation for certain OD pairs, the question was raised, whether it is possible to find automatically all groups of OD pairs for which the superstation representation is applicable. This discussed in **Section 8.2.2**.

The main objective for station grouping in this thesis was to increase the sample size of OJTs for route choice estimation with the finite mixture model. The main limitation here is that the concept of superstations is applicable only for the OD pairs, where the first/last journey leg for all reasonable routes is on the same line. However, in reality, there are many OD pairs, where the first/last journey leg is on different lines, and hence the concept of superstations in not applicable. Therefore, it was understood, that for a more comprehensive understanding on route choice, it is advisable to set multiple objectives for station grouping. This is discussed in **Section 8.2.3**.

8.1.3 Adjustment according to fail-to-board delays

The model proposed in this thesis accounted for the fail-to-board delay at the origin and interchange station.

For the fail-to-board delay at the origin station the OJTs were further adjusted in the process of their aggregation. A crucial point here was that deducting all fail-to-board delays from the OJT dataset would be an optimistic, but not realistic view about the likelihood of passengers' boarding, assuming that fail-to-board delays had not occurred. As a consequence of this, the finite mixture model applied on that dataset of CCOJTs had worse convergence (i.e. greater difference between the solutions for different settings of tolerance thresholds) compared to when it was applied on the CCOJTs without adjustment according to fail-to-board delays. In order to address this question, a different approach is proposed in **Section 8.2.4**, which follows the concept presented in **Chapter 7**.

Apart from this, there are also other possible reasons why the adjustment according to fail-to-board delays did not bring the expected benefit:

- The quasi-dynamic approach could infer only the average fail-to-board delay values for the 15 minute time intervals, but not their actual distribution.
- The assumption made for the actual capacity of trains may not be realistic

In case the fail-to-board delay occurs at the interchange station, the route choice estimates of the finite mixture model were updated with this additional information in a Bayesian framework. It was understood that the updated route choice probabilities did not give a better match to the actual LU results than without the adjustment due to the following limitations of the model:

- The Scheduled Journey Time (SJT) of routes was inferred from timetables and station layouts
- The fail-to-board delay was inferred from the Rolling Origin Destination Survey (RODS) data

To overcome these limitations, an improved model is proposed in **Section 8.2.5**.

8.2 Further work

In response to these issues (cf. **Section 8.1**) it is proposed for further research:

- an improved route choice generation model that uses as an input both route specific and OD specific attributes;

- an algorithm that finds all groups of OD pairs, for which the superstation representation is applicable
- the extension of the concept of superstations for nearby stations;
- consideration of fail-to-board delays at the origin station with the concept of updating priors of route choice probabilities
- an improved model that can rely more on automated data sources

As it follows, these improvements are presented.

8.2.1 An improved route choice set generation model

An improved route choice set generation algorithm is proposed for further research that – in addition to the route specific attributes (generalised cost) – accounts also for the OD specific attributes, such as the presence of a direct route, the number of available directions at the origin and destination station, OD minimum travel time and the presence of an express line (cf. **Section 5.3.3**). In order to properly formulate this function, it is necessary to calibrate the model with additional types of OD pairs:

1) with one reasonable route (i.e. find the attribute cut-off value between the shortest and the second shortest route);
2) with reasonable routes that have three or more journey legs and
3) from/to/between LU outer zones

The first type of OD pair needs to be examined, as it is expected that in those cases the generalised cost proportion of the second shortest route is quite high, therefore it is not included in the observed choice set of passengers. These are mostly OD pairs with a direct route and the indirect routes do not have any attractive attributes (cf. **Section 4.7.4.2**). Analogously, for the second type of OD pair one would be interested to know what attractive attributes a route with three or more journey legs has that passengers still consider that option, when routes with less journey legs are also available (cf. **Section 5.6 Case 3**). Finally, the third type of OD pair exhibits a case, where the on-board times are much longer and the interchange times are relatively shorter with respect to the total journey time. Therefore, it is expected that in these cases different cut-off values will be between the reasonable and unreasonable routes (cf. **Section 2.3.1**).

As the program code is already ready, it could be applied automatically for all OD pairs of the LU. It requires only the input of the RODS data for validation.

Another question to discuss is that in the process of route choice set generation, the oversimplified assumption was made, that the generalised cost of a route depends only on their journey time and interchange properties (cf. (4-2)). However in reality, it may occur that there are other routes that passengers consider in their choice set with higher journey time and less favourable interchange properties, because those routes have other attributes that attract them (e.g. they look shorter on the map or they have new LU fleet). This issue could be treated by including those attributes in the generalised cost function.

8.2.2 An algorithm to find all superstations in a metro network

Once the previously proposed route choice set generation model (cf. **Section 8.2.1**) is calibrated, it could be applied for all OD pairs of the LU to find those which have multiple reasonable routes. Following this – in line with the **Definition** of superstations (cf. **Section 5.3.1**) – an algorithm can be written that can automatically find those groups of OD pairs, for which the first/last journey leg is on the same line both among the reasonable routes and across the OD pairs. This can be done easily as the relevant information is already stored in the network model of the LU (cf. **Section 4.4**).

8.2.3 Extension of the concept of superstations for nearby stations

It is proposed to extend the concept of superstations for nearby stations so that – in addition to the objective of data availability – it can comply with the following objectives:

- Including in the model that a passenger has a set of attractive entry/exit stations near his/her true origin/destination
- Reducing network complexity

These objectives are especially relevant in the LU network, as in the inner zone many stations are within walking distance (12 minutes (Transport for London, 2010)). A typical example for this are the stations around the **Bank**/**Monument** station complex in the City of London. Identifying all groups of stations with these properties, could lead to the superstation map of the LU (see **Figure 8-1**). This is equivalent to a simplified network representation (**Figure 8-2**), and applying a Transit Assignment Model (TAM) on that could give a more comprehensive picture on route choice with significantly less computational time.

The objective of data availability is not fully in line with the objectives for grouping nearby stations; as while the former requires that the stations should be on the same line (regardless their distance), the latter requires that they should be in physical proximity (regardless whether they are on the same line). Therefore, the overall aim of the modeller is to find the optimal scenario of these objectives.

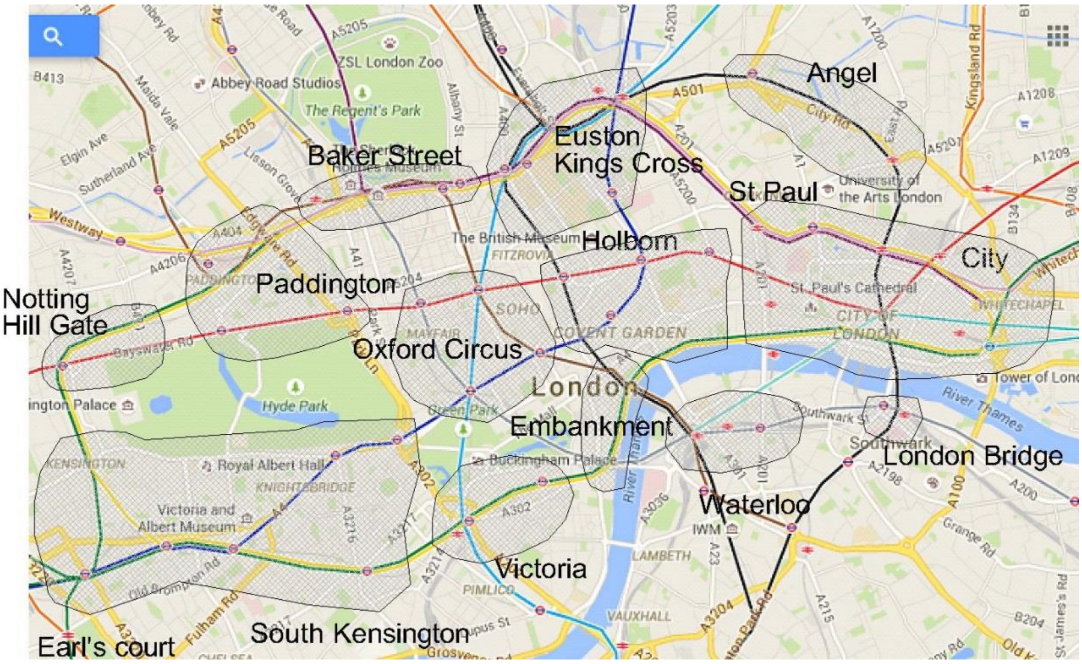


Figure 8-1 The superstation map of the London Underground according to the extended concept of superstations

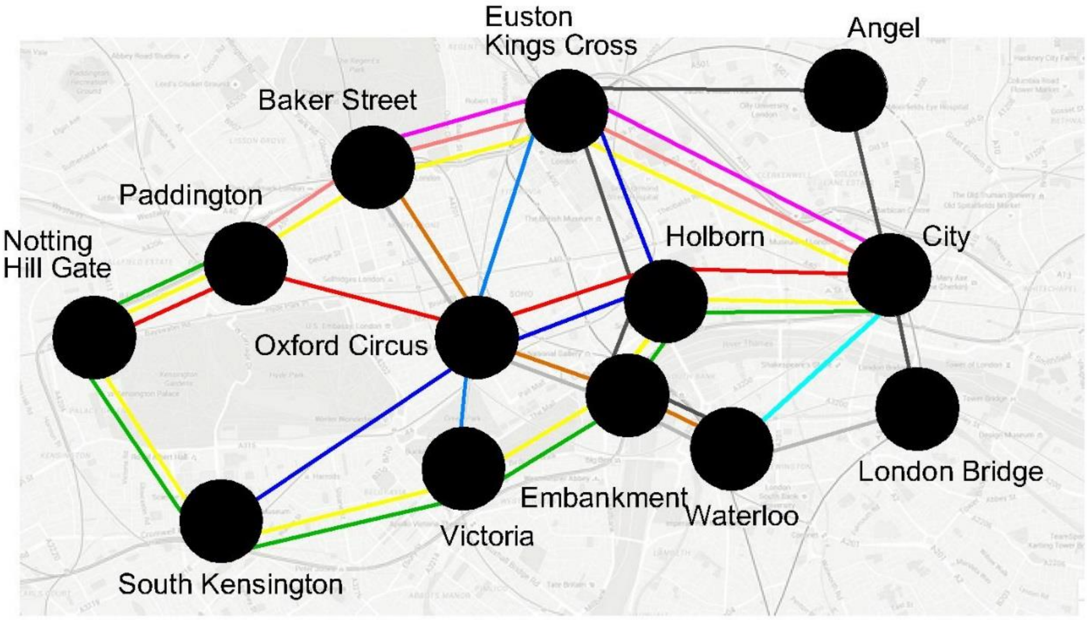


Figure 8-2 Application of the superstation map of the LU for transit assignment modelling

Looking at, for example, the **Victoria** – **Holborn** OD pair: According to the extended concept of superstations, the origin station: **Victoria** can be grouped with the nearby **St James' Park** station and the destination station: **Holborn** with the **Tottenham Court Road**, **Leicester Square**, **Covent Garden**, **Chancery Lane** and **Temple** stations. This way the route choice set between the origin and destination superstations will be

1) Victoria – Central (via **Oxford Circus**)
2) Victoria – Piccadilly (via **Green Park**)
3) **Circle** – Northern (via **Embankment**)
4) **Circle**

This example illustrates that with the inclusion of nearby stations, there are additional routes in the choice set of passengers. Practically this means that if the true origin of the passenger is somewhere near **Victoria** station and the true destination is somewhere near **Holborn** station – in addition to the previously discussed routes (route 1 and 2) – it might be convenient for him/her to take the **Circle** line to **Embankment** and interchange to the **Northern line** (route 3), or to **Temple** and walk to the true destination (route 4).

With this concept of superstation representation, route choice could be estimated in the following way: For some station-to-station OD pairs, the route choice can be known explicitly, as there is only one reasonable route (e.g. **Victoria** – **Temple**, **Victoria** – **Covent Garden**), while for other OD pairs there are more reasonable routes (e.g. **Victoria** – **Holborn**, **Victoria** – **Leicester Square**). In the latter case – similarly to the original concept of superstations – the route choice could be estimated with the finite mixture models.

Through this specific example it could be further understood that the main objective of this superstation representation is to give a more comprehensive picture on route choice by including those passengers who decide to take a direct service (i.e. **Circle** line) to a station near their true destination (i.e. **Temple**). For this OD pair it was not possible to comply with the objective of data availability, because – unlike the original concept of superstations – it was not possible to group the **Victoria** station with the other stations on the Victoria line as they are not in its physical proximity.

It is important to note that the station grouping presented on **Figure 8-1** is one, but not the only possibility to group the stations. For example – looking into the opposite direction – **Victoria** station could be grouped with **Sloane Square** and **Hyde Park** stations instead of **St James's Park** station. Therefore this extended concept of superstations would necessitate not only to group the nearby stations, but also to find the optimal configuration among all possibilities. Furthermore, dealing with nearby stations, would require additional information on the true origin/destination of passengers and the surface walk time to/from their corresponding entry/exit stations. These tasks would bring the necessity to apply geospatial analysis.

8.2.4 A different approach for considering fail-to-board delays at the origin station

It is proposed to extend the concept in **Chapter 7** also for the cases, when the fail-to-board delay occurs at the origin stations. In other words, looking at the outcomes of **Chapter 6** and **Chapter 7** it was understood that updating the route choice probabilities is a more justifiable approach than adjusting the OJT values.

Within the framework of this thesis, for **Case 1** and **Case 2** the common pattern is that for all reasonable routes the first journey leg is on the same line. In this context, in the time intervals when fail-to-board delays occur at the origin station, it occurs on all routes. Therefore in this specific case the fail-to-board delay (δ^{fail}) and hence the expected journey time (δ^{EXP}) is not anymore a function of the chosen route (r), only of the entry time ($T_{q,Ii}^{entry}$). This way the original Bayesian network structure (cf. **Figure 7-1**) can be further simplified (see **Figure 8-3**).

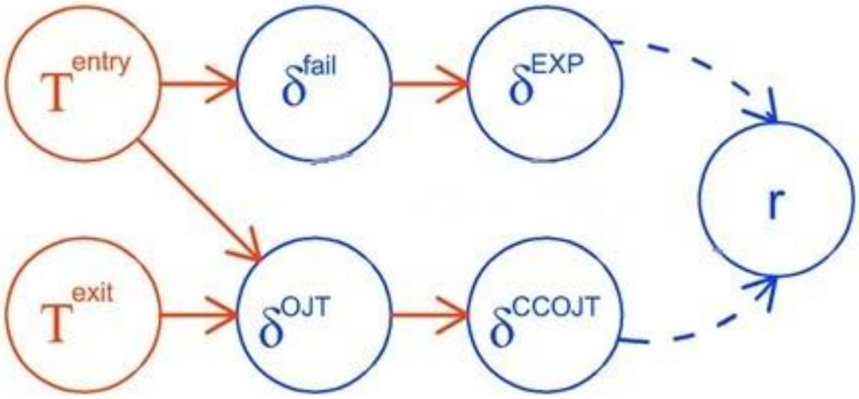


Figure 8-3 Bayesian network structure to illustrate passengers probabilistic route choices – in case fail-to-board delay is the same on all routes

In these cases, equation (7-1) and (7-2) simplifies to the following

$$T_{q,IJ}^{plat-o} = T_{q,Ii}^{entry} + t_{1,Ii}^{acc} \qquad (8\text{-}1)$$

where $T_{q,IJ}^{plat-o}$ denotes the arrival time stamp of passenger q at the departure platform of the origin station. From this, the time interval (t) of arriving at the congested platform can be explicitly known, and hence the corresponding fail-to-board delay $(t_{q,IJ}^{fail}$, cf. **Section 6.3**) and Expected Journey Time $(t_{q,k,IJ}^{\mathrm{EXP}}$, cf. equation (7-3)) can be inferred. Based on this, the updated route choice probabilities $(Pr(choice_{qr,IJ}|\Delta_{q,IJ}, \Delta_{qr,IJ})$, cf. **Section 7.4**) as well as the reference time of routes $(t_{r,IJ}^{REF})$ and the aggregate values of route choice $(\omega_{r,IJ}^{upd})$ (cf. **Section 7.5**) can be obtained.

Looking at equation (7-6) it is expected that in the time intervals of the peak of peak (i.e. 8:15-9:00) – when fail-to-board delays occur at the origin station and hence $t_{q,k,IJ}^{\mathrm{EXP}}$ is higher for both routes – the $Pr(choice_{qr,IJ}|\Delta_{q,IJ}, \Delta_{qr,IJ})$ will be closer to each other for the different routes than it is for the priors; where the higher $\delta_{q,IJ}^{\mathrm{CCOJT}}$ observations were associated by default with the route that has longer SJT $(t_{k,IJ}^{SJT})$, which meant that the corresponding route choice probabilities $(\omega_{r,IJ})$ were also higher (c.f. **Figure 5-13** and **Figure 5-19**).

8.2.5 Relying on automatically collected data for inferring crowding and the journey time of routes

It is proposed for further research to improve the adjustment and matching process, so that it can completely move away from manual surveys and rely more on automatically collected data sources:

- Estimate passenger flows also from smart card data instead of RODS data
- Infer on-board and wait time of trains from their actual departure and arrival times instead of timetables
- Infer access egress interchange (AEI) times in function of station crowding

In theory, from smart card data both the OD demand and route choice can be inferred, hence passenger flow can be determined for all links. However – in the congested case – this would require to solve this problem as a TAM for the entire network, which would require much additional modelling work due to the size and complexity of the LU network.

The actual departure/arrival time of trains can be understood from the TfL open data website (cf. **Section 3.6.2.1**). Staying at the frequency based context it would be possible to infer from that data source the distribution of on-board and wait times. Furthermore, also for the AEI times more realistic estimates could be obtained, if they were modelled not only with their mean values based on the information on station layouts (cf. **Section 3.6.2.2**); but also with their distribution which is in relation with the crowding experienced along the station passageways (following the concept in **Section 6.3**).

References

Arthur, D. and Vassilvitskii, S. 2007. k-means++: The advantages of careful seeding. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*: Society for Industrial and Applied Mathematics, pp.1027-1035.

Azevedo, J.A., Madeira, J.J.E.S., Martins, E.Q.V. and Pires, F.M.A. 1990. A shortest paths ranking algorithm.

Azevedo, J.A., Santos Costa, M.E.O., Silvestre Madeira, J.J.E.R. and Vieira Martins, E.Q. 1993. An algorithm for the ranking of shortest paths. *European Journal of Operational Research.* **69**(1), pp.97-106.

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.

Başar, G. and Bhat, C. 2004. A parameterized consideration set model for airport choice: an application to the San Francisco Bay Area. *Transportation Research Part B: Methodological.* **38**(10), pp.889-904.

Bekhor, S., Toledo, T. and Prashker, J. 2006. Implementation issues of route choice models in path-based algorithms. In: *11th international conference on travel behaviour research, Kyoto, Japan*.

Bellman, R. 1958. On a routing problem. *Quarterly of applied mathematics.* **16**(1), pp.87-90.

Ben-Akiva, M., Bergman, M., Daly, A.J. and Ramaswamy, R. 1984. Modeling inter-urban route choice behaviour. In: *Proceedings of the 9th International Symposium on Transportation and Traffic Theory, VNU Press, Utrecht*, pp.299-330.

Ben-Akiva, M. and Bierlaire, M. 1999. Discrete Choice Methods and their Applications to Short Term Travel Decisions. In: Hall, R.W. ed. *Handbook of Transportation Science.* Boston, MA: Springer US, pp.5-33.

Bergantino, A.S., Capurso, M., Dekker, T. and Hess, S. 2019. Allowing for Heterogeneity in the Consideration of Airport Access Modes: The Case of Bari Airport. *Transportation Research Record: Journal of the Transportation Research Board.* p036119811882512.

Bovy, P.H.L. 2009. On Modelling Route Choice Sets in Transportation Networks: A Synthesis. *Transport Reviews.* **29**(1), pp.43-68.

Cantillo, V. and Ortúzar, J.d.D. 2005. A semi-compensatory discrete choice model with explicit attribute thresholds of perception. *Transportation Research Part B: Methodological.* **39**(7), pp.641-657.

Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A. 1996. A modified Logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In: *Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Pergamon, Lyon, France*. pp.697–711.

Cascetta, E. and Papola, A. 2001. Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand. *Transportation Research Part C: Emerging Technologies.* **9**(4), pp.249-263.

Cea, J.d. and Fernández, E. 1993. Transit Assignment for Congested Public Transport Systems: An Equilibrium Model. *Transportation Science.* **27**(2), pp.133-147.

Cepeda, M., Cominetti, R. and Florian, M. 2006. A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria. *Transportation Research Part B: Methodological.* **40**(6), pp.437-459.

Chamundeswari, G., Varma, G.P. and Satyanarayana, C. 2012. An experimental analysis of k-means using Matlab. *International Journal of Engineering Research Technology (IJERT) ISSN*.

Chan, J. 2007. *Rail Transit OD Matrix Estimation and Journey Time Reliability Metrics Using Automated Fare Data*. Master of Science in Transportation thesis, Massachusetts Institute of Technology.

Chriqui, C., & Robillard, P. 1975. Common bus lines. *Transportation Science*. **9**, pp.115-121.

Chu, K. and Chapleau, R. 2010. Augmenting Transit Trip Characterization and Travel Behavior Comprehension. *Transportation Research Record: Journal of the Transportation Research Board*. **2183**, pp.29-40.

Chu, K.K. 2010. *Leveraging data from a smart card automatic fare collection system for public transit planning*. thesis, École Polytechnique de Montréal.

Cominetti, R. and Correa, J. 2001. Common-Lines and Passenger Assignment in Congested Transit Networks. *Transportation Science*. **35**(3), pp.250-267.

Connors, R.D. and Watling, D.P. 2014. Assessing the Demand Vulnerability of Equilibrium Traffic Networks via Network Aggregation. *Networks and Spatial Economics*. **15**(2), pp.367-395.

Cui, A. 2006. *Bus passenger origin-destination matrix estimation using automated data collection systems*. thesis, Massachusetts Institute of Technology.

de la Barra, T., Perez, B. and Anez, J. 1993. Multidimensional path search and assignment. In: *PTRC Summer Annual Meeting, 21st, 1993, University of Manchester, United Kingdom*.

Dempster, A.P., Laird, N.M. and Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. **39**(1), pp.1-38.

Dial, R.B. 1971. A probabilistic multipath traffic assignment model which obviates path enumeration. *Transportation Research*. **5**(2), pp.83-111.

Dijkstra, E.W. 1959. A note on two problems in connexion with graphs. *Numerische mathematik*. **1**(1), pp.269-271.

Domencich, T. and McFadden, D. 1975. Urban travel demand; a behavioural analysis.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*, pp.226-231.

Fifer, S., Rose, J. and Greaves, S. 2014. Hypothetical bias in Stated Choice Experiments: Is it a problem? And if so, how do we deal with it? *Transportation Research Part A: Policy and Practice*. **61**, pp.164-177.

Fiorenzo-Catalano, S., Van Nes, R. and Bovy, P.H. 2004. Choice set generation for multi-modal travel analysis. *European journal of transport and infrastructure research EJTIR, 4 (2).*

Ford, L.R.J. 1956. *Network flow theory*. RAND CORP SANTA MONICA CA.

Forgy, E.W. 1965. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*. **21**, pp.768-769.

Freemark, Y. 2013. *Assessing Journey Time Impacts of Disruptions on London's Piccadilly Line*. Master of Science in Transportation thesis, Massachusetts Institute of Technology.

Frigge, M., Hoaglin, D.C. and Iglewicz, B. 1989. Some Implementations of the Boxplot. *The American Statistician*. **43**(1), pp.50-54.

Frühwirth-Schnatter, S. 2006. *Finite mixture and Markov switching models*. Springer Science & Business Media.

Fu, Q. 2014. *Modelling route choice behaviour with incomplete data: an application to the London Underground*. PhD thesis, University of Leeds.

Fu, Q., Liu, R. and Hess, S. 2012. A Review on Transit Assignment Modelling Approaches to Congested Networks: A New Perspective. *Procedia - Social and Behavioral Sciences.* **54**, pp.1145-1155.

Fujiyama, T. and Tyler, N. 2010. Predicting the walking speed of pedestrians on stairs. *Transportation Planning and Technology.* **33**(2), pp.177-202.

Gan, L. and Jiang, J. 1999. A Test for Global Maximum. *Journal of the American Statistical Association.* **94**(447), pp.847-854.

Gaundry, M.J.I. and Dagenais, M.G. 1979. The dogit model. *Transportation Research Part B: Methodological.* **13**(2), pp.105-111.

Gentile, G. and Noekel, K. 2016. *Modelling public transport passenger flows in the era of intelligent transport systems.*

Gordillo, F. 2006. *The Value of Automated Fare Collection Data for Transit Planning: An Example of Rail Transit OD Matrix Estimation.* Master of Science in Transportation thesis, Massachusetts Institute of Technology.

Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H.M. and Attanucci, J.P. 2013. Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. *Transportation Research Record: Journal of the Transportation Research Board.* **2343**(-1), pp.17-24.

Guo, W. and Lu, X. 2016. London underground: Neighbourhood centrality and relation to urban geography. In: *2016 IEEE International Smart Cities Conference (ISC2), 12-15 Sept. 2016*, pp.1-7.

Guo, Z. 2008. *Transfers and path choice in urban public transport systems*. thesis, Massachusetts Institute of Technology.

Guo, Z. 2011. Mind the map! The impact of transit maps on path choice in public transit. *Transportation Research Part A: Policy and Practice.* **45**(7), pp.625-639.

Hart, P.E., Nilsson, N.J. and Raphael, B. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics.* **4**(2), pp.100-107.

Hassan, M.N., Rashidi, T.H., Waller, S.T., Nassir, N. and Hickman, M.J.J.o.P.T. 2016. Modeling Transit Users Stop Choice Behavior: Do Travelers Strategize? *Journal of Public Transportation.* **19**(3), p6.

Hickman, M.D. and Bernstein, D.H. 1997. Transit Service and Path Choice Models in Stochastic and Time-Dependent Networks. *Transportation Science.* **31**(2), pp.129-146.

Holleczek, T., Anh, D.T., Yin, S., Jin, Y., Antonatos, S., Goh, H.L., Low, S. and Shi-Nash, A. 2015. Traffic Measurement and Route Recommendation System for Mass Rapid Transit (MRT). In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia.* 2788590: ACM, pp.1859-1868.

Hong, L., Li, W. and Zhu, W. 2017. Assigning Passenger Flows on a Metro Network Based on Automatic Fare Collection Data and Timetable. *Discrete Dynamics in Nature and Society.* **2017**, pp.1-10.

Hong, S.-P., Min, Y.-H., Park, M.-J., Kim, K.M. and Oh, S.M. 2015. Precise estimation of connections of metro passengers from Smart Card data. *Transportation.* **43**(5), pp.749-769.

Hörcher, D., Graham, D.J. and Anderson, R.J. 2017. Crowding cost estimation with large scale smart card and vehicle location data. *Transportation Research Part B: Methodological.* **95**, pp.105-125.

Horowitz, J.L. and Louviere, J.J. 1995. What is the role of consideration sets in choice modeling? *International Journal of Research in Marketing.* **12**(1), pp.39-54.

Ingvardson, J.B., Nielsen, O.A., Raveau, S. and Nielsen, B.F. 2018. Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis. *Transportation Research Part C: Emerging Technologies.* **90**, pp.292-306.

Jafari, E. and Boyles, S.D. 2016. Improved bush-based methods for network contraction. *Transportation Research Part B: Methodological.* **83**, pp.298-313.

Jánošíková, Ľ., Slavík, J. and Koháni, M. 2014. Estimation of a route choice model for urban public transport using smart card data. *Transportation Planning and Technology.* **37**(7), pp.638-648.

Kaufman, L. and Rousseeuw, P.J. 2009. *Finding groups in data: an introduction to cluster analysis.* John Wiley & Sons.

Kieu, L.-M., Bhaskar, A. and Chung, E. 2015a. A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data. *Transportation Research Part C: Emerging Technologies.* **58**, pp.193-207.

Kieu, L.M., Bhaskar, A. and Chung, E. 2015b. Passenger Segmentation Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems.* **16**(3), pp.1537-1548.

Koutsopoulos, H.N., Noursalehi, P., Zhu, Y. and Wilson, N.H.M. 2017. Automated data in transit: Recent developments and applications. In: *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 26-28 June 2017*, pp.604-609.

Kuhlman, W. 2015. *The construction of purpose-specific OD matrices using public transport smart card data*. thesis, TU Delft.

Kurauchi, F., Bell, M.G.H. and Schmöcker, J.-D. 2003. Capacity Constrained Transit Assignment with Common Lines. *Journal of Mathematical Modelling and Algorithms.* **2**(4), pp.309-327.

Kusakabe, T. and Asakura, Y. 2014. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies.* **46**, pp.179-191.

Kusakabe, T., Iryo, T. and Asakura, Y. 2010. Estimation method for railway passengers' train choice behavior with smart card transaction data. *Transportation.* **37**(5), pp.731-749.

Leahy, C., Batley, R. and Chen, H. 2015. Toward an automated methodology for the valuation of reliability. *Journal of Intelligent Transportation Systems.* **20**(4), pp.334-344.

Lee, M. and Sohn, K. 2015. Inferring the route-use patterns of metro passengers based only on travel-time data within a Bayesian framework using a reversible-jump Markov chain Monte Carlo (MCMC) simulation. *Transportation Research Part B: Methodological.* **81**, pp.1-17.

Lee, S., Hickman, M. and Tong, D. 2013. Development of a temporal and spatial linkage between transit demand and land-use patterns. *Journal of Transport and Land Use.* **6**(2), p33.

Li, Z. and Hensher, D.A. 2011. Crowding and public transport: A review of willingness to pay evidence and its relevance in project appraisal. *Transport Policy.* **18**(6), pp.880-887.

Luo, D., Cats, O. and van Lint, H. 2017. Constructing Transit Origin–Destination Matrices with Spatial Clustering. *Transportation Research Record: Journal of the Transportation Research Board.* **2652**(1), pp.39-49.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* **1**, p17.

Manski, C.F. 1977. The structure of random utility models. *Theory and Decision.* **8**(3), pp.229-254.

Martínez, F., Aguila, F. and Hurtubia, R. 2009. The constrained multinomial logit: A semi-compensatory choice model. *Transportation Research Part B: Methodological.* **43**(3), pp.365-377.

McLachlan, G. and Peel, D. 2000. *Finite mixture models: Wiley series in probability and mathematical statistics*. John Wiley & Sons, Inc.

McLachlan, G.J. and Krishnan, T. 2007. *The EM algorithm and extensions.* John Wiley & Sons.

Meschini, L., Gentile, G. and Papola, N. 2007. A frequency based transit model for dynamic traffic assignment to multimodal networks. In: *17th International Symposium on Transportation and Traffic Theory, 23rd-25th July 2007, London, United Kingdom*.

Morency, C., Trépanier, M. and Agard, B. 2007. Measuring transit use variability with smart-card data. *Transport Policy.* **14**(3), pp.193-203.

Munizaga, M.A. and 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**, pp.9-18.

Nádudvari, T., Liu, R. and Balijepalli, C. 2016. The reasonable route choice set in large and complex metro networks; an implementation of the K-shortest path algorithm for the London Underground. In: *Proceedings of the 21st International Conference of Hong Kong Society for Transportation Studies, 10th-12th December 2016, Hong Kong, China*. Hong Kong Society for Transportation Studies, pp.247-254.

Nádudvari, T., Liu, R. and Hess, S. 2015. Modelling passengers' route choice behaviour on the London Underground: application of two choice modelling approaches. In: *In: Proceedings of the 47th Annual Conference of Universities' Transport Study Group, 5th-7th January 2015, London, United Kingdom*.

Nassir, N., Hickman, M. and Ma, Z.-L. 2015a. Activity detection and transfer identification for public transit fare card data. *Transportation.* **42**(4), pp.683-705.

Nassir, N., Hickman, M., Malekzadeh, A. and Irannezhad, E. 2015b. Modeling Transit Passenger Choices of Access Stop. *Transportation Research Record: Journal of the Transportation Research Board.* **2493**(1), pp.70-77.

Nassir, N., Hickman, M., Malekzadeh, A. and Irannezhad, E. 2016. A utility-based travel impedance measure for public transit network accessibility. *Transportation Research Part A: Policy and Practice.* **88**, pp.26-39.

Nguyen, S. and Pallottino, S. 1988. Equilibrium traffic assignment for large scale transit networks. *European Journal of Operational Research.* **37**(2), pp.176-186.

Nuzzolo, A., Crisalli, U. and Rosati, L. 2012. A schedule-based assignment model with explicit capacity constraints for congested transit networks. *Transportation Research Part C: Emerging Technologies.* **20**(1), pp.16-33.

Nuzzolo, A., Russo, F. and Crisalli, U. 2001. A Doubly Dynamic Schedule-based Assignment Model for Transit Networks. *Transportation Science.* **35**(3), pp.268-285.

Opsahl, T., Agneessens, F. and Skvoretz, J. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks.* **32**(3), pp.245-251.

Ortega-Tong, M.A. 2013. Classification of London's Public Transport Users Using Smart Card Data. *MIT Thesis.*

Ortúzar, J.d.D. and Willumsen, L.G. 2011. *Modelling transport*. Chichester, West Sussex: John Wiley & Sons.

Parkes, S.D., Jopson, A. and Marsden, G. 2016. Understanding travel behaviour change during mega-events: Lessons from the London 2012 Games. *Transportation Research Part A: Policy and Practice.* **92**, pp.104-119.

Paul, E.C. 2010. *Estimating train passenger load from automated data systems : application to London Underground*. Master of Science in Transportation thesis, Massachusetts Institute of Technology.

Pelletier, M.-P., Trépanier, M. and Morency, C. 2011. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies.* **19**(4), pp.557-568.

Prato, C.G. 2009. Route choice modeling: past, present and future research directions. *Journal of Choice Modelling.* **2**(1), pp.65-100.

Ramming, M. 2002. *Network Knowledge and Route Choice. PhD.* MIT, Cambridge, MA, Unpublished.

Raveau, S., Guo, Z., Muñoz, J.C. and Wilson, N.H.M. 2014. A behavioural comparison of route choice on metro networks: Time, transfers, crowding, topology and socio-demographics. *Transportation Research Part A: Policy and Practice.* **66**, pp.185-195.

Richardson, S. and Green, P.J. 1997. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* **59**(4), pp.731-792.

Rodriguez, A. and Laio, A. 2014. Clustering by fast search and find of density peaks. *Science.* **344**(6191), pp.1492-1496.

Ross, L. 2017. *Measuring Travel Time Reliability under Disruption Conditions for the London Underground*. Master of Science thesis, University of Leeds.

Schmöcker, J.-D. 2006. *Dynamic Capacity Constrained Transit Assignment* Thesis submitted for the degree of Doctor of Philosophy thesis, Imperial College London.

Schmöcker, J.-D., Bell, M.G.H. and Kurauchi, F. 2008. A quasi-dynamic capacity constrained frequency-based transit assignment model. *Transportation Research Part B: Methodological.* **42**(10), pp.925-945.

Schmöcker, J.-D., Fonzone, A., Shimamoto, H., Kurauchi, F. and Bell, M.G.H. 2011. Frequency-based transit assignment considering seat capacities. *Transportation Research Part B: Methodological.* **45**(2), pp.392-408.

Seaborn, C., Attanucci, J. and Wilson, N. 2009. Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data. *Transportation Research Record: Journal of the Transportation Research Board.* **2121**, pp.55-62.

Silva, D. 2017. *Quantifying Journey Time Variability and Understanding its Impact On Passenger Decision Making For Bus Travel In London*. Master of Science thesis, University of Leeds.

Spiess, H. and Florian, M. 1989. Optimal strategies: A new assignment model for transit networks. *Transportation Research Part B: Methodological.* **23**(2), pp.83-102.

Sun, G., Xiong, Y. and Zhu, Y. 2017. How the Passengers Flow in Complex Metro Networks? In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA*. 3085527: ACM, pp.1-6.

Sun, L., Lee, D.-H., Erath, A. and Huang, X. 2012. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China*. 2346519: ACM, pp.142-148.

Sun, L., Lu, Y., Jin, J.G., Lee, D.-H. and Axhausen, K.W. 2015. An integrated Bayesian approach for passenger flow assignment in metro networks. *Transportation Research Part C: Emerging Technologies.* **52**, pp.116-131.

Sun, Y. and Xu, R. 2012. Rail Transit Travel Time Reliability and Estimation of Passenger Route Choice Behavior. *Transportation Research Record: Journal of the Transportation Research Board.* **2275**, pp.58-67.

Swait, J. 2001. A non-compensatory choice model incorporating attribute cutoffs. *Transportation Research Part B: Methodological.* **35**(10), pp.903-928.

Swait, J. and Ben-Akiva, M. 1987. Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological.* **21**(2), pp.91-102.

Tamblay, S., Galilea, P., Iglesias, P., Raveau, S. and Muñoz, J.C. 2016. A zonal inference model based on observed smart-card transactions for Santiago de Chile. *Transportation Research Part A: Policy and Practice.* **84**, pp.44-54.

Tamblay, S., Muñoz, J.C. and Ortúzar, J.d.D. 2018. Extended Methodology for the Estimation of a Zonal Origin-Destination Matrix: A Planning Software Application Based on Smartcard Trip Data. *Transportation Research Record: Journal of the Transportation Research Board.*

Teklu, F. 2007. A Stochastic Process Approach for Frequency-based Transit Assignment with Strict Capacity Constraints. *Networks and Spatial Economics.* **8**(2-3), pp.225-240.

Tirachini, A., Hurtubia, R., Dekker, T. and Daziano, R.A. 2017. Estimation of crowding discomfort in public transport: Results from Santiago de Chile. *Transportation Research Part A: Policy and Practice.* **103**, pp.311-326.

Tirachini, A., Sun, L., Erath, A. and Chakirov, A. 2016. Valuation of sitting and standing in metro trains using revealed preferences. *Transport Policy.* **47**, pp.94-104.

Tong, C.O. and Wong, S.C. 1999. A stochastic transit assignment model using a dynamic schedule-based network. *Transportation Research Part B: Methodological.* **33**(2), pp.107-121.

Tong, C.O.C.O. 1986. A schedule-based transit network model.

Transport for London. 2010. *Measuring Public Transport Accessibility Levels, (PTALs), Summary.* Transport for London.

Transport for London. 2017. *Review of the TfL WiFi pilot.* London, UK: Transport for London.

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.

Utsunomiya, M., Attanucci, J. and Wilson, N. 2006. Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning. *Transportation Research Record: Journal of the Transportation Research Board.* **1971**, pp.119-126.

Viggiano, C., Koutsopoulos, H.N., Wilson, N.H.M. and Attanucci, J. 2016. Journey-based characterization of multi-modal public transportation networks. *Public Transport.* **9**(1-2), pp.437-461.

Vovsha, P. 1997. Application of Cross-Nested Logit Model to Mode Choice in Tel Aviv, Israel, Metropolitan Area. *Transportation Research Record: Journal of the Transportation Research Board.* **1607**, pp.6-15.

Wahaballa, A.M., Kurauchi, F., Yamamoto, T. and Schmöcker, J.-D. 2017. Estimation of Platform Waiting Time Distribution Considering Service Reliability Based on Smart Card Data and Performance Reports. *Transportation Research Record: Journal of the Transportation Research Board.* **2652**, pp.30-38.

Wang, W., Attanucci, J. and Wilson, N. 2011. Bus passenger origin-destination estimation and related analyses using automated data collection systems.

Watling, D.P., Rasmussen, T.K., Prato, C.G. and Nielsen, O.A. 2018. Stochastic user equilibrium with a bounded choice model. *Transportation Research Part B: Methodological.* **114**, pp.254-280.

Wen, C.-H. and Koppelman, F.S. 2001. The generalized nested logit model. *Transportation Research Part B: Methodological.* **35**(7), pp.627-641.

Whelan, G.A. and Crockett, J. 2009. An Investigation of the Willingness to Pay to Reduce Rail Overcrowding. In: *International Choice Modelling Conference 2009.*

Williams, H.C.W.L. 1977. On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit. *Environment and Planning A: Economy and Space.* **9**(3), pp.285-344.

Wu, J.H., Florian, M., & Marcotte, P. 1994. Transit equilibrium assignment: a model and solution algorithms. *Transportation Science.* **28**, pp.193-203.

Xu, R.-h. and Zhou, F. 2012. Model of Passenger Flow Assignment for Urban Rail Transit Based on Entry and Exit Time Constraints. *Transportation Research Record: Journal of the Transportation Research Board.* **2284**(-1), pp.57-61.

Xu, X., Xie, L., Li, H. and Qin, L. 2018. Learning the route choice behavior of subway passengers from AFC data. *Expert Systems with Applications.* **95**, pp.324-332.

Yen, J.Y. 1971. Finding the K Shortest Loopless Paths in a Network. *Management Science.* **17**(11), pp.712-716.

Young, M.A. and Blainey, S.P. 2017. Development of railway station choice models to improve the representation of station catchments in rail demand models. *Transportation Planning and Technology.* **41**(1), pp.80-103.

Yu, B., Zhu, H., Cai, W., Ma, N., Kuang, Q. and Yao, B. 2013. Two-phase optimization approach to transit hub location – the case of Dalian. *Journal of Transport Geography.* **33**, pp.62-71.

Zhao, J., Frumin, M., Wilson, N. and Zhao, Z. 2013. Unified estimator for excess journey time under heterogeneous passenger incidence behavior using smartcard data. *Transportation Research Part C: Emerging Technologies.* **34**, pp.70-88.

Zhao, J., Rahbee, A. and Wilson, N.H.M. 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.

Zhong, C., Arisona, S.M., Huang, X., Batty, M. and Schmitt, G. 2014. Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science.* **28**(11), pp.2178-2199.

Zhu, W. and Xu, R. 2016. Generating route choice sets with operation information on metro networks. *Journal of Traffic and Transportation Engineering (English Edition).* **3**(3), pp.243-252.

Zhu, Y. 2017. *Passenger-to-Itinerary Assignment Model Based on Automated Data.* thesis, Northeastern University.

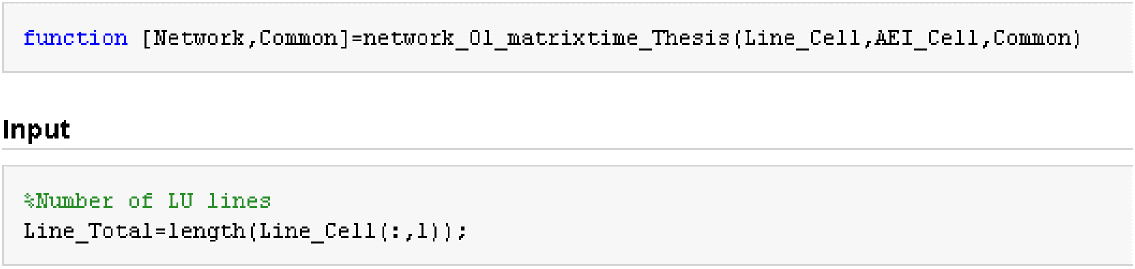
Zhu, Y., Koutsopoulos, H.N. and Wilson, N.H.M. 2017. A probabilistic Passenger-to-Train Assignment Model based on automated data. *Transportation Research Part B: Methodological.* **104**, pp.522-542.

Zhu, Y., Koutsopoulos, H.N. and Wilson, N.H.M. 2018. Inferring left behind passengers in congested metro systems from automated data. *Transportation Research Part C: Emerging Technologies.* **94**, pp.323-337.

Appendix A
Matlab code for creating matrix of link times

Section 4.4.1 presented the definition of nodes and links, as well as the allocation rules for the values in the matrix of link times. The case study network (i.e. LU inner zone with 9 lines and 68 stations) is represented with 280 nodes and 722 links.

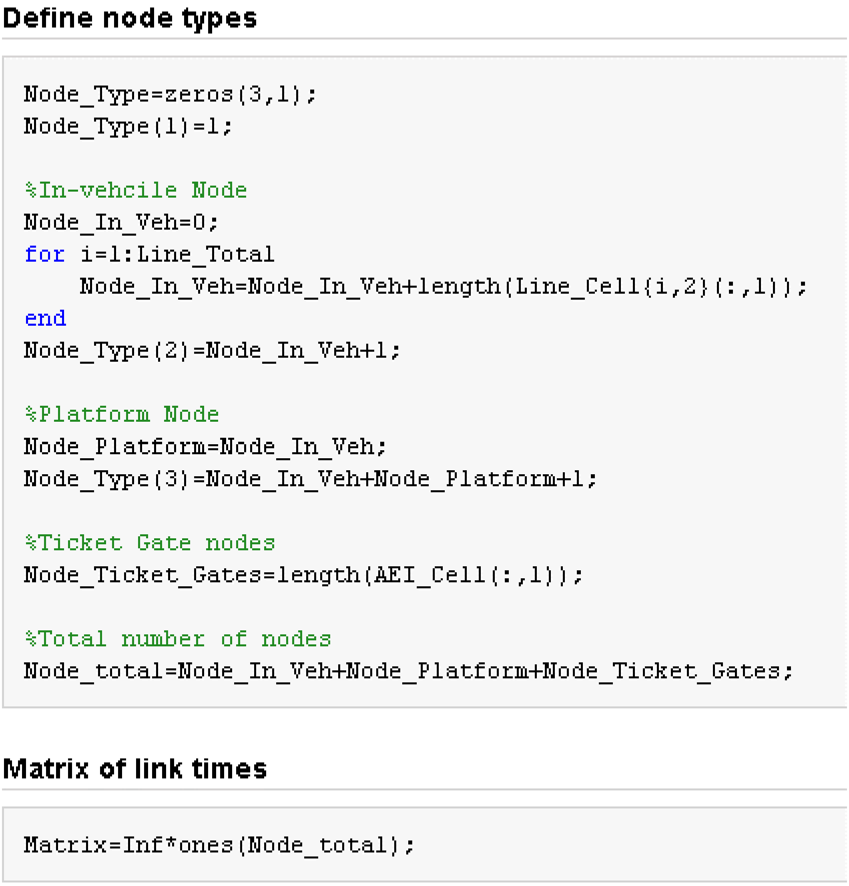
To fill the matrix of link times manually would be time consuming and it could easily give place to errors. To avoid this, a Matlab code was written to create and fill this matrix automatically based on the input data for the lines and stations (cf. **Section 3.6.2**):



```
function [Network,Common]=network_01_matrixtime_Thesis(Line_Cell,AEI_Cell,Common)
```

Input

```
%Number of LU lines
Line_Total=length(Line_Cell(:,1));
```

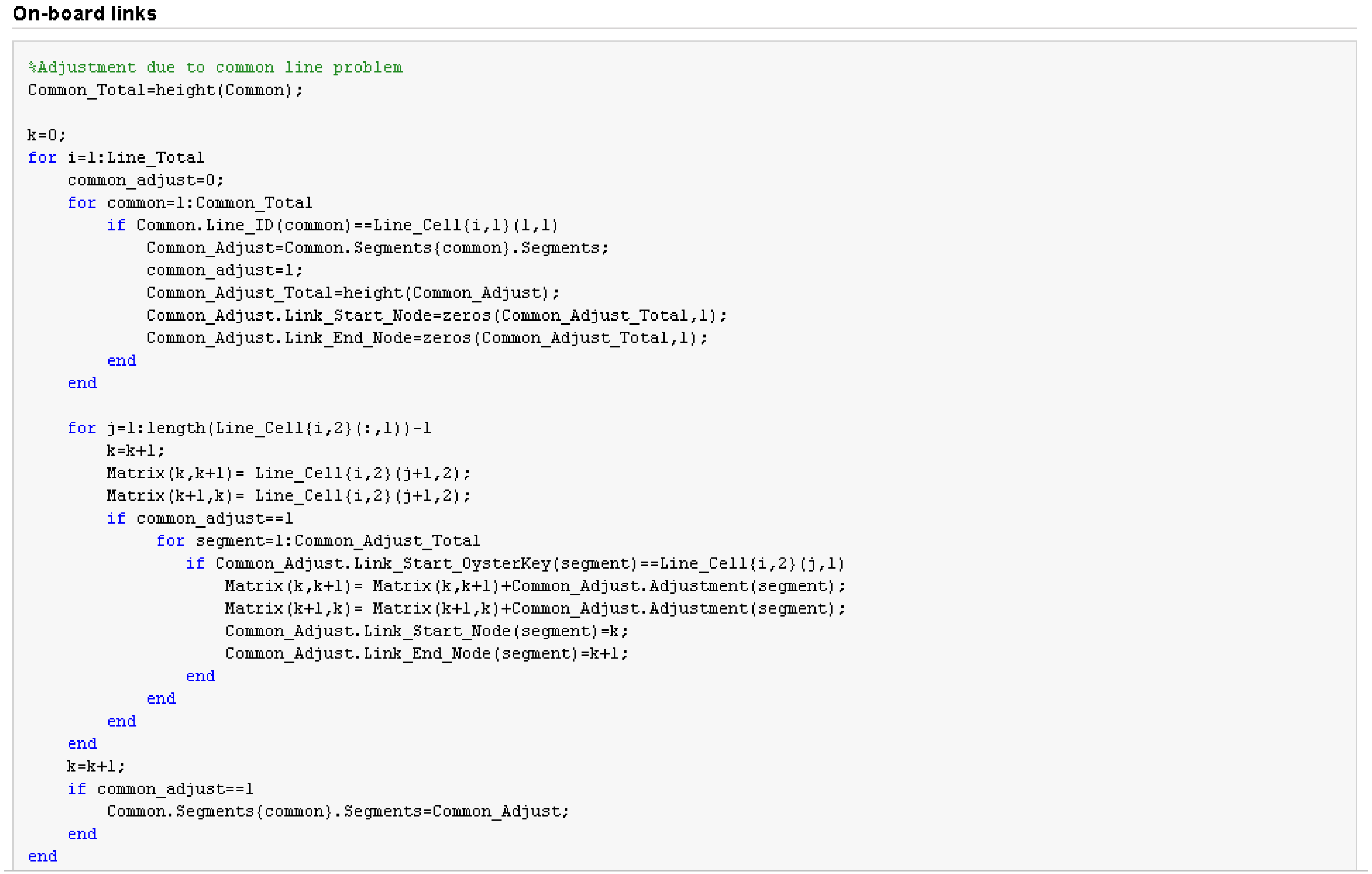


Define node types

```
Node_Type=zeros(3,1);
Node_Type(1)=1;

%In-vehcile Node
Node_In_Veh=0;
for i=1:Line_Total
    Node_In_Veh=Node_In_Veh+length(Line_Cell{i,2}(:,1));
end
Node_Type(2)=Node_In_Veh+1;

%Platform Node
Node_Platform=Node_In_Veh;
Node_Type(3)=Node_In_Veh+Node_Platform+1;

%Ticket Gate nodes
Node_Ticket_Gates=length(AEI_Cell(:,1));

%Total number of nodes
Node_total=Node_In_Veh+Node_Platform+Node_Ticket_Gates;
```

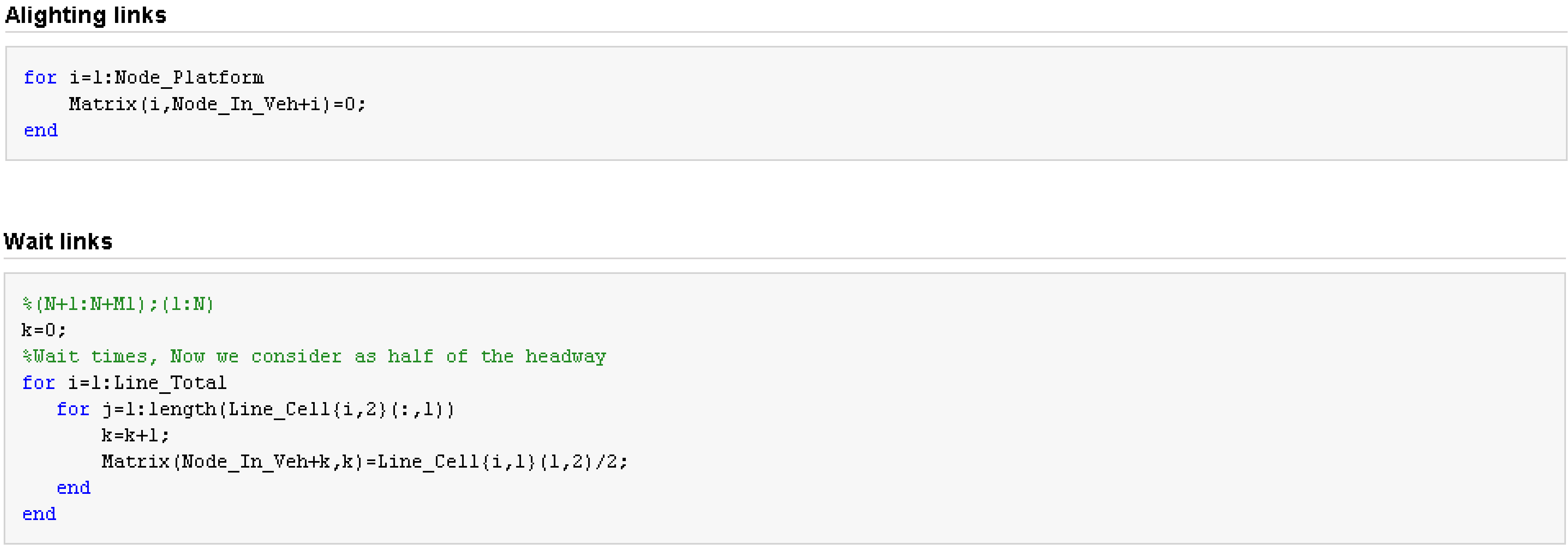
Matrix of link times

```
Matrix=Inf*ones(Node_total);
```



On-board links

```
%Adjustment due to common line problem
Common_Total=height(Common);

k=0;
for i=1:Line_Total
    common_adjust=0;
    for common=1:Common_Total
        if Common.Line_ID(common)==Line_Cell{i,1}(1,1)
            Common_Adjust=Common.Segments{common}.Segments;
            common_adjust=1;
            Common_Adjust_Total=height(Common_Adjust);
            Common_Adjust.Link_Start_Node=zeros(Common_Adjust_Total,1);
            Common_Adjust.Link_End_Node=zeros(Common_Adjust_Total,1);
        end
    end

    for j=1:length(Line_Cell{i,2}(:,1))-1
        k=k+1;
        Matrix(k,k+1)= Line_Cell{i,2}(j+1,2);
        Matrix(k+1,k)= Line_Cell{i,2}(j+1,2);
        if common_adjust==1
             for segment=1:Common_Adjust_Total
                if Common_Adjust.Link_Start_OysterKey(segment)==Line_Cell{i,2}(j,1)
                    Matrix(k,k+1)= Matrix(k,k+1)+Common_Adjust.Adjustment(segment);
                    Matrix(k+1,k)= Matrix(k+1,k)+Common_Adjust.Adjustment(segment);
                    Common_Adjust.Link_Start_Node(segment)=k;
                    Common_Adjust.Link_End_Node(segment)=k+1;
                end
            end
        end
    end
    k=k+1;
    if common_adjust==1
        Common.Segments{common}.Segments=Common_Adjust;
    end
end
```

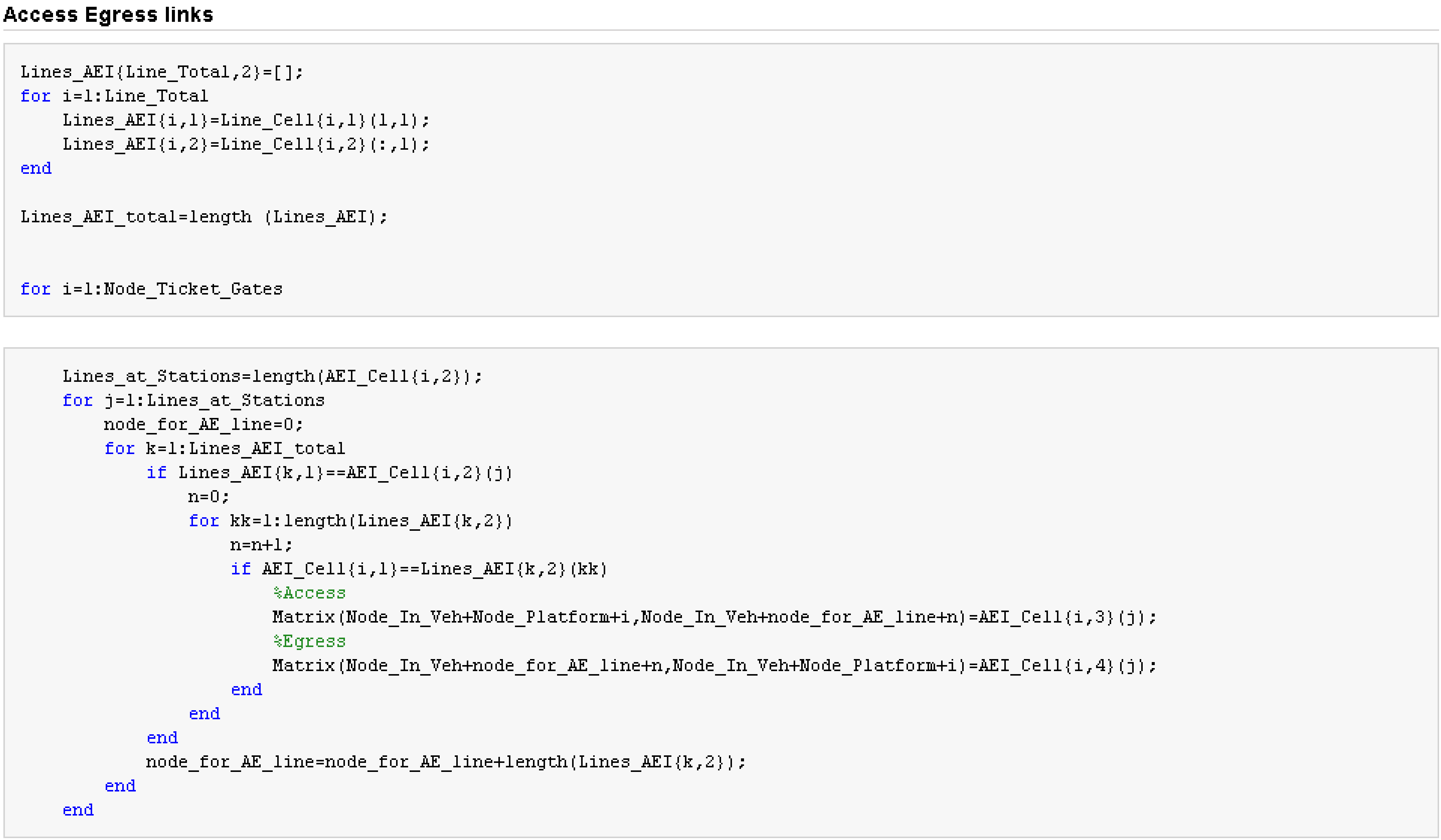


Alighting links

```
for i=1:Node_Platform
    Matrix(i,Node_In_Veh+i)=0;
end
```

Wait links

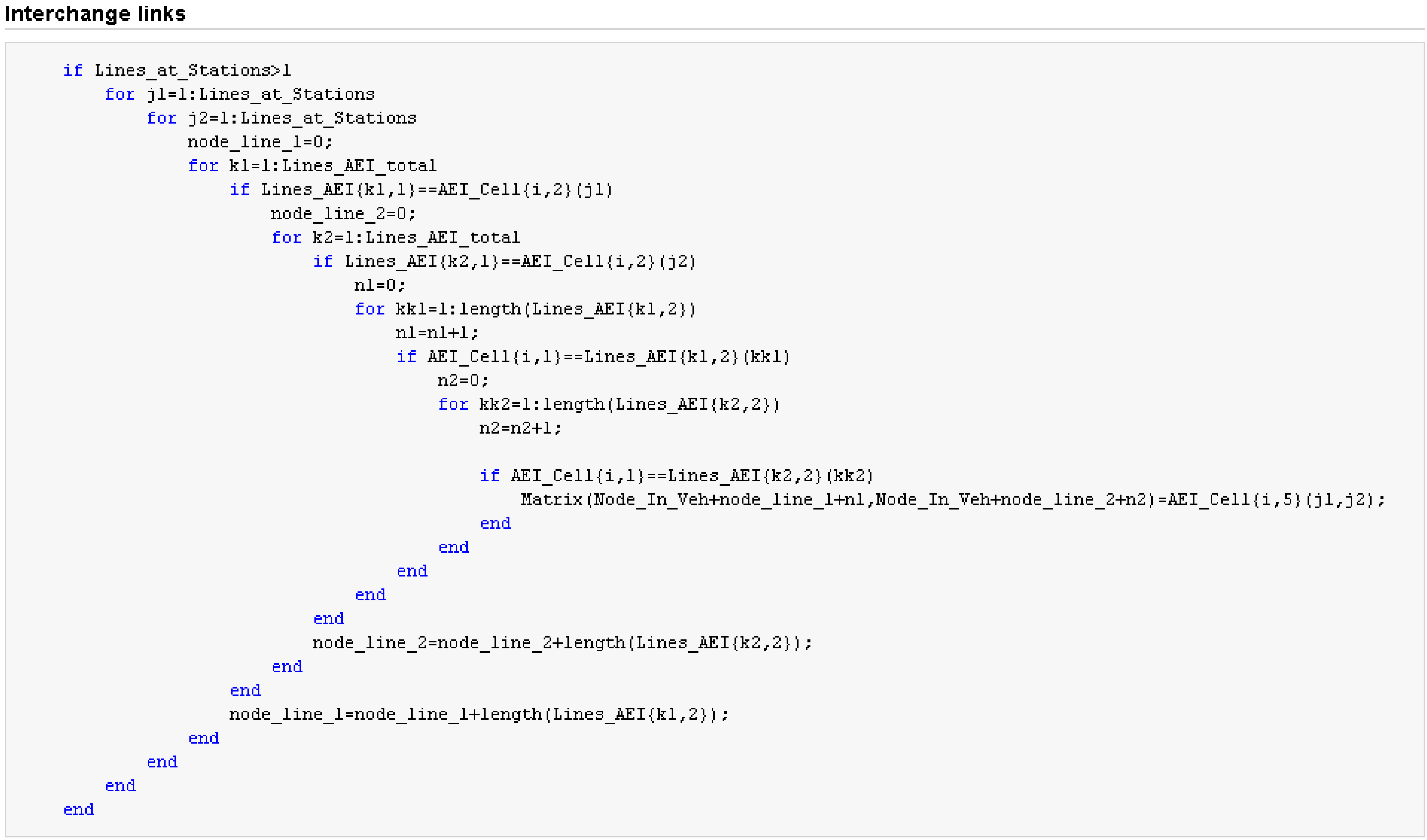
```
%(N+1:N+M1);(1:N)
k=0;
%Wait times, Now we consider as half of the headway
for i=1:Line_Total
   for j=1:length(Line_Cell{i,2}(:,1))
       k=k+1;
       Matrix(Node_In_Veh+k,k)=Line_Cell{i,1}(1,2)/2;
   end
end
```



Access Egress links

```
Lines_AEI{Line_Total,2}=[];
for i=1:Line_Total
    Lines_AEI{i,1}=Line_Cell{i,1}(1,1);
    Lines_AEI{i,2}=Line_Cell{i,2}(:,1);
end

Lines_AEI_total=length (Lines_AEI);


for i=1:Node_Ticket_Gates
```

```
    Lines_at_Stations=length(AEI_Cell{i,2});
    for j=1:Lines_at_Stations
        node_for_AE_line=0;
        for k=1:Lines_AEI_total
            if Lines_AEI{k,1}==AEI_Cell{i,2}(j)
                n=0;
                for kk=1:length(Lines_AEI{k,2})
                    n=n+1;
                    if AEI_Cell{i,1}==Lines_AEI{k,2}(kk)
                        %Access
                        Matrix(Node_In_Veh+Node_Platform+i,Node_In_Veh+node_for_AE_line+n)=AEI_Cell{i,3}(j);
                        %Egress
                        Matrix(Node_In_Veh+node_for_AE_line+n,Node_In_Veh+Node_Platform+i)=AEI_Cell{i,4}(j);
                    end
                end
            end
            node_for_AE_line=node_for_AE_line+length(Lines_AEI{k,2});
        end
    end
```



Interchange links

```
if Lines_at_Stations>1
    for j1=1:Lines_at_Stations
        for j2=1:Lines_at_Stations
            node_line_1=0;
            for k1=1:Lines_AEI_total
                if Lines_AEI{k1,1}==AEI_Cell{i,2}(j1)
                    node_line_2=0;
                    for k2=1:Lines_AEI_total
                        if Lines_AEI{k2,1}==AEI_Cell{i,2}(j2)
                            n1=0;
                            for kk1=1:length(Lines_AEI{k1,2})
                                n1=n1+1;
                                if AEI_Cell{i,1}==Lines_AEI{k1,2}(kk1)
                                    n2=0;
                                    for kk2=1:length(Lines_AEI{k2,2})
                                        n2=n2+1;

                                        if AEI_Cell{i,1}==Lines_AEI{k2,2}(kk2)
                                            Matrix(Node_In_Veh+node_line_1+n1,Node_In_Veh+node_line_2+n2)=AEI_Cell{i,5}(j1,j2);
                                        end
                                    end
                                end
                            end
                        end
                        node_line_2=node_line_2+length(Lines_AEI{k2,2});
                    end
                end
                node_line_1=node_line_1+length(Lines_AEI{k1,2});
            end
        end
    end
end
```

```
end
```

Output

```
Network{1,1}=Node_Type;
%Matrix of link times
Network{1,3}=Matrix;
```

Appendix B
The Dijkstra algorithm

The Dijkstra (1959) algorithm, calculates the shortest path by going through the following steps:

1. Set an initial value of **"distance from origin"** for all nodes:
   0 for initial node,
   ∞ for all other nodes

2. Set the origin node as current node
   Set all other nodes **unvisited**

3. Update **"distance from origin"** for the neighbours of the current node

   3.1. Calculate the **"distance from origin"** via the current node

   3.2. If the newly calculated **"distance from origin"** is smaller than the current value,
   assign the newly calculated value for that node
   set the current node as its **parent** node.

   3.3. Once 3.1 and 3.2 is done for all neighbours of the current node,
   mark the current node as **visited.**

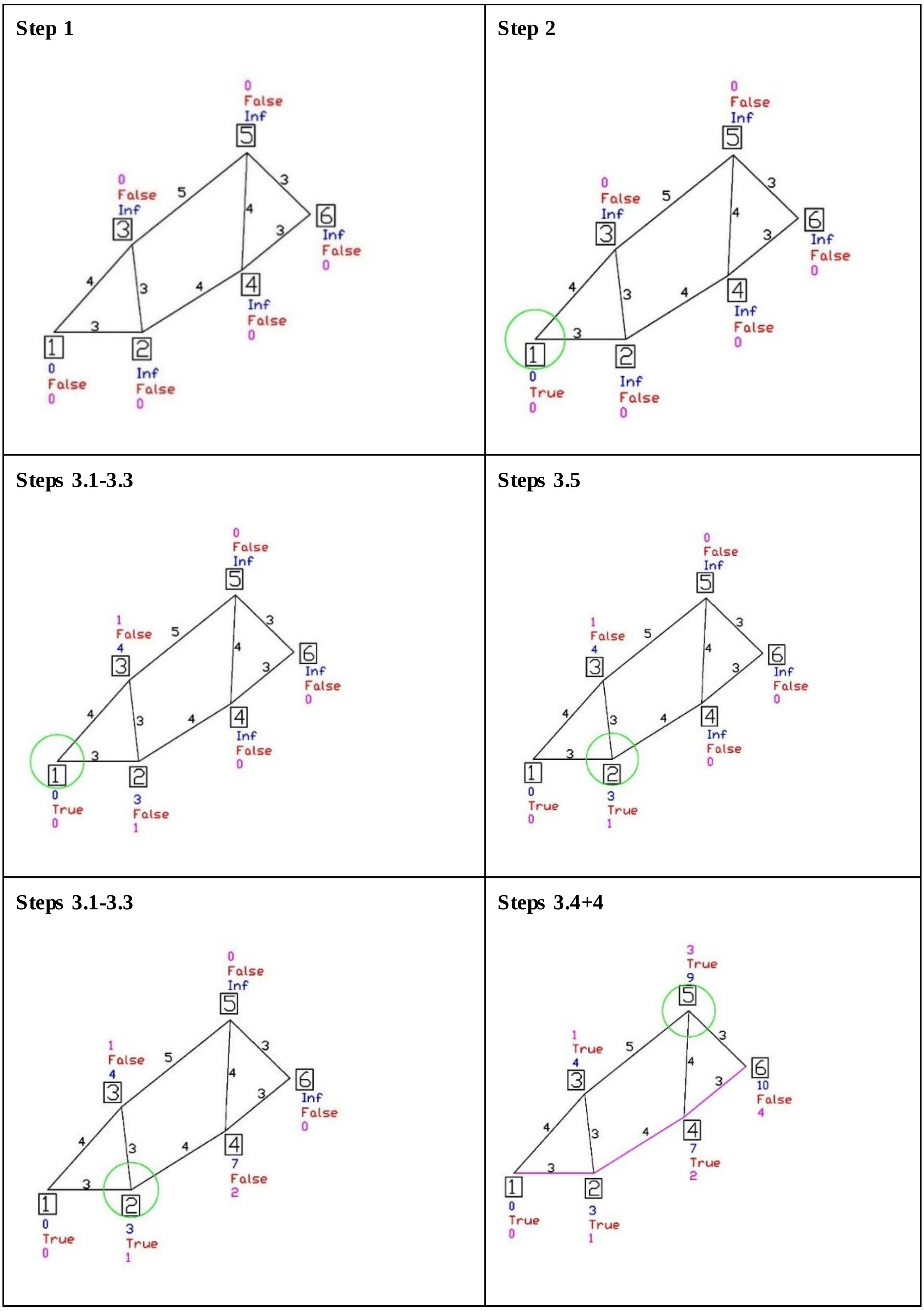
   3.4. If the destination node is marked as **visited**, stop

   3.5. Find the **unvisited** node with the smallest **"distance from origin"**
   Set it at current node
   Go back to step 3.1

4. The shortest path will be given as the sequence of **parent** nodes from destination to origin

This is illustrated on **Figure B-1** through a small example network. The Dijkstra algorithm was applied in Matlab as a sub-function of the K shortest algorithm on the LU inner zone network (cf. **Section 4.6**). The program code is available from the Matlab file exchange website[30].

[30] https://uk.mathworks.com/matlabcentral/fileexchange/5550-dijkstra-shortest-path-routing



Step 1

Step 2

Steps 3.1-3.3

Steps 3.5

Steps 3.1-3.3

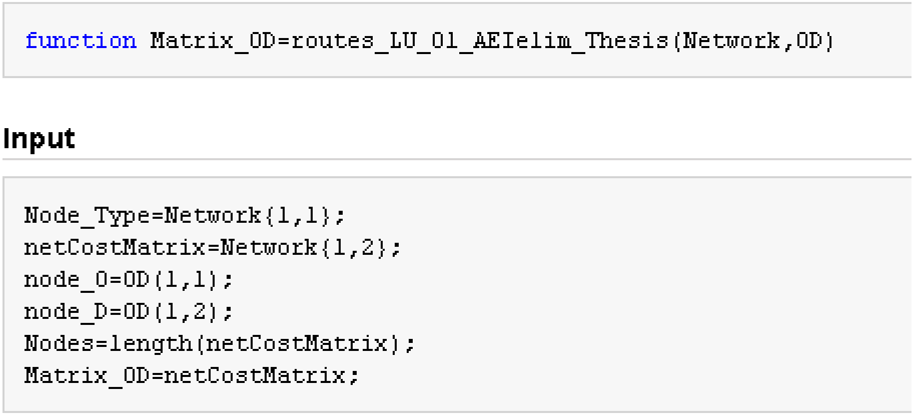
Steps 3.4+4

Figure B-1 The Dijkstra (1959) algorithm

Appendix C
Matlab code for the proposed modifications to the K shortest path algorithm

Section 4.6.2 explained the proposed modifications to the K shortest path algorithm to avoid that the results should give route variants, which differ only in their access egress interchange (AEI) movements within the stations (see **Figure 4-8** a). The Matlab code for these modifications are presented as follows

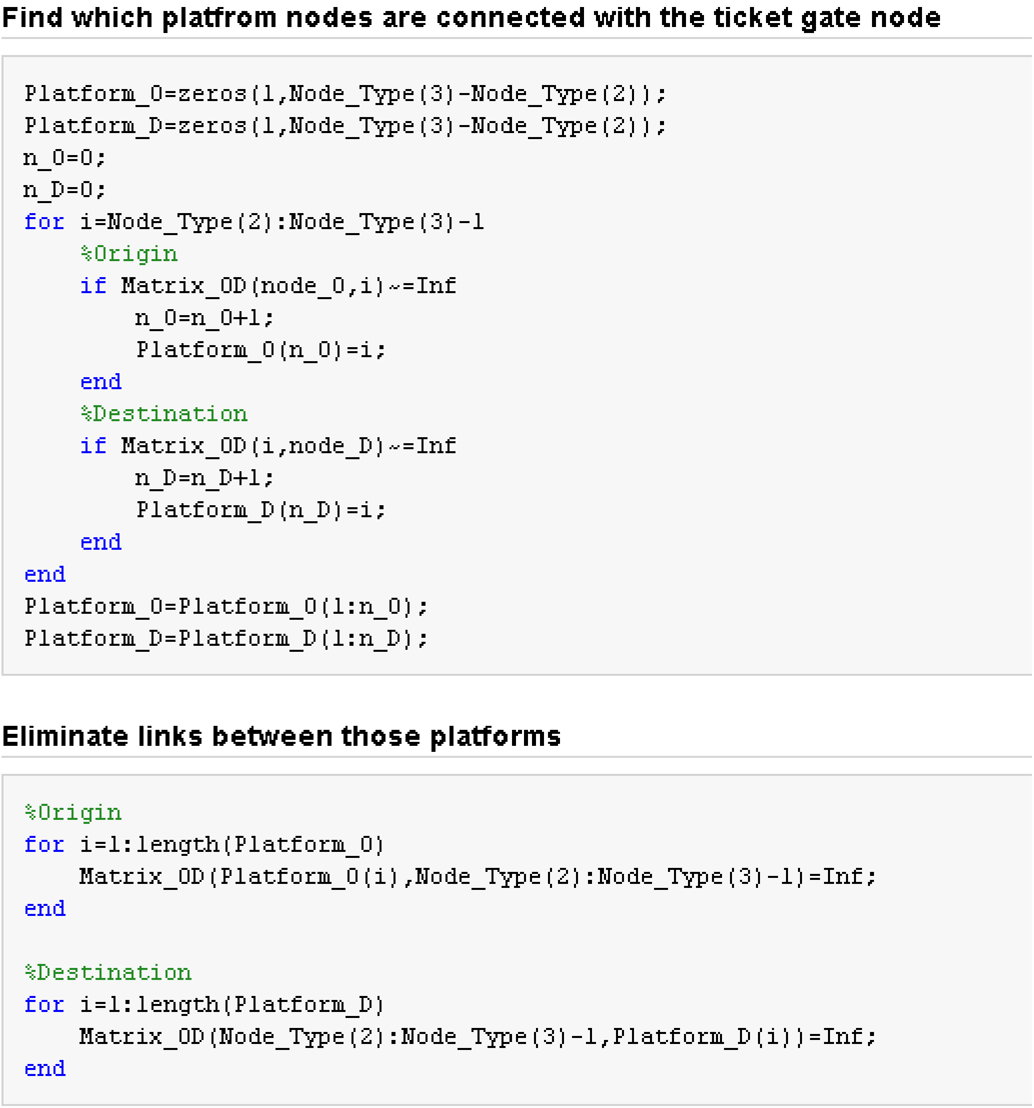
C.1 Algorithm to eliminate interchange links at origin and destination stations and access and egress links at all other stations



```
function Matrix_OD=routes_LU_01_AEIelim_Thesis(Network,OD)
```

Input

```
Node_Type=Network{1,1};
netCostMatrix=Network{1,2};
node_O=OD(1,1);
node_D=OD(1,2);
Nodes=length(netCostMatrix);
Matrix_OD=netCostMatrix;
```

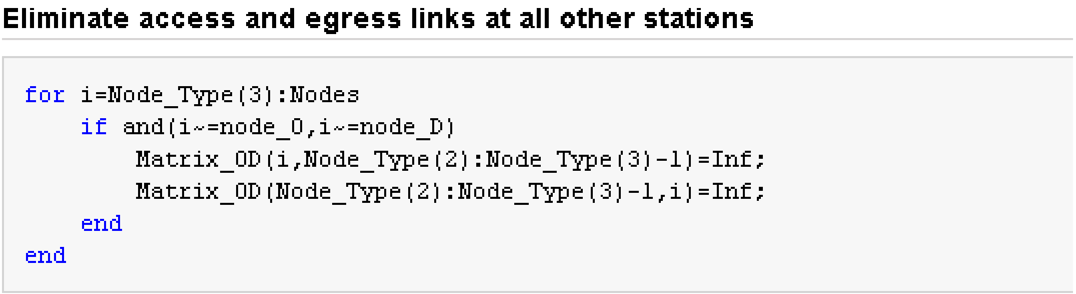


Find which platfrom nodes are connected with the ticket gate node

```
Platform_O=zeros(1,Node_Type(3)-Node_Type(2));
Platform_D=zeros(1,Node_Type(3)-Node_Type(2));
n_O=0;
n_D=0;
for i=Node_Type(2):Node_Type(3)-1
    %Origin
    if Matrix_OD(node_O,i)~=Inf
        n_O=n_O+1;
        Platform_O(n_O)=i;
    end
    %Destination
    if Matrix_OD(i,node_D)~=Inf
        n_D=n_D+1;
        Platform_D(n_D)=i;
    end
end
Platform_O=Platform_O(1:n_O);
Platform_D=Platform_D(1:n_D);
```

Eliminate links between those platforms

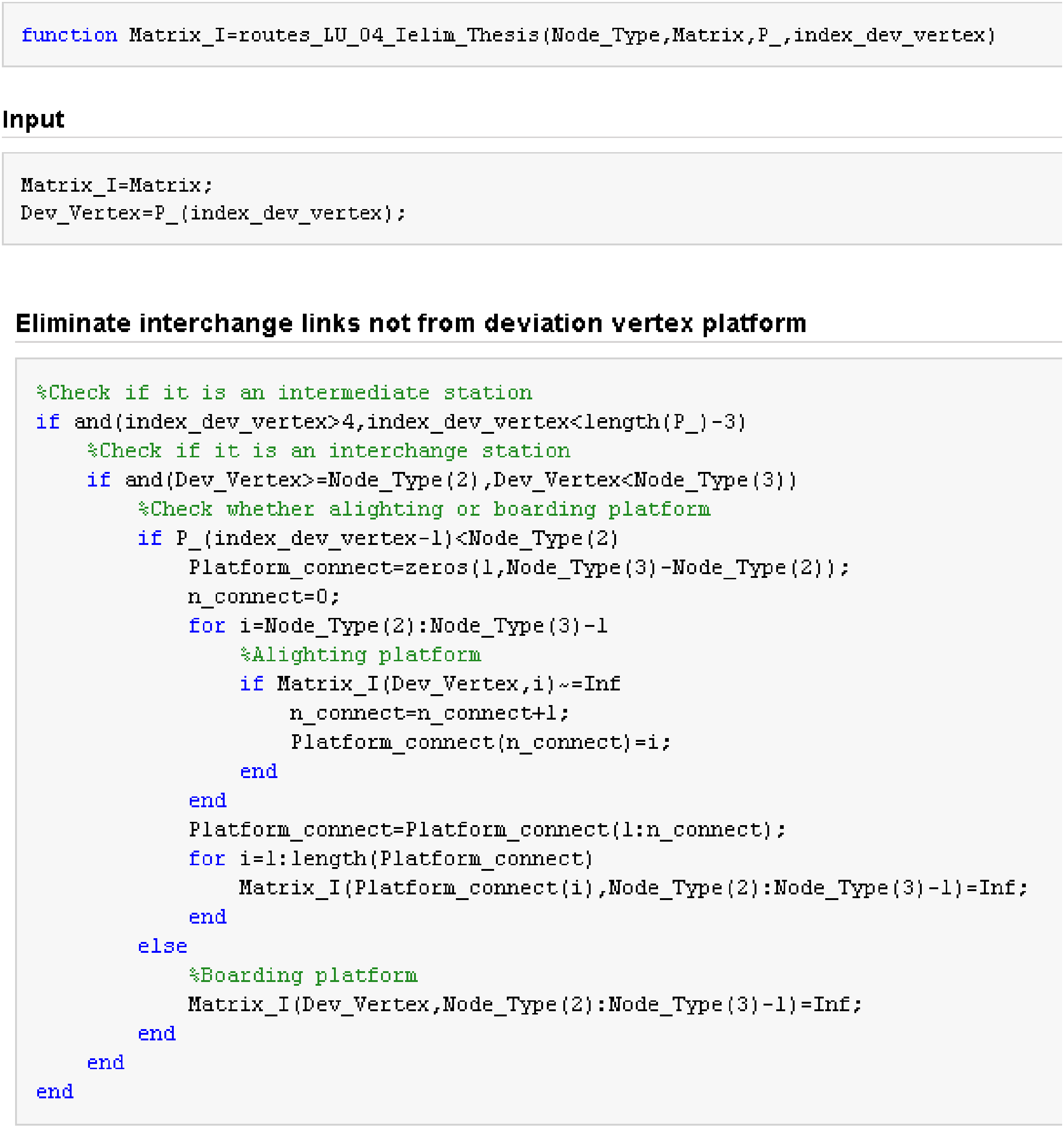
```
%Origin
for i=1:length(Platform_O)
    Matrix_OD(Platform_O(i),Node_Type(2):Node_Type(3)-1)=Inf;
end

%Destination
for i=1:length(Platform_D)
    Matrix_OD(Node_Type(2):Node_Type(3)-1,Platform_D(i))=Inf;
end
```



Eliminate access and egress links at all other stations

```
for i=Node_Type(3):Nodes
    if and(i~=node_O,i~=node_D)
        Matrix_OD(i,Node_Type(2):Node_Type(3)-1)=Inf;
        Matrix_OD(Node_Type(2):Node_Type(3)-1,i)=Inf;
    end
end
```

C.2 Algorithm to eliminate links at interchange stations which does not start from the deviation vertex



```
function Matrix_I=routes_LU_04_Ielim_Thesis(Node_Type,Matrix,P_,index_dev_vertex)
```

Input

```
Matrix_I=Matrix;
Dev_Vertex=P_(index_dev_vertex);
```

Eliminate interchange links not from deviation vertex platform

```
%Check if it is an intermediate station
if and(index_dev_vertex>4,index_dev_vertex<length(P_)-3)
    %Check if it is an interchange station
    if and(Dev_Vertex>=Node_Type(2),Dev_Vertex<Node_Type(3))
        %Check whether alighting or boarding platform
        if P_(index_dev_vertex-1)<Node_Type(2)
            Platform_connect=zeros(1,Node_Type(3)-Node_Type(2));
            n_connect=0;
            for i=Node_Type(2):Node_Type(3)-1
                %Alighting platform
                if Matrix_I(Dev_Vertex,i)~=Inf
                    n_connect=n_connect+1;
                    Platform_connect(n_connect)=i;
                end
            end
            Platform_connect=Platform_connect(1:n_connect);
            for i=1:length(Platform_connect)
                Matrix_I(Platform_connect(i),Node_Type(2):Node_Type(3)-1)=Inf;
            end
        else
            %Boarding platform
            Matrix_I(Dev_Vertex,Node_Type(2):Node_Type(3)-1)=Inf;
        end
    end
end
```

Appendix D
The results of the K shortest algorithm for the case study OD pairs

In **Section 4.6.3**, the results were presented for the **Victoria** – **Holborn** OD pair. Here the results are presented for the other case study OD pairs

Table D-1 The 10 shortest routes for **Euston – St James's Park** (OD 2) with their journey time and generalised cost, observed routes (Rolling Origin Destination Survey, RODS) are highlighted with green

Route								Time		Generalised cost		
ID	Line 1	IC 1	Line 2	IC 2	Line 3	IC 3	Line 4	Total	IC	Total	AEI	Proportion
k, ij $ij = 2$								$t_{k,ij}$ [min]	$t^{ic}_{k,ij}$ [min]	$c_{k,ij}$ [min]	$c^{AEI}_{k,ij}$ [min]	$\delta c_{k,ij}$
1	Victoria	Victoria	Circle					16.7	2.0	27.7	14.9	1.00
2	Northern (CX)	Embankment	Circle					20.0	1.3	30.9	12.1	1.11
3	Victoria	Green Park	Jubilee	Westminster	Circle			22.1	5.5	38.9	23.1	1.40
4	Victoria	Oxford Circus	Bakerloo	Embankment	Circle			23.7	3.6	39.5	19.8	1.42
5	Northern (CX)	Warren Street	Victoria	Victoria	Circle			20.9	5.2	40.7	24.0	1.47
6	Northern (Bank)	Kings Cross	Victoria	Victoria	Circle			23.9	5.2	41.4	22.7	1.49
7	Northern (Bank)	Bank	Circle					32.2	3.5	43.4	15.5	1.56
8	Victoria	Warren Street	Northern (CX)	Embankment	Circle			24.6	4.5	43.7	24.0	1.58
9	Victoria	Oxford Circus	Bakerloo	Baker Street	Jubilee	Westminster	Circle	28.9	4.7	46.7	22.0	1.68
10	Northern (CX)	Waterloo	Bakerloo	Embankment	Circle			26.3	4.1	47.2	23.5	1.70

Table D-2 The 10 shortest routes for **Victoria** – **Liverpool Street** (OD 3) with their journey time and generalised cost, observed routes (Rolling Origin Destination Survey, RODS) are highlighted with green

Route						Time		Generalised cost		
ID	**Line 1**	**IC 1**	**Line 2**	**IC 2**	**Line 3**	**Total**	**IC**	**Total**	**AEI**	**Proportion**
k, ij $ij = 3$						$t_{k,ij}$ [min]	$t^{ic}_{k,ij}$ [min]	$c_{k,ij}$ [min]	$c^{AEI}_{k,ij}$ [min]	$\delta c_{k,ij}$
1	Victoria	Oxford Circus	Central			23.2	3.4	34.1	17.2	1.00
2	Circle					28.5	0.0	34.2	4.6	1.00
3	Victoria	Kings Cross	Circle			28.0	5.3	38.9	18.0	1.14
4	Circle	Bank	Central			28.5	5.8	39.8	18.9	1.17
5	Victoria	Green Park	Jubilee	Bond Street	Central	26.9	5.0	44.4	24.6	1.30
6	Victoria	Green Park	Piccadilly	Holborn	Central	29.5	7.1	47.2	26.4	1.39
7	Victoria	Kings Cross	Northern (Bank)	Moorgate	Circle	29.0	5.3	47.7	24.9	1.40
8	Victoria	Oxford Circus	Bakerloo	Baker Street	Circle	32.7	3.5	48.0	19.2	1.41
9	Victoria	Euston	Northern (Bank)	Moorgate	Circle	29.3	5.6	48.8	26.1	1.43
10	Circle	Embankment	Northern (CX)	Tottenham Court Rd	Central	30.4	4.2	49.4	23.7	1.45

Table D-3 The 10 shortest routes for **Angel** – **Waterloo** (OD 4)with their journey time and generalised cost, observed routes (Rolling Origin Destination Survey, RODS) are highlighted with green

Route						Time		Generalised cost		
ID	**Line 1**	**IC 1**	**Line 2**	**IC 2**	**Line 3**	**Total**	**IC**	**Total**	**AEI**	**Proportion**
k, ij $ij = 4$						$t_{k,ij}$ [min]	$t^{ic}_{k,ij}$ [min]	$c_{k,ij}$ [min]	$c^{AEI}_{k,ij}$ [min]	$\delta c_{k,ij}$
1	**Northern (Bank)**	**London Bridge**	**Jubilee**			23.9	3.2	33.9	19.1	1.00
2	**Northern (Bank)**	**Bank**	**Waterloo & City**			23.5	5.3	35.6	19.8	1.05
3	**Northern (Bank)**	**Euston**	**Northern (CX)**			25.3	2.6	37.1	18.3	1.09
4	**Northern (Bank)**	**Kings Cross**	**Victoria**	**Oxford Circus**	**Bakerloo**	26.9	4.4	40.5	20.7	1.19
5	**Northern (Bank)**	**Euston**	**Victoria**	**Oxford Circus**	**Bakerloo**	27.1	4.6	43.3	23.5	1.28
6	**Northern (Bank)**	**Kings Cross**	**Victoria**	**Green Park**	**Jubilee**	29.5	5.7	46.1	27.3	1.36
7	**Northern (Bank)**	**Kings Cross**	**Victoria**	**Warren Street**	**Northern (CX)**	29.0	6.3	46.9	27.1	1.38
8	**Northern (Bank)**	**Kings Cross**	**Victoria**	**Euston**	**Northern (CX)**	30.2	6.5	48.1	27.4	1.42
9	**Northern (Bank)**	**Euston**	**Northern (CX)**	**Embankment**	**Bakerloo**	28.6	4.6	48.8	27.1	1.44
10	**Northern (Bank)**	**Euston**	**Victoria**	**Green Park**	**Jubilee**	29.8	6.0	48.9	30.1	1.44

Table D-4 The 10 shortest routes for **Liverpool Street** – **Green Park** (OD 5) with their journey time and generalised cost, observed routes (Rolling Origin Destination Survey, RODS) are highlighted with green

Route						Time		Generalised cost		
ID	Line 1	IC 1	Line 2	IC 2	Line 3	Total	IC	Total	AEI	Proportion
k, ij $ij = 5$						$t_{k,ij}$ [min]	$t^{ic}_{k,ij}$ [min]	$c_{k,ij}$ [min]	$c^{AEI}_{k,ij}$ [min]	$\delta c_{k,ij}$
1	Central	Oxford Circus	Victoria			21.4	2.9	32.8	16.9	1.00
2	Central	Bond Street	Jubilee			24.3	3.2	34.9	18.0	1.06
3	Central	Holborn	Piccadilly			24.1	3.4	35.8	18.0	1.09
4	Circle	Kings Cross	Victoria			26.3	5.3	36.1	16.2	1.10
5	Circle	Baker Street	Jubilee			29.3	3.8	40.3	17.4	1.23
6	Circle	Kings Cross	Piccadilly			30.2	5.0	40.6	16.8	1.24
7	Central	Bank	Northern (Bank)	London Bridge	Jubilee	25.7	6.6	43.7	27.9	1.33
8	Circle	Moorgate	Northern (Bank)	Kings Cross	Victoria	27.3	5.3	44.0	22.3	1.34
9	Circle	Moorgate	Northern (Bank)	London Bridge	Jubilee	26.8	5.3	44.0	24.3	1.34
10	Circle	Westminster	Jubilee			31.9	2.9	44.7	15.2	1.37

Table D-5 The 10 shortest routes for **Euston** – **South Kensington** (OD 6) with their journey time and generalised cost, observed routes (Rolling Origin Destination Survey, RODS) are highlighted with green

Route						Time		Generalised cost		
ID	Line 1	IC 1	Line 2	IC 2	Line 3	Total	IC	Total	AEI	Proportion
k, ij $ij = 6$						$t_{k,ij}$ [min]	$t^{ic}_{k,ij}$ [min]	$c_{k,ij}$ [min]	$c^{AEI}_{k,ij}$ [min]	$\delta c_{k,ij}$
1	Victoria	Victoria	Circle			20.1	2.0	31.5	16.7	1.00
2	Victoria	Green Park	Piccadilly			25.7	3.7	36.2	19.4	1.15
3	Northern (CX)	Leicester Square	Piccadilly			26.5	1.5	38.0	16.3	1.21
4	Northern (CX)	Embankment	Circle			27.4	1.3	38.6	13.8	1.23
5	Victoria	Kings Cross	Piccadilly			32.4	3.4	42.8	19.0	1.36
6	Victoria	Oxford Circus	Bakerloo	Piccadilly Circus	Piccadilly	27.4	2.9	43.1	22.3	1.37
7	Northern (Bank)	Kings Cross	Piccadilly			32.6	4.7	43.2	19.3	1.37
8	Northern (CX)	Warren Street	Victoria	Victoria	Circle	24.3	5.2	44.5	25.7	1.41
9	Northern (Bank)	Kings Cross	Victoria	Victoria	Circle	27.3	5.2	45.2	24.4	1.43
10	Victoria	Green Park	Jubilee	Westminster	Circle	29.5	5.5	46.6	24.9	1.48

Table D-6 The 10 shortest routes for **Victoria** – **Waterloo** (OD 7) with their journey time and generalised cost, observed routes (Rolling Origin Destination Survey, RODS) are highlighted with green

Route						Time		Generalised cost		
ID	**Line 1**	**IC 1**	**Line 2**	**IC 2**	**Line 3**	**Total**	**IC**	**Total**	**AEI**	**Proportion**
k, ij $ij = 7$						$t_{k,ij}$ [min]	$t^{ic}_{k,ij}$ [min]	$c_{k,ij}$ [min]	$c^{AEI}_{k,ij}$ [min]	$\delta c_{k,ij}$
1	Victoria	Oxford Circus	Bakerloo			17.5	1.2	24.0	10.2	1.00
2	Victoria	Green Park	Jubilee			16.1	2.6	25.6	16.7	1.07
3	Circle	Embankment	Northern (CX)			15.6	1.3	26.0	13.3	1.08
4	Circle	Westminster	Jubilee			17.2	2.9	27.0	16.2	1.13
5	Circle	Embankment	Bakerloo			16.4	2.4	27.7	14.9	1.15
6	Victoria	Warren Street	Northern (CX)			23.7	3.2	34.4	16.6	1.43
7	Victoria	Oxford Circus	Bakerloo	Embankment	Northern (CX)	21.2	3.2	37.1	20.4	1.55
8	Victoria	Oxford Circus	Bakerloo	Baker Street	Jubilee	26.9	1.8	37.4	15.6	1.56
9	Victoria	Euston	Northern (CX)			26.9	3.3	37.7	16.8	1.57
10	Victoria	Green Park	Piccadilly	Piccadilly Circus	Bakerloo	20.3	5.5	39.3	25.5	1.63

Appendix E
Influence of the selection of the centroid on the CCOJT distribution

Following the analogy in **Section 5.4.2**, here it is discussed how the CCOJT distribution varies if a different origin or destination superstation centroid is chosen. Let them be called Ic' and Jc' respectively. This way, the on-board time between the previously and the newly chosen centroids is $t^{ob}_{(Ic)(Ic')}$ for the origin superstation and $t^{ob}_{(Jc)(Jc')}$ for the destination superstation (**Figure E-1**). Based on these considerations, the equivalent entry and exit time stamp at the superstation centroid can be written as:

$$T^{entry}_{(Ic)(Ic')} = T^{entry}_{(Ii)(Ic)} + t^{ob}_{(Ic)(Ic')} \tag{E-1}$$

and

$$T^{exit}_{(Jj)(Jc')} = T^{exit}_{(Jc)(Jc)} + t^{ob}_{(Jc)(Jc')} \tag{E-2}$$

respectively.

Let $CCOJT'_{(Ii)(Jj)}$ be the CCOJT corresponding to each station-to-station OD pair based on the newly selected centroids. Following the analogy in equation (5-6), this can be calculated as:

$$CCOJT'_{(Ii)(Jj)} = T^{exit}_{(Jj)(Jc')} - T^{entry}_{(Ii)(Ic')} \tag{E-3}$$

Substituting equation (E-1) and (E-2) into (E-3):

$$\begin{aligned} CCOJT'_{(Ii)(Jj)} &= T^{exit}_{(Jc)(Jc)} + t^{ob}_{(Jc)(Jc')} - \left(T^{entry}_{(Ii)(Ic)} + t^{ob}_{(Ic)(Ic')}\right) \\ &= T^{exit}_{(Jc)(Jc)} - T^{entry}_{(Ii)(Ic)} + t^{ob}_{(Jc)(Jc')} - t^{ob}_{(Ic)(Ic')} \end{aligned} \tag{E-4}$$

Substituting equation (5-6) into (E-4):

$$CCOJT'_{(Ii)(Jj)} = CCOJT_{(Ii)(Jj)} + t^{ob}_{(Jc)(Jc')} - t^{ob}_{(Ic)(Ic')} \tag{E-5}$$

Looking at equation (E-5) it can be understood that it contains the same term of $t^{ob}_{(Jc)(Jc')} - t^{ob}_{(Ic)(Ic')}$ for all possible entry (Ii) and exit (Jj) stations. In other words, the $CCOJT'_{(Ii)(Jj)}$ obtained with the newly selected superstation centroids is just the previously defined $CCOJT_{(Ii)(Jj)}$ shifted with the corresponding on-board times $(t^{ob}_{(Jc)(Jc')} - t^{ob}_{(Ic)(Ic')})$. Therefore the shape of $CCOJT'_{(Ii)(Jj)}$ distribution is not different from the shape of $CCOJT_{(Ii)(Jj)}$ and hence the route choice estimates with the finite mixture model would be also identical.

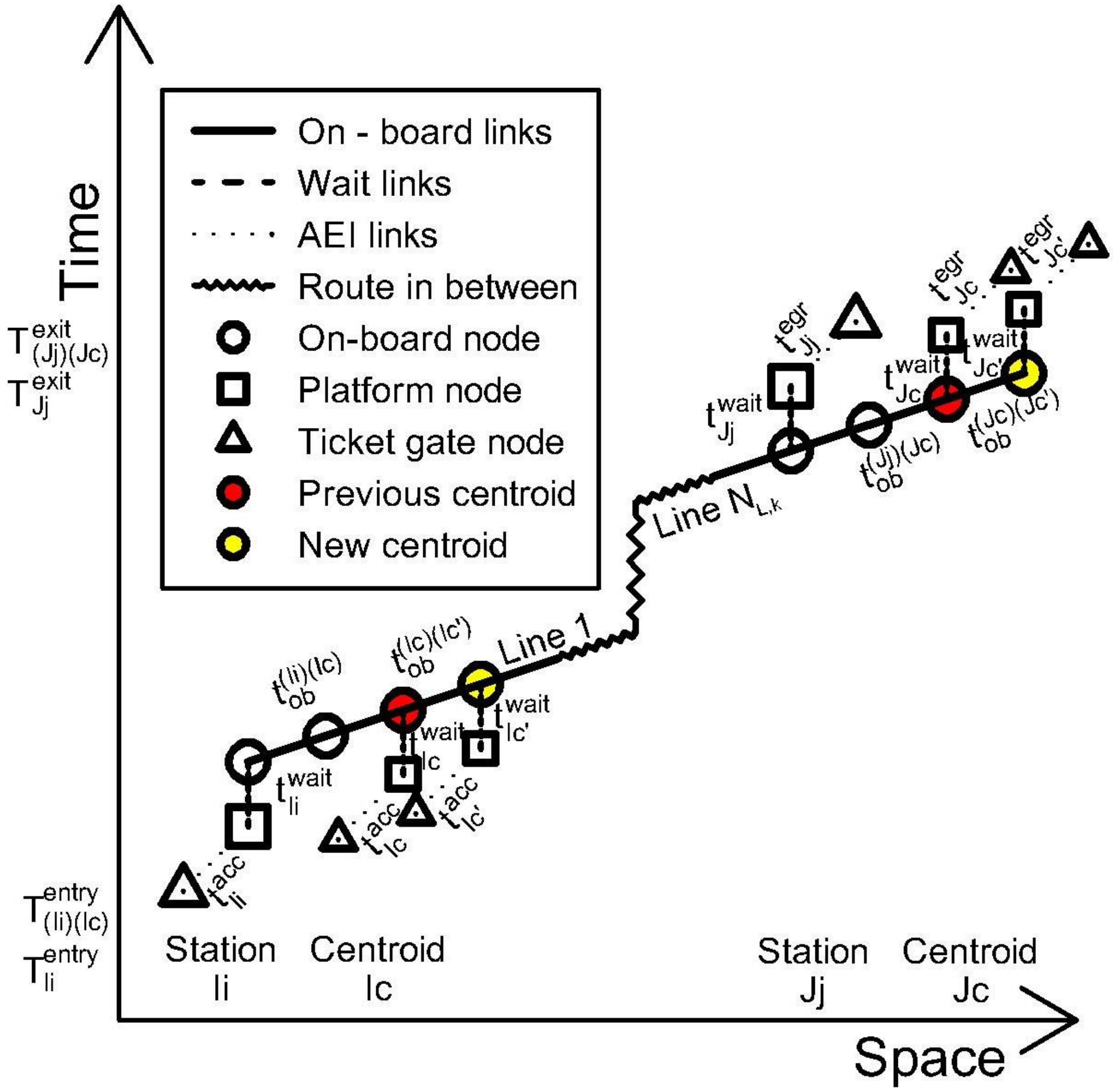


Figure E-1 Adjustment of the Oyster entry/exit times at the origin/destination superstations with a newly selected centroid, representation on a diachronic graph

Appendix F
Application of the finite mixture model on the CCOJTs of superstation-to-superstation OD pairs

The case studies for the application of the finite mixture model on the CCOJTs of superstation-to-superstation OD pairs were presented in **Section 5.6**. The detailed results with different seeds and tolerance thresholds are reported here.

Case 1 Victoria South - Holborn

The settings described in **Section 3.3.1** were used for centroid initialisation (K-means ++), distances (Euclidean square) and update methods (online phase). Conducting trials with different seeds for the random number generator it gave two possible solutions for $\mu_{r,IJ}^{KMS}$, $\sigma_{r,IJ}^{KMS}$ and $\omega_{r,IJ}^{KMS}$ (**Table F-1**)

Table F-1 Results of the K-means clustering algorithm with different seeds for **Victoria South – Holborn**; a) Seed=1, b) Seed=2, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	18.0	2.0	70.7%
2	24.0	2.5	29.3%

a)

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	16.0	1.9	57.4%
2	22.0	3.3	42.6%

b)

Figure F-1 and **Figure F-2** presents the estimated mean (μ_1^{MIX}) and proportion (ω_1^{MIX}) for mixture component labelled with $r = 1$.

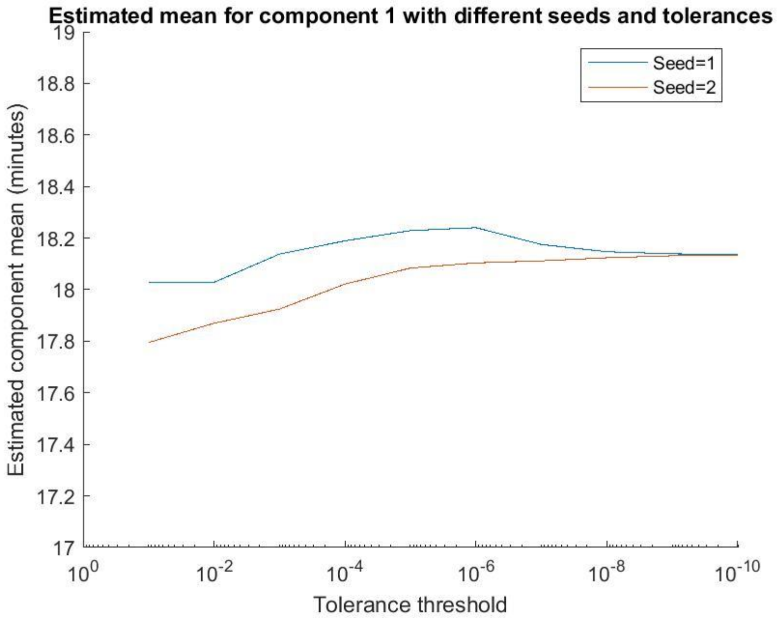


Figure F-1 Estimated mean for mixture component 1, given different initial values and tolerance thresholds for **Victoria South** – **Holborn**, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

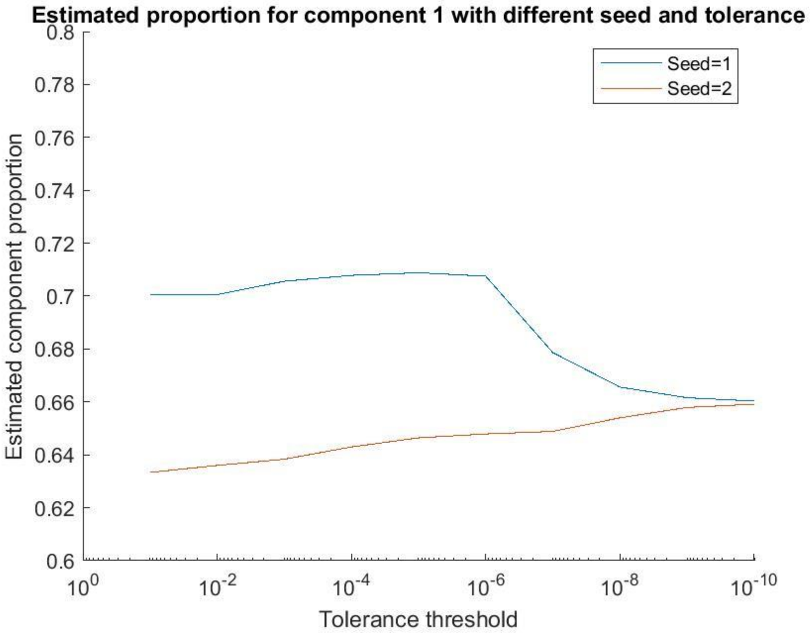


Figure F-2 Estimated proportion for mixture component 1, given different initial values and tolerance thresholds for **Victoria South** – **Holborn**, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Figure F-3 presents the log-likelihood (equation (3-9)) for each initial value (seed) and tolerance threshold.

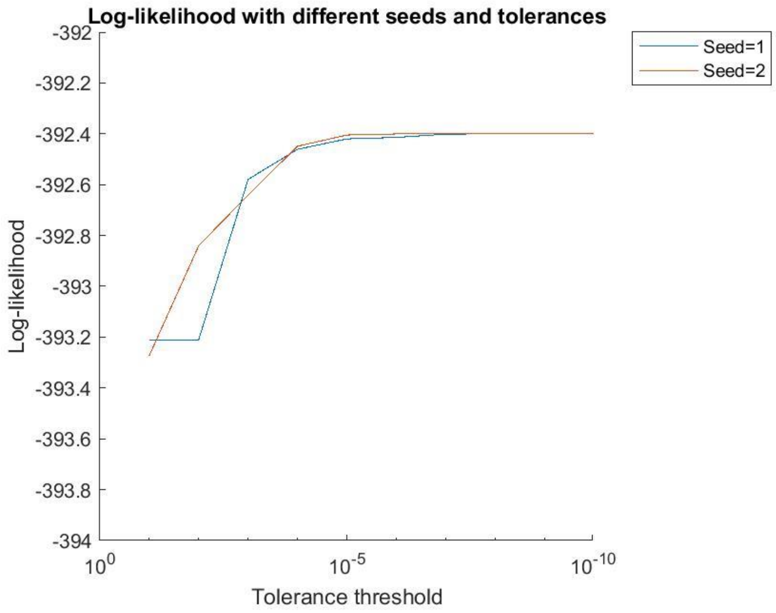


Figure F-3 Log-likelihood,

given different initial values and tolerance thresholds for **Victoria South** – **Holborn** OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Case 2 Central East – Green Park

The settings described in **Section 3.3.1** were used for centroid initialisation (K-means ++), distances (Euclidean square) and update methods (online phase). Conducting trials with different seeds for the random number generator it gave two possible solutions for $\mu_{r,IJ}^{KMS}$, $\sigma_{r,IJ}^{KMS}$ and $\omega_{r,IJ}^{KMS}$ (**Table F-2**)

Table F-2 Results of the K-means clustering algorithm

for **Central East– Green Park**;

OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	20.0	2.3	82.6%
2	29.5	4.0	17.4%

Figure F-4 and **Figure F-5** presents the estimated mean (μ_1^{MIX}) and proportion (ω_1^{MIX}) for mixture component labelled with $r = 1$.

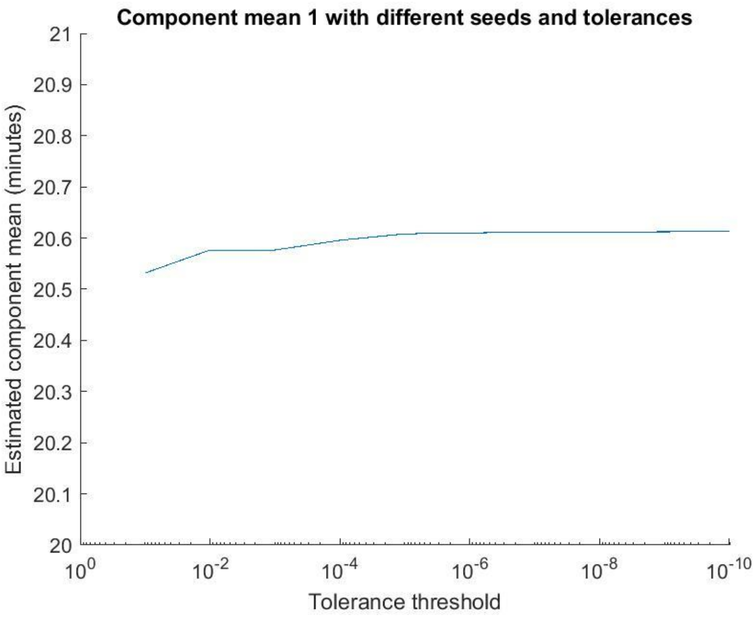


Figure F-4 Estimated mean for mixture component 1,

given different initial values and tolerance thresholds for **Central East– Green Park**,

OJTs adjusted to superstation centroid, but not according to fail-to-board delays

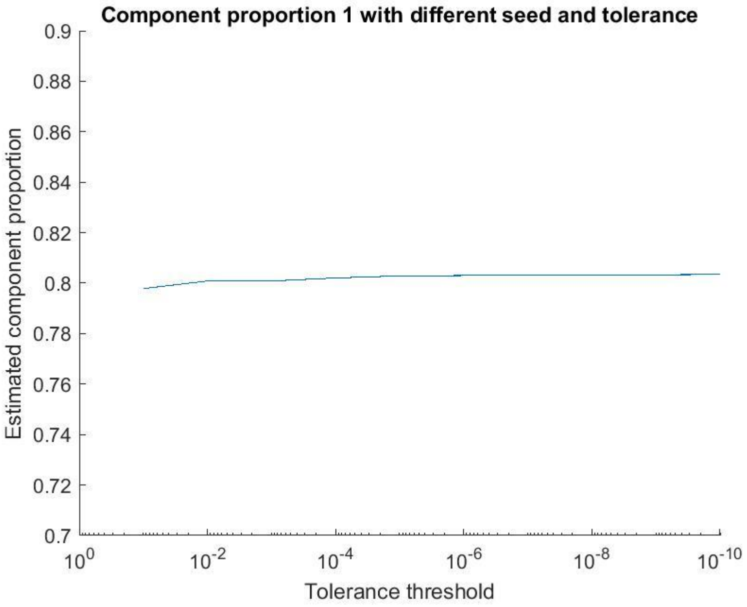


Figure F-5 Estimated proportion for mixture component 1,
given different initial values and tolerance thresholds for Central East– Green Park,
OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Figure F-6 presents the log-likelihood (equation (3-9)) for each initial value (seed) and tolerance threshold.

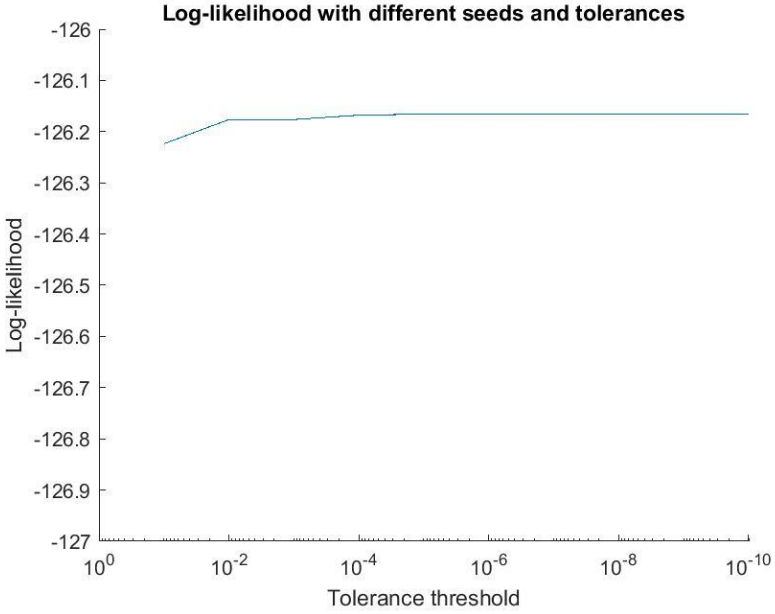


Figure F-6 Log-likelihood,
given different initial values and tolerance thresholds for Central East– Green Park
OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Case 3 Jubilee West – Jubilee Central

The settings described in **Section 3.3.1** were used for centroid initialisation (K-means ++), distances (Euclidean square) and update methods (online phase). Conducting trials with different seeds for the random number generator it gave two possible solutions for $\mu_{r,IJ}^{KMS}$, $\sigma_{r,IJ}^{KMS}$ and $\omega_{r,IJ}^{KMS}$ (**Table F-3**)

Table F-3 Results of the K-means clustering algorithm with different seeds for **Jubilee West** – **Jubilee Central**; a) Seed=1, b) Seed=2, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	41.0	3.6	85.0%
2	55.0	5.7	15.0%

a)

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	41.0	2.9	74.1%
2	50.5	6.2	25.9%

b)

Figure F-7 and **Figure F-8** presents the estimated mean (μ_1^{MIX}) and proportion (ω_1^{MIX}) for mixture component labelled with $r = 1$.

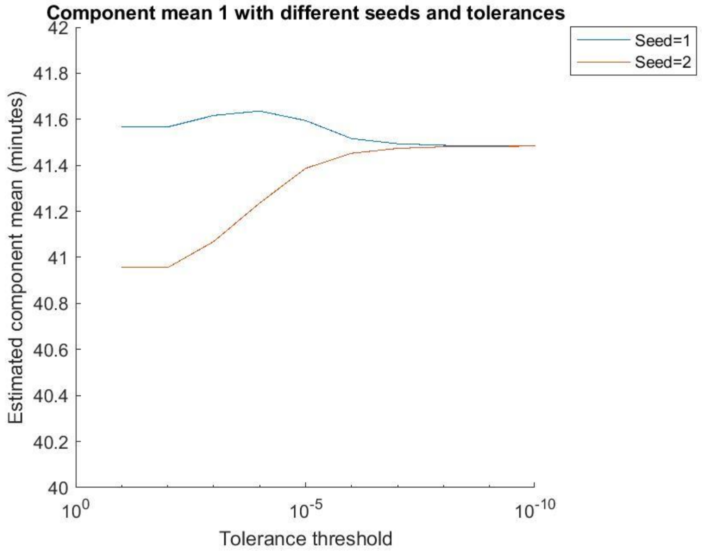


Figure F-7 Estimated mean for mixture component 1, given different initial values and tolerance thresholds for Jubilee West– Jubilee Central, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

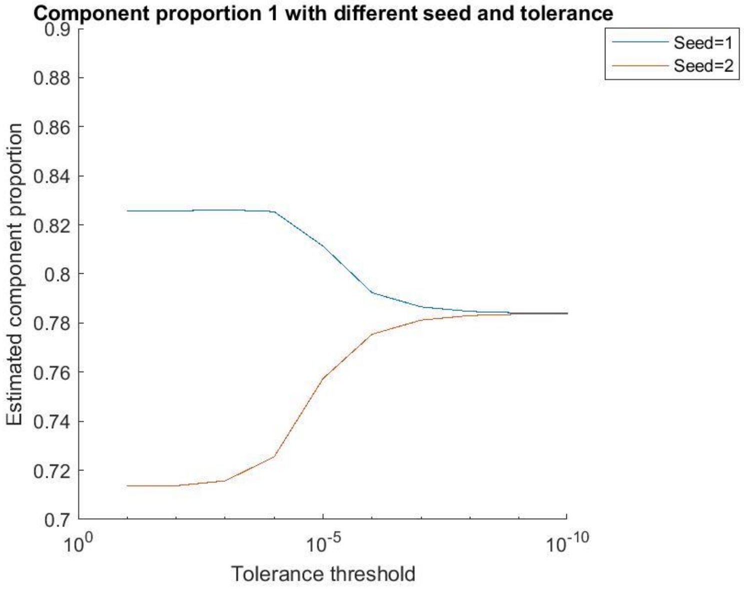


Figure F-8 Estimated proportion for mixture component 1, given different initial values and tolerance thresholds for Jubilee West – Jubilee Central, OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Figure F-9 presents the log-likelihood (equation (3-9)) for each initial value (seed) and tolerance threshold.

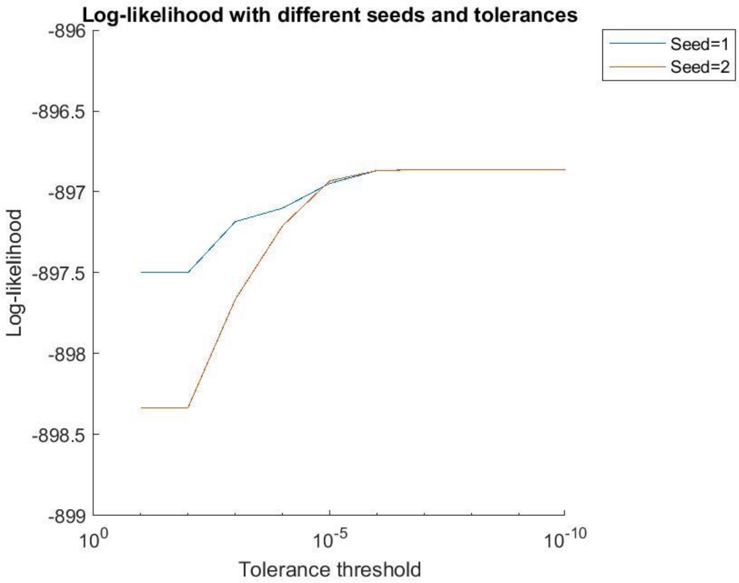


Figure F-9 Log-likelihood, given different initial values and tolerance thresholds for **Jubilee West** – **Jubilee Central**

OJTs adjusted to superstation centroid, but not according to fail-to-board delays

Appendix G
Application of the finite mixture model on the CCOJTs adjusted according to fail-to-board delays

The case studies for the application of the finite mixture model on the CCOJT dataset adjusted according to fail-to-board delays were presented in **Section 6.6**. The detailed description of the settings and of the results are reported here.

Case 1 Victoria South - Holborn

Within the dataset of CCOJTs adjusted according to fail-to-board delays ($CCOJT_{IJ}^{fail}$), all entries could be considered as valid data, because the upper outer fence (cf. **Section 3.2.1**) resulted 33 minutes, while the maximum CCOJT value is 30 minutes. This valid dataset is denoted by $CCOJT_{IJ}^{fail,0}$.

As two reasonable routes were assumed for the superstation-to-station OD pair, route choice was estimated as a two-component ($N_R = 2$) finite mixture distribution. Therefore, the K-means clustering algorithm was applied on the $CCOJT^{fail,0}$ dataset with two clusters and with the settings described in **Section 3.3.1** (K-means ++ for centroid initialisation, Euclidean square for distances and online phase update method) to produce the initial values for the EM algorithm. Conducting trials with different seeds for the random number generator it gave two possible solutions for $\mu_{r,IJ}^{KMS}$, $\sigma_{r,IJ}^{KMS}$ and $\omega_{r,IJ}^{KMS}$ (**Table G-1**).

Table G-1 Results of the K-means clustering algorithm with different seeds for **Victoria South** – **Holborn**; a) Seed=1, b) Seed=2, OJTs adjusted to superstation centroid and according to fail-to-board delays

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	18.0	1.8	74.1%
2	23.0	2.2	25.9%

a)

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	17.0	1.6	62.6%
2	22.0	2.5	37.4%

b)

Using these initial values, the EM algorithm was run with different settings for the tolerance threshold (cf. **Section 3.3.2**). **Figure G-1** and **Figure G-2** presents the estimated mean ($\mu_{1,IJ}^{MIX}$) and proportion ($\omega_{1,IJ}^{MIX}$) for mixture component labelled with $r = 1$.

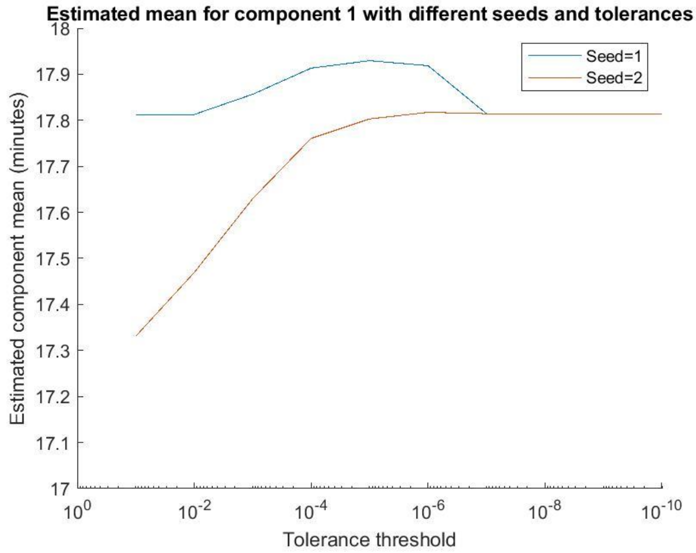


Figure G-1 Estimated mean for mixture component 1,

given different initial values and tolerance thresholds, for **Victoria South** – **Holborn**, OJTs adjusted to superstation centroid and according to fail-to-board delays

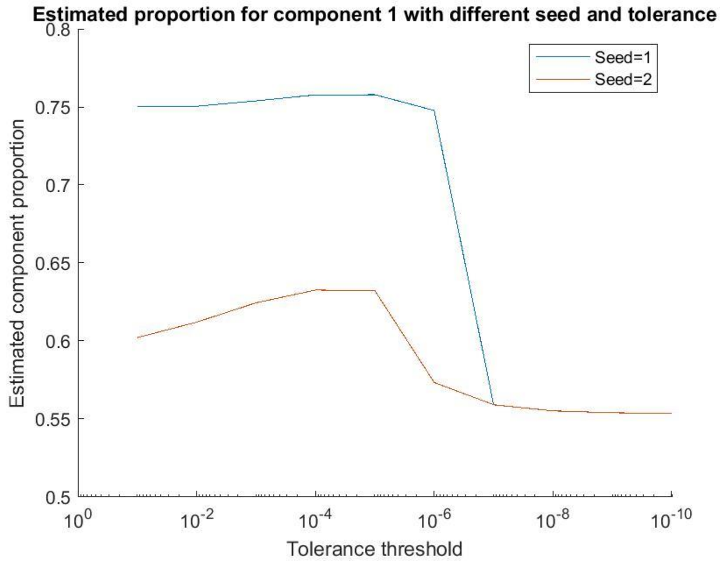


Figure G-2 Estimated proportion for mixture component 1,

given different initial values and tolerance thresholds, for **Victoria South** – **Holborn**, OJTs adjusted to superstation centroid and according to fail-to-board delays

Figure G-3 presents the log-likelihood (equation (3-9)) for each initial value (seed) and tolerance threshold.

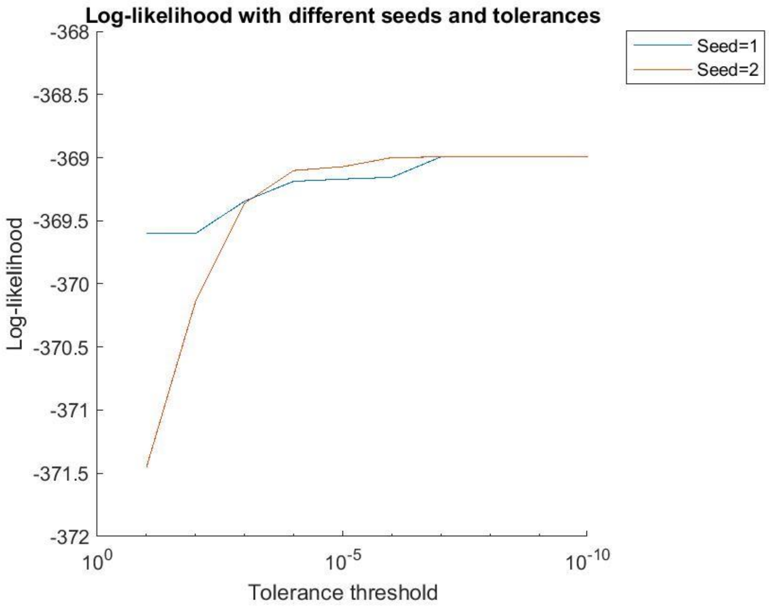


Figure G-3 Log-likelihood,

given different initial values and tolerance thresholds, for Victoria South – Holborn OJTs adjusted to superstation centroid and according to fail-to-board delays

From these results it is understood, that when the tolerance threshold is 1e-06 or greater, the EM algorithm converges to a root close to the initial value for seed 1. But when the tolerance threshold is 1e-07 or smaller, the EM algorithm converges to a root around 17.8 minutes for the mean and 55.3% for the component proportion for both seeds (cf. **Figure G-1** and **Figure G-2**). Similar properties could be observed for the other mixture component (labelled with $r = 2$). The log-likelihood shows a considerable jump between the tolerance threshold of 1e-02 and 1e-03 for both seeds (**Figure G-3**). Among the estimates, the one with seed 1 and tolerance threshold 1e-06 gives the best approximation to RODS results (cf. **Table 6-5**).

Case 2 Central East – Green Park

Within the dataset of CCOJTs adjusted according to fail-to-board delays ($CCOJT_{IJ}^{fail}$), all entries could be considered as valid data, because the upper outer fence (cf. **Section 3.2.1**) resulted 35.3 minutes, while the maximum CCOJT value is 28 minutes. This valid dataset is denoted by $CCOJT_{IJ}^{fail,0}$.

As two reasonable routes were assumed for the superstation-to-station OD pair, route choice was estimated as a two-component ($N_R = 2$) finite mixture distribution. Therefore, the K-means clustering algorithm was applied on the $CCOJT_{IJ}^{fail,0}$ dataset with two clusters and with the settings described in **Section 3.3.1** (K-means ++ for centroid initialisation, Euclidean square for distances and online phase update method) to produce the initial values for the EM algorithm. Conducting trials with different seeds for the random number generator it gave two possible solutions for $\mu_{r,IJ}^{KMS}$, $\sigma_{r,IJ}^{KMS}$ and $\omega_{r,IJ}^{KMS}$ (**Table G-2**).

Table G-2 Results of the K-means clustering algorithm with different seeds for Central East– **Green Park**; a) Seed=1, b) Seed=2, OJTs adjusted to superstation centroid and according to fail-to-board delays

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	18.0	1.8	74.5%
2	24.5	2.1	25.5%

a)

Label	K-means clustering		
r []	$\mu_{r,IJ}^{KMS}$ [min]	$\sigma_{r,IJ}^{KMS}$ [min]	$\omega_{r,IJ}^{KMS}$ [%]
1	17.5	1.4	51.1%
2	22.0	2.6	48.9%

b)

Using these initial values, the EM algorithm was run with different settings for the tolerance threshold (cf. **Section 3.3.2**). **Figure G-4** and **Figure G-5** presents the estimated mean ($\mu_{1,IJ}^{MIX}$) and proportion ($\omega_{1,IJ}^{MIX}$) for mixture component labelled with $r = 1$.

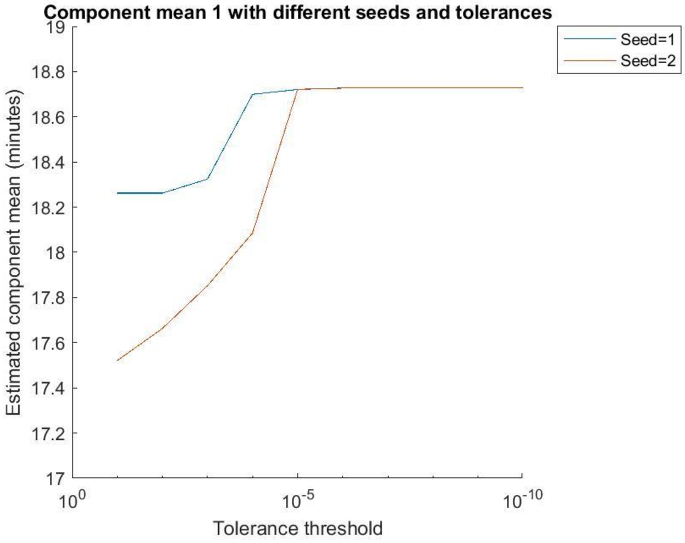


Figure G-4 Estimated mean for mixture component 1, given different initial values and tolerance thresholds, for Central East– Green Park, OJTs adjusted to superstation centroid and according to fail-to-board delays

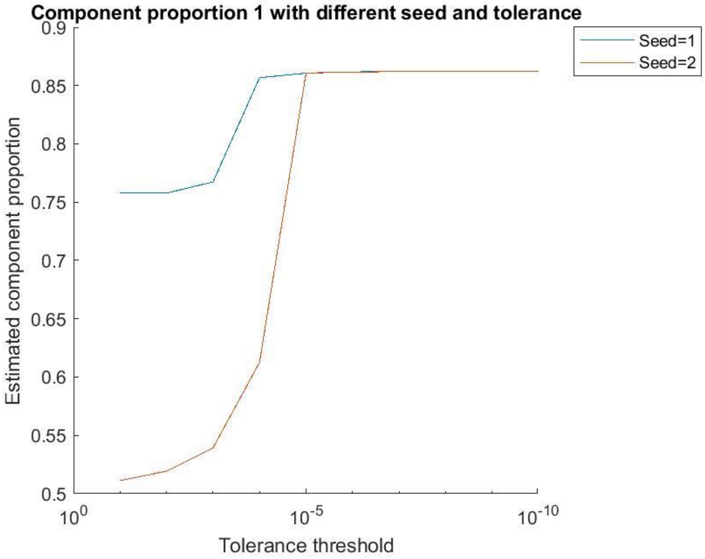


Figure G-5 Estimated proportion for mixture component 1, given different initial values and tolerance thresholds, for Central East– Green Park, OJTs adjusted to superstation centroid and according to fail-to-board delays

Figure G-6 presents the log-likelihood (equation (3-9)) for each initial value (seed) and tolerance threshold.

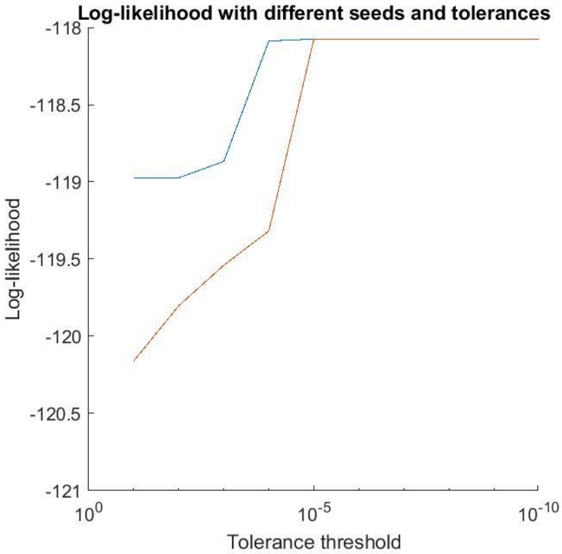


Figure G-6 Log-likelihood,

given different initial values and tolerance thresholds, for **Central East**– **Green Park** OJTs adjusted to superstation centroid and according to fail-to-board delays

From these results it is understood, that when the tolerance threshold is 1e-04 or greater, the EM algorithm converges to different roots for seed 1 and 2; but when it is 1e-05 or smaller, it converges to the same root for the two seeds: 18.7 minutes for the mean journey time (cf. **Figure G-4**) and 86.2% for the proportion of component 1 (cf. **Figure G-5**). It starts plateauing from the tolerance threshold value of 1e-07. Similar properties could be observed for the other mixture component (labelled with $r = 2$). The log-likelihood exhibits a considerable jump between the tolerance threshold of 1e-03 and 1e-04 for seed 1 and between 1e-04 and 1e-05 for seed 2 (**Figure G-6**). Due to these considerations, the estimate with seed 1 and tolerance threshold 1e-07 was chosen (cf. **Table 6-10**).