Modelling route choice behaviour with incomplete data: an application to the London Underground

Qian Fu

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

The University of Leeds
Institute for Transport Studies

September 2014

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

Fu, Q., Liu, R. and Hess, S. 2012. On considering journey time variability and passengers' path choice: an empirical study using Oyster card data on the London Underground, paper presented at the *Fifth International Symposium on Transportation Network Reliability*, Hong Kong, China, 18th-19th December 2012.

Fu, Q., Liu, R. and Hess, S. 2012. A review on transit assignment modelling approaches to congested networks: a new perspective, paper presented at the *15th Meeting of the EURO Working Group on Transportation*, Paris, France, 10th-13th September 2012. Published in: Procedia – Social and Behavioral Sciences, Vol. 54, pp.1145-1155.

Chapter 3 and **Chapter 4** of the thesis are based on the work of the conference paper below, of which the author drafted the paper and the co-authors provided commentary and comments throughout:

Fu, Q., Liu, R. and Hess, S. 2014. A Bayesian modelling framework for individual passenger's probabilistic route choices: a case study on the London Underground, paper presented by poster at the *Transportation Research Board 93rd Annual Meeting*, Washington, D.C., USA, 12th-16th January 2014.

The following three conference papers laid the groundwork for the thesis:

Fu, Q. 2014. A Bayesian modelling framework for individual passenger's probabilistic route choices: a case study on the London Underground, paper presented at the *46th Annual Conference of the Universities' Transport Study Group*, Newcastle upon Tyne, UK, 6th-8th January 2014.

Fu, Q. 2012. Bayesian inference of passengers' path choices with incomplete data – an application for the London Underground with Oyster card, paper presented at the *17th International Conference of Hong Kong Society for Transportation Studies*, Hong Kong, China, 15th-17th December 2012.

Fu, Q. 2012. Understanding the travel patterns on the London rail network: the use of Oyster card data, paper presented at the *44th Annual Conference of the Universities' Transport Study Group*, Aberdeen, UK, 4th-6th January 2012.

Acknowledgements

I owe my deepest gratitude to my parents and all my family for their great encouragement and love to me. Their absolute complete faith and trust in me always inspires me to be stronger and spurs me on to overcome challenges as well as difficulties.

And certainly, I would love to express my heartfelt appreciation and thanks to my two supervisors, Ronghui Liu and Stephane Hess, for their valuable advice and patience during the past four years, especially when I lost my train of thought that got all tangled up later in a sequence of incomplete thesis chapters. I am more than grateful for their contributions of precious time and inspirational ideas that did help me to cut through the confusion. The thesis would not have been possible without their guidance. In addition, I must specially mention the 'deadline plan' drawn up with Stephane's help at the final stage before the thesis was submitted, which eventually drove the thesis to reach the end.

Furthermore, I am enormously thankful to Jan-Dirk Schmöcker and Richard Connors for being my examiners; and I deeply appreciate all their insightful comments on the thesis and useful suggestions for its improvement. Also, I would like to thank Mark Wardman for his supervision of the whole viva process.

Moreover, many thanks to the China Scholarship Council and the University of Leeds for funding my PhD study, and to Andrew Gaitskell, Maunder Geoffrey and Duncan Horne from the Transport for London for their considerable support on provision of the essential data used in this research work.

What is more, I do feel profoundly indebted to loads of people who are always too overly generous with their kindness: Haiyue Yuan, a trusted and dear old friend, who positively encouraged me to study in the UK and helped me too much with my PhD application; Leigang Cao and Mingfu Guan, who always came to my aid whenever there was a harsh time during the course of the past four years; Christopher Kelsey, who not only taught me how to play squash, but also volunteered to help me move home out of the tenth floor when all the lifts were unavailable; Tiong Sing Khien and Evona Teh, who unfortunately kept hearing

about my endless complaints when we met up, but who are just like my brother and sister always cheering me up; Shuang Liang, who was willing to stand out to act as a 'solicitor' when I received the most unfair treatment ever in my life; Anthony Magee, the best office mate, who really saved my spine when I was homeless for a couple of weeks; and of course, Arwa Sayegh, Julián Arellana, Tamas Nadudvari, Jay & Joan Cohen, Sheriff Idriss-Yahya, Mahmoud Al-Khazaleh, Mathilde Francheteau, Tamara Muroiwa, Nurul Hidayati, Stefanie Peer, Chaiyon Tongsukkaeng, and Jue Wang, who made me convince myself that I am a human, and at least not a bad one.

In particular, I would also like to thank all the lovely and likeable companions of the SF group, who made for the most interesting academic year ever since I came to Leeds. Special thanks to Rui Mo, the key promoter for this group, as well as Tang-Po Chang, Chun-Yen Cheng, Juvena Ng, Jeff Tjiong, Xiao Wang, Yi-Fan Wen, Zhonghao Xie, Yao Yao, Yue Yin and Guoqing Zhang. I truly cherish all the memories of our get-together, which keep making me feel so cheerful.

And last but not least, here I still wish to place on record my grateful thanks to Paul Adams, Afzal Ahmed, Segun Aluko, Chandra Balijepalli, Zahara Batool, Valerio Benedetto, Joanne Davies, Fangce Guo, Keith Harrison, Daryl Hibberd, Calvin Jephcote, Andrew Koh, Andyka Kusuma, Frank Lai, Chao Li, Junfang Li, Zhengyang Ling, Mojtaba Moharrer, Ying Nan, Nobuhiro Sanko, Padma Seetharaman, Nikolas Thomopoulos, Andrew Tomlinson, Yue Zhang, as well as those who I carelessly omitted to mention in these two pages …

It is all those kind-hearted people who let me believe that I am able to complete my PhD work because they are always here by my side.

Qian Fu
付谦

Abstract

This thesis develops a modelling framework for learning route choice behaviour of travellers on an underground railway system[1], with a major emphasis on the use of smart-card data.

The motivation for this topic comes from two respects. On the one hand, in a metropolis, particularly those furnished with massive underground services (*e.g.* London, Beijing and Paris), severe passenger-traffic congestion may often occur, especially during rush hours. In order to support the public transport managers in taking actions that are more effective in smoothening the passenger flows, there is bound to be a need for better understanding of the passengers' routing behaviour when they are travelling on such public transport networks. On the other hand, a wealth of travel data is nowadays readily obtainable, largely owing to the widespread implementation of **a**utomatic **f**are **c**ollection systems (AFC) as well as popularity of smart cards on the public transport. Nevertheless, a core limitation of such data is that the actual route-choice decisions taken by the passengers might not be available, especially when their journeys involve alternative routes and/or within-station interchanges. Mostly, the AFC systems (*e.g.* the Oyster system in London) record only data of passengers' entry and exit, rather than their route choices. We are thus interested in whether it is possible to analytically infer the route-choice information based on the 'incomplete' data.

Within the scope of this thesis, passengers' single journeys are investigated on a station basis, where sufficiently large samples of the smart-card users' travel records can be gained. With their journey time data being modelled by simple finite mixture distributions, Bayesian inference is applied to estimate posterior probabilities for each route that a given passenger might have chosen from all

[1] The 'underground' system is also known as the 'Tube' (especially in London), 'metro' (*e.g.* in Moscow, Paris, Shanghai, Madrid and Santiago), 'subway' (*e.g.* in Beijing, New York City and Seoul), 'mass rapid transit' (especially in Singapore), and 'U-Bahn' (especially in Germany), *etc*. *See also* "List of metro systems" (Wikipedia, the free encyclopedia, 2014), available online at https://en.wikipedia.org/wiki/List_of_metro_systems; last accessed on 30th September 2014.

possible alternatives. We learn the route-choice probabilities of every individual passenger in any given sample, conditional on an observation of the passenger's journey time. Further to this, the estimated posterior probabilities are also updated for each passenger, by taking into account additional information including their entry times as well as the timetables. To understand passengers' actual route choice behaviour, we then make use of adapted discrete choice model, replacing the conventional dependent variable of actual route choices by the posterior choice probabilities for different possible outcomes. This proposed methodology is illustrated with seven case studies based in the area of central zone of the London Underground network, by using the Oyster smart-card data. Two standard mixture models, *i.e.* the probability distributions of Gaussian and log-normal mixtures, are tested, respectively. The outcome demonstrates a good performance of the mixture models. Moreover, relying on the updated choice probabilities in the estimation of a multinomial logit latent choice model, we show that we could estimate meaningful relative sensitivities to the travel times of different journey segments. This approach thus allows us to gain an insight into passengers' route choice preferences even in the absence of observations of their actual chosen routes.

Table of Contents

Acknowledgements ... **iii**

Abstract ... **v**

Table of Contents ... **vii**

List of Tables ... **xi**

List of Figures ... **xiii**

List of Abbreviations ... **xv**

Chapter 1 Introduction ... **1**

1.1 Background ... 1

1.2 Research scope and objectives ... 4

1.3 Methodological framework ... 4

1.4 Outline ... 6

1.5 Contributions ... 8

Chapter 2 Modelling route choices on public transport ... **9**

2.1 Introduction ... 9

2.2 The foundation of modelling route choices ... 11

2.2.1 Transit network and alternative routes ... 11

2.2.2 Journey cost ... 13

2.2.3 Fundamental behavioural assumptions ... 14

2.3 Route choice behaviours ... 16

2.3.1 Moving through passageways ... 17

2.3.2 Waiting and failures of boarding ... 18

2.3.3 Travelling and on-board crowdedness ... 21

2.4 Discussions ... 22

2.4.1 Route choice and journey time variability ... 22

2.4.2 Data for route choices ... 24

2.5 Summary and conclusions ... 27

Chapter 3 Bayesian inference of probabilistic route choices ... **31**

3.1 Introduction ... 31

3.2 Problem description ... 34

3.3 Finite mixture model for journey time distribution ... 41

3.3.1 Model formulation 43

3.3.2 Incomplete data..46

3.3.3 Model estimation..48

3.4 Inference of passenger traffic distribution..52

3.4.1 *Naive* inference..53

3.4.2 *Effective* inference 54

3.5 Interpretation and validation of mixture model estimates....................55

3.5.1 Expectation of journey time for a given pair of O-D....................56

3.5.2 General principles 64

3.6 Summary and conclusions..67

Chapter 4 Application of mixture models of route choices: case studies on the London Underground 71

4.1 Introduction..71

4.2 Data description 73

4.2.1 The Oyster-card data..73

4.2.2 Data for computation of route average journey time....................76

4.2.3 Data from Rolling Origin and Destination Survey 77

4.3 Case studies on the London Underground..77

4.3.1 Two alternative routes (*Case-1* – *Case-3*)..81

4.3.1.1 *Case-1*: Victoria – Holborn 81

4.3.1.2 *Case-2*: Euston – St. James's Park 91

4.3.1.3 *Case-3*: Victoria – Liverpool Street 98

4.3.2 More than two alternative routes (*Case-4* and *Case-6*)............106

4.3.2.1 A case of three routes (*Case-4*): Angel – Waterloo.......107

4.3.2.2 A case of four routes (*Case-6*): Euston – South Kensington..116

4.4 Summary and conclusions..124

Chapter 5 Updating route choice probabilities..127

5.1 Introduction..127

5.2 Correlation between passengers' entry and trains' timetable............131

5.3 Updating the posterior route-choice probabilities..134

5.3.1 Factoring additional condition..134

5.3.2 Conditional journey time distribution 136

5.3.3 Deriving updated posterior probabilities 138

5.4 Implementation of the updating approach 139

5.4.1 Data issue 140

5.4.2 An example based on *Case-1* 141

5.5 Summary and conclusions 147

Chapter 6 A latent route choice model 149

6.1 Introduction 149

6.2 Choice modelling 151

6.2.1 Standard logit choice probability 151

6.2.2 Route choice models 152

6.2.3 Data for choice modelling 154

6.3 Latent choice probability 156

6.4 Empirical examples on the London Underground network 158

6.4.1 Data description 159

6.4.2 Utility function specification 160

6.4.3 Estimation results 161

6.4.4 An extended example 165

6.5 Summary and conclusions 169

Chapter 7 Concluding remarks and future research 173

7.1 Summary of the thesis 173

7.2 Directions for future research 176

7.2.1 Refinement of naive Bayesian modelling framework 178

7.2.2 Route-choice set generation 180

7.2.3 Refinement and validation of latent route-choice model 180

Bibliography 183

Appendix A Explanatory notes on logical dependency and relationship between route choice and journey time 193

Appendix B Standard Tube map (© Transport for London) 197

Appendix C Application of mixture models: *Case-5* and *Case-7* 199

C.1 *Case-5*: Liverpool Street – Green Park 199

C.2 *Case-7*: Victoria – Waterloo 206

Appendix D Updated posterior probabilities for *Case-2* – *Case-7* 213

D.1 Cases of two alternative routes 213

D.1.1 *Case-2*: Euston – St. James's Park 213

D.1.2 *Case-3*: Victoria – Liverpool Street 215

D.2 Cases of three alternative routes.. 217

D.2.1 *Case-4*: Angel – Waterloo... 217

D.2.2 *Case-5*: Liverpool Street – Green Park.. 219

D.3 Cases of four alternative routes 222

D.3.1 *Case-6*: Euston – South Kensington 222

D.3.2 *Case-7*: Victoria – Waterloo 225

Appendix E Notation 229

E.1 Symbols based on the English alphabet.. 229

E.2 Symbols based on the Greek alphabet.. 234

List of Tables

Table 4.1 Time-bands set by TfL 75

Table 4.2 An introductory summary of the LU case studies 80

Table 4.3 ***Case-1***: parameter estimates of GM and LNM models 84

Table 4.4 ***Case-1***: inferences of proportion of passenger traffic 86

Table 4.5 ***Case-1***: expected journey times of simulated samples 87

Table 4.6 ***Case-1***: matching mixture components with routes 88

Table 4.7 ***Case-1***: goodness-of-fit test result 90

Table 4.8 ***Case-2***: parameter estimates of GM and LNM models 93

Table 4.9 ***Case-2***: inferences of proportion of passenger traffic 95

Table 4.10 ***Case-2***: expected journey times of simulated samples 95

Table 4.11 ***Case-2***: matching mixture components with routes 96

Table 4.12 ***Case-2***: goodness-of-fit test result 98

Table 4.13 ***Case-3***: parameter estimates of GM and LNM models 101

Table 4.14 ***Case-3***: inferences of proportion of passenger traffic 103

Table 4.15 ***Case-3***: expected journey times of simulated samples 104

Table 4.16 ***Case-3***: matching mixture components with routes 105

Table 4.17 ***Case-3***: goodness-of-fit test result 106

Table 4.18 ***Case-4***: parameter estimates of GM and LNM models 110

Table 4.19 ***Case-4***: inferences of proportion of passenger traffic 112

Table 4.20 ***Case-4***: expected journey times of simulated samples 113

Table 4.21 ***Case-4***: matching mixture components with routes 114

Table 4.22 ***Case-4***: goodness-of-fit test result 116

Table 4.23 ***Case-6***: parameter estimates of GM and LNM models 119

Table 4.24 ***Case-6***: inferences of proportion of passenger traffic 121

Table 4.25 ***Case-6***: expected journey times of simulated samples 121

Table 4.26 ***Case-6***: matching mixture components with routes 122

Table 4.27 ***Case-6***: goodness-of-fit test result 124

Table 5.1 Summary of sample sizes and elected mixture models 140

Table 5.2 ***Case-1***: proportion of passenger traffic 147

Table 6.1 Estimation results for MNL models using $\mathbf{\Pi}^{\mathtt{MIX}}_{\Delta_{5\%}}$ 162

Table 6.2 Estimation results for MNL models using $\boldsymbol{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ 163

Table 6.3 Estimation results for an additional MNL example using $\boldsymbol{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$ 166

Table 6.4 Estimation results for an additional MNL example using $\boldsymbol{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ 167

Table 6.5 Goodness of fit of models of ***Test-1*** and ***Test-2*** 167

Table C.1 ***Case-5***: parameter estimates of GM and LNM models 201

Table C.2 ***Case-5***: inferences of proportion of passenger traffic 203

Table C.3 ***Case-5***: expected journey times of simulated samples 203

Table C.4 ***Case-5***: matching mixture components with routes 204

Table C.5 ***Case-5***: goodness-of-fit test result 206

Table C.6 ***Case-7***: parameter estimates of GM and LNM models 207

Table C.7 ***Case-7***: inferences of proportion of passenger traffic 209

Table C.8 ***Case-7***: expected journey times of simulated samples 209

Table C.9 ***Case-7***: matching mixture components with routes 210

Table C.10 ***Case-7***: goodness-of-fit test result 210

Table D.1 ***Case-2***: proportion of passenger traffic for each route 213

Table D.2 ***Case-3***: proportion of passenger traffic for each route 215

Table D.3 ***Case-4***: proportion of passenger traffic for each route 217

Table D.4 ***Case-5***: proportion of passenger traffic for each route 220

Table D.5 ***Case-6***: proportion of passenger traffic for each route 222

Table D.6 ***Case-7***: proportion of passenger traffic for each route 225

List of Figures

Figure 1.1 Methodological framework of the thesis 5

Figure 3.1 A single O-D network with multiple travel routes 35

Figure 4.1 Tube map for Zone-1 (© 2013 Transport for London) 72

Figure 4.2 ***Case-1***: the LU network 81

Figure 4.3 ***Case-1***: summary of OJT^{OBS} data 82

Figure 4.4 ***Case-1***: posterior probabilities of route choices 86

Figure 4.5 ***Case-1***: estimated mixture distributions of OJT 89

Figure 4.6 ***Case-2***: the LU network 91

Figure 4.7 ***Case-2***: summary of OJT^{OBS} data 92

Figure 4.8 ***Case-2***: posterior probabilities of route choices 94

Figure 4.9 ***Case-2***: estimated mixture distributions of OJT 97

Figure 4.10 ***Case-3***: the LU network 99

Figure 4.11 ***Case-3***: summary of OJT^{OBS} data 100

Figure 4.12 ***Case-3***: posterior probabilities of route choices 102

Figure 4.13 ***Case-3***: estimated mixture distributions of OJT 105

Figure 4.14 ***Case-4***: the LU network 108

Figure 4.15 ***Case-4***: summary of OJT^{OBS} data 109

Figure 4.16 ***Case-4***: posterior probabilities of route choices 111

Figure 4.17 ***Case-4***: estimated mixture distributions of OJT 115

Figure 4.18 ***Case-6***: the LU network 117

Figure 4.19 ***Case-6***: summary of OJT^{OBS} data 118

Figure 4.20 ***Case-6***: posterior probabilities of route choices 119

Figure 4.21 ***Case-6***: estimated mixture distributions of OJT 123

Figure 5.1 Bayesian-network structures for probabilistic route choices 128

Figure 5.2 The causality between T^{ENT} and T^{EXT} 131

Figure 5.3 A simplified structure of passengers' journey 133

Figure 5.4 ***Case-1***: comparisons between OJT_q^{OBS} and $\delta_{qr}^{\mathrm{EXP}}$, given $\Delta_{5\%}$ 142

Figure 5.5 ***Case-1***: comparisons between π_{qr}^{MIX} (LNM) and π_{qr}^{UMM} 144

Figure 7.1 A modelling framework to be developed for future research 177

Figure C.1 ***Case-5***: the LU network 200

Figure C.2 ***Case-5***: summary of OJT^{OBS} data 201

Figure C.3 ***Case-5***: posterior probabilities of route choices 202

Figure C.4 ***Case-5***: estimated mixture distributions of OJT 205

Figure C.5 ***Case-7***: the LU network 206

Figure C.6 ***Case-7***: summary of OJT^{OBS} data 207

Figure C.7 ***Case-7***: posterior probabilities of route choices 208

Figure C.8 ***Case-7***: estimated mixture distributions of OJT 211

Figure D.1 ***Case-2***: comparisons between OJT_q^{OBS} and $\delta_{qr}^{\mathrm{EXP}}$, given $\Delta_{5\%}$ 213

Figure D.2 ***Case-2***: comparisons between π_{qr}^{MIX} (LNM) and π_{qr}^{UMM} 214

Figure D.3 ***Case-3***: comparisons between OJT_q^{OBS} and $\delta_{qr}^{\mathrm{EXP}}$, given $\Delta_{5\%}$ 215

Figure D.4 ***Case-3***: comparisons between π_{qr}^{MIX} (LNM) and π_{qr}^{UMM} 216

Figure D.5 ***Case-4***: comparisons between OJT_q^{OBS} and $\delta_{qr}^{\mathrm{EXP}}$, given $\Delta_{5\%}$ 217

Figure D.6 ***Case-4***: comparisons between π_{qr}^{MIX} (GM) and π_{qr}^{UMM} 218

Figure D.7 ***Case-5***: comparisons between OJT_q^{OBS} and $\delta_{qr}^{\mathrm{EXP}}$, given $\Delta_{5\%}$ 219

Figure D.8 ***Case-5***: comparisons between π_{qr}^{MIX} (GM) and π_{qr}^{UMM} 220

Figure D.9 ***Case-6***: comparisons between OJT_q^{OBS} and $\delta_{qr}^{\mathrm{EXP}}$, given $\Delta_{5\%}$ 222

Figure D.10 ***Case-6***: comparisons between π_{qr}^{MIX} (GM) and π_{qr}^{UMM} 223

Figure D.11 ***Case-7***: comparisons between OJT_q^{OBS} and $\delta_{qr}^{\mathrm{EXP}}$, given $\Delta_{5\%}$ 225

Figure D.12 ***Case-7***: comparisons between π_{qr}^{MIX} (GM) and π_{qr}^{UMM} 226

List of Abbreviations

ORR	**O**ffice of **R**ail **R**egulation
LU	**L**ondon **U**nderground
O-D	**o**rigin and **d**estination
AFC	**a**utomatic **f**are **c**ollection
MNL	**m**ulti**n**omial **l**ogit
TfL	**T**ransport **f**or **L**ondon
GPS	**g**lobal **p**ositioning **s**ystem
PDF	**p**robability **d**ensity **f**unction
EM	**E**xpectation-**M**aximisation
CI	**c**onfidence **i**nterval
CL	**c**onfidence **l**evel
AIC	**A**kaike's **i**nformation **c**riterion
BIC	**B**ayesian **i**nformation **c**riterion
NRMSE	**n**ormalized **r**oot **m**ean **s**quare **e**rror
OJT	**O**yster **j**ourney **t**ime
IQR	**i**nter**q**uartile **r**ange
AEI	**A**ccess, **E**gress and **I**nterchange
DE	**D**irect **E**nquiries
RODS	**R**olling **O**rigin and **D**estination **S**urvey
GM	**G**aussian **m**ixture
LNM	**l**og-**n**ormal **m**ixture
EB	**e**ast**b**ound
WB	**w**est**b**ound
NB	**n**orth**b**ound
SB	**s**outh**b**ound
RP	**r**evealed **p**reference

Chapter 1
Introduction

1.1 Background

Public transport[1] in almost every metropolis, such as in London, Beijing and Paris, to name but a few, has furnished travellers with a highly sophisticated and interconnected mass transit system. Meanwhile, a boom in travel demand in the urban areas, particularly a surge in passenger traffic during a certain period (*e.g.* rush hour), could make cumulative impact on regularity as well as reliability of the transit services. An in-depth understanding of the passengers' travel behaviour in the network is interestingly significant to transit planning, operations and the travel demand management.

In the case of an underground rail system[2], such as the **L**ondon **U**nderground (LU), a number of stations may serve for 'interchanges' at which two or more 'transit lines' intersect and/or the service directions change. A transit line, or simply a line, refers to a fleet of trains running along a particular 'route' that links two terminal stations within the network (*cf.* Ortúzar and Willumsen, 2011, p.376), with one terminal being an origin and another a destination. Given this definition, there could be either a single or multiple alternative routes ready for carrying passenger traffic in both directions between a pair of **o**rigin and **d**estination (O-D) stations. Each of the alternatives is referred to as a travel route (or simply a route), which is composed of one or several route sections; and a route section can be a portion of a route, which is between two adjacent interchange stations (*cf.* De Cea and Fernández, 1993).

[1] The terms 'public transport' and 'transit' will be used interchangeably in this thesis.

[2] According to the **O**ffice of **R**ail **R**egulation (ORR) (2014), "an underground system is defined as an electric railway public transport network (a metro or subway system) that runs both above and underground" (*cf.* Footnote 1 on page v). The term 'underground' being referred to throughout this thesis is in line with this definition.

Suppose that there is availability of a few routes for passengers travelling on the underground network. All the passengers choose from among available lines to complete their journeys; and they might need to transfer between different lines that are serving the same specific route. It is noteworthy that the total travel time through a certain route would vary (within a day as well as between days) for many reasons, such as engineering work and adjustment for operation schemes. Furthermore, there are likely to be both similarities as well as differences in the passengers' perceptions and sensitivities to different attributes shared by the alternative routes. Such attributes, in addition to the travel time that we have mentioned already, could also involve fare, number of interchanges, preference of a certain line and so forth. Moreover, for any given O-D pair, the passengers' choice sets of the travel routes can differ from individual-to-individual. That is to say, there are also differences among their 'route-choice tasks', in each of which one individual must choose one and only one of the alternative routes from his/her own route-choice set for a certain pair of O-D stations. In view of these facts, effective approaches to reproducing and analysing the passengers' route choice behaviour are certainly attracting interest from public-transport planners and operators. This is because such modelling instruments could provide those professionals with necessary knowledge of passenger-flow distributions across the network, and thereby assisting them in identifying traffic bottlenecks and delivering a more efficient transit service, especially at rush hour. And what is more important, such information is vital for system managers to grasp the usage patterns of different transit lines, thereby offering insight into the network utilization, especially for dealing with planned and unexpected disruptions. Additionally, the aggregate passenger traffic or ridership on the different lines can, if needed, serve as evidence for transport authorities to cope with the settlement of fare revenues among stakeholders, such as multiple line operating companies.

On the other hand, in a bid to maintain or strengthen the operational efficiency, it is also of critical importance that the policy makers have a collection of facts and data of the passengers' evolving travel behaviour. From the perspective of modelling, the passengers' travel behaviour is learnt from diverse mathematical models of their route choices, which could inform the passengers' relative sensitivities to a range of factors underlying their decision-making process. So far, numerous studies have been devoted to developing these sorts of models, which may be broadly divided into two very different approaches: the route choice modules used for transit assignment models and the discrete choice modelling approaches. Although the former methodology, which will be discussed in **Chapter 2**, might reproduce the choice process in more detail, the latter, which will be discussed in **Chapter 6**, may have more advantages in understanding the behaviour and causes of the passengers' choices. However, it is also noted that the development of such models must rely on analysts observing data of each individual passenger's actual route choice. In other words, only when the real data is explicitly presented, parameters for the model can then be estimated.

Usually, the real route-choice data could be acquired via conducting manual surveys and passive monitoring. Either way can be very costly to gather data of sufficiently large samples; and in some circumstances, the data might be inexplicable due to a lack of accuracy or even loss of key information. In this connection, the availability, as well as the accessibility, of the data about each individual's actual choice would act as a prime determinant of developing the route choice models that offer predictive value.

In another regard, the **a**utomatic **f**are **c**ollection (AFC) system driven by smart cards on the public transport can gather a wealth of individual passengers' travel data, which is readily accessible.[3] Nevertheless, the route-choice information is still not available in its database; commonly, just entry and exit stations are recorded. This is indeed worth our best thinking and efforts in exploiting that data in connection with travel demand forecasting and management for public transport, especially because of the huge amounts of individual journey histories being recorded. In addition, it leads us to envisage the possibility of finding out the information relevant to the individuals' route choices from the AFC database. Again, however, given the fact that there is no firm or any direct evidence of passengers' route choices, any route-choice information gained from the smart-card data would have to be represented in a probabilistic setting. That is, the actual chosen route of each individual could only be known up to a choice

[3] The travel data recorded by the smart-card system is hereafter referred to as smart-card data. More details about it are described in **Chapter 2**.

probability. This thus presents us with a new research topic that is going to be addressed and discussed in this thesis.

1.2 Research scope and objectives

From the issues pointed out in the previous section, this research is principally aimed at developing a route choice model in order to gain an understanding of route choice behaviour of public-transport users. Our interest and efforts are focused only on the underground system. As has also been mentioned in the previous section, the data shortage is a main obstacle to the model development. It is our initiative to explore the possibility of digging for information about passengers' route choices from the smart-card data. In this regard, two additional objectives, which serve as the prerequisites for the stated aim stated, are to:

1. examine the connection between the smart-card data and passengers' route choices; and
2. analyse and discuss the ways of learning the information, particularly about every passenger's route choices and choice probabilities.

Further (depending on the excavated information), another aim of the thesis is to reveal the traffic loadings of passengers on different possible travel routes.

Moreover, this thesis conducts case studies on the LU using the Oyster smart-card data[4] from the Oyster system implemented across the public-transport network in London. It is expected that the established approach would be adaptable and applicable to other similar underground network.

1.3 Methodological framework

Following the background introduced above, the methodological framework of this thesis is illustrated in **Figure 1.1** (*see next page*).

[4] More details about the Oyster smart-card system and its data are described in **Section 4.2.1**.

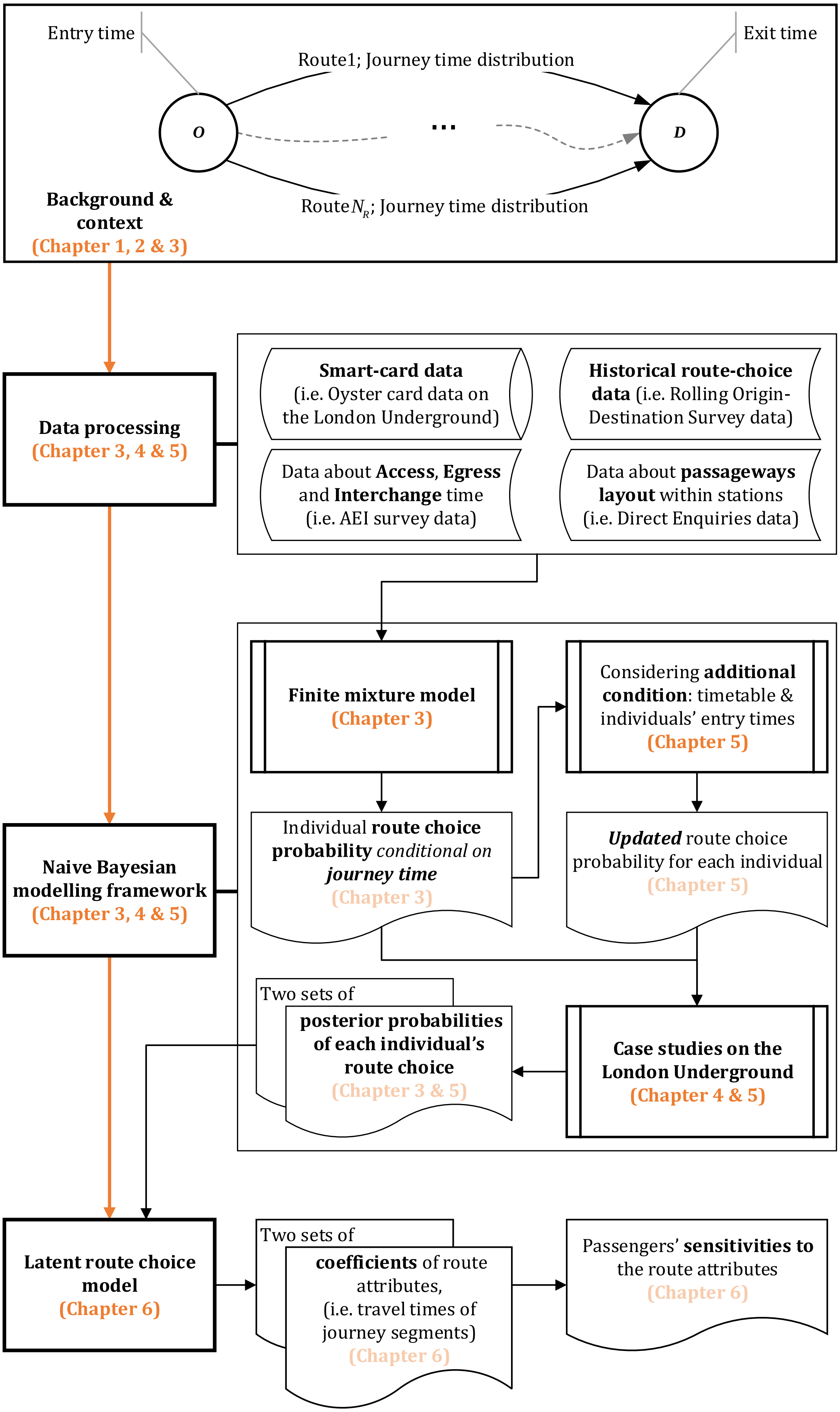

Entry time
Exit time
Route1; Journey time distribution
O
...
D
Route N_R; Journey time distribution
Background & context
(Chapter 1, 2 & 3)
Data processing
(Chapter 3, 4 & 5)
Smart-card data
(i.e. Oyster card data on the London Underground)
Historical route-choice data (i.e. Rolling Origin-Destination Survey data)
Data about Access, Egress and Interchange time
(i.e. AEI survey data)
Data about passageways layout within stations
(i.e. Direct Enquiries data)
Finite mixture model
(Chapter 3)
Considering additional condition: timetable & individuals' entry times
(Chapter 5)
Individual route choice probability conditional on journey time
(Chapter 3)
Updated route choice probability for each individual
(Chapter 5)
Naive Bayesian modelling framework
(Chapter 3, 4 & 5)
Two sets of
posterior probabilities of each individual's route choice
(Chapter 3 & 5)
Case studies on the London Underground
(Chapter 4 & 5)
Latent route choice model
(Chapter 6)
Two sets of
coefficients of route attributes,
(i.e. travel times of journey segments)
(Chapter 6)
Passengers' sensitivities to the route attributes
(Chapter 6)

Figure 1.1 Methodological framework of the thesis.

The directed lines coloured in **orange** show the structure of the thesis, and the black ones demonstrate the flow of data. From the top of this framework, **Chapter 1**, **Chapter 2** and **Chapter 3** set the context in which the research problem in the thesis is addressed. The problem includes two strands: a naive Bayesian framework for the modelling of passengers' route choices, which is elaborated in **Chapter 3**, **Chapter 4** and **Chapter 5**; and a latent route choice model, which is elaborated in **Chapter 6**.

The processing of a bundle of available data used in this thesis is in parallel with the delivery of **Chapter 3**, **Chapter 4** as well as **Chapter 5**. As part of the naive Bayesian modelling framework, the finite mixture model is elaborated in **Chapter 3**; and its application (including the model estimation, interpretation and validation) is demonstrated in **Chapter 4**, using data from the LU network. Such data includes the Oyster smart-card data, historical route-choice data, and walking time data (for access, egress and interchange) as well as data for layout of the passageways within the underground stations. Then, the estimates of the mixture model is updated in **Chapter 5**, by incorporating additional information.

Finally, the outputs from the naive Bayesian modelling framework serve as inputs for estimation of the latent route choice model expounded in **Chapter 6**. That is, in view of the fact that the actual route choices of passengers are not observed, their route choice probabilities are used as the data for model estimation.

1.4 Outline

Given the context presented in this introductory chapter, the remaining part of this thesis is constructed as follows:

Chapter 2 presents a review of studies on the modelling of passengers' route choice behaviour on underground systems, focusing particularly on the route choice modules that serve as the core for transit assignment models. Different behavioural assumptions on passengers' choice decision-making processes are compared and discussed. The issues relevant to route-choice data and choice modelling are also pointed out.

Chapter 3 provides a completely different viewpoint of representing the route choices of passengers for a given O-D pair. The chapter elaborates on the applicability of a finite mixture model to allow for a probabilistic representation, namely, posterior probabilities, of each individual passenger's route choice, given the observation of his/her journey time.

Chapter 4 demonstrates the application of the mixture models proposed from its precious chapter, **Chapter 3**, with two different types of standard mixtures. The chapter presents a range of case studies based on the LU network, taking advantage of the Oyster smart-card data together with ancillary information available for the LU system (as shown in **Figure 1.1**, p.5). A comparison of the estimation results from two the types of mixture models is presented.

Chapter 5 proposes an approach to update individually each passenger's route-choice probabilities in order to obtain relatively more robust estimates, which is still based on Bayes' theorem. Relying on the estimates from **Chapter 4**, the chapter involves more evidence, that is, the timetable as well as each individual's actual entry time. A comparison of the individual route-choice probabilities before and after the update is presented.

Chapter 6 demonstrates a new approach to the development of a discrete choice model by using the estimated posterior probabilities of passengers' route choices, instead of their actual route choices, which is referred to as a latent choice model. The chapter uses the two sets of posterior estimates, which are derived from the case studies in its previous two chapters, to test the proposed approach separately, by estimating a simple **m**ulti**n**omial **l**ogit (MNL) model. A comparison between the estimation results is presented.

Chapter 7 concludes this thesis with a summary of main limitations of the methodological framework illustrated in **Figure 1.1** (*see* p.5). Furthermore, the chapter also provides a set of recommendations for improving its structure and important avenues for future research (illustrated with **Figure 7.1**, p.177).

1.5 Contributions

This thesis makes worthwhile contributions to the modelling and understanding of the passengers' route choice behaviour within the context of the underground system. They are achieved in four respects as follow.

The work of this thesis

I. establishes a preliminary methodological framework for the modelling and understanding of passengers' route choice behaviours without actual route-choice data;

II. assesses and demonstrates applicability of the finite mixture models for discovering passengers' route choices at both the aggregate and the individual levels;

III. attains initial development of a latent route choice model, which allows for the estimation of discrete choice models without actual route-choice data; and

IV. further explores potentialities of the use of smart-card data on public transport.

Chapter 2
Modelling route choices on public transport

2.1 Introduction

For decades, the modelling and the prediction of passengers' route choices – as well as that of passenger-traffic distribution over public transport network, have long been a challenging subject for transport planners and researchers. Many specialists, especially the modellers, have continuously strived to build and refine various effective platforms for developing more and more efficient mathematical approaches. By far, a wide spectrum of mathematical models for the route choice on the public transport have been established, which are mostly serving as a vital module for tackling transit assignment problems. In that regard, a transit assignment model is devoted to reproduce the passengers' route choice behaviour at each of decision-making points along their journeys, hence their route choices and the traffic between any O-D pairs of a given transit network. Additionally, it may also act as an assessment tool for validation and analyses of operation schemes for the transit system. On that basis, this chapter scrutinises a diverse range of route choice models built in the numerous existing transit assignment models, which are later referred to as route-choice modules, and further explains the homogeneity and heterogeneity of underlying factors and choice behaviour addressed by those models. This is based mainly on the surveys reported by Fu *et al.* (2012b), which identified issues that remain outstanding in gaining a deeper insight into passengers' route choice behaviour. The principal aim of the review is to elucidate the essential aspects of the route-choice decision-making process, so as to lay the foundation for further exploration of solutions to handle the crux of the research problem on how the choice behaviour can be better understood.

On the whole, the transit assignment problem has been well inspected from two distinct standpoints: the frequency-based approach (*e.g.* Chriqui and Robillard, 1975; Nguyen and Pallottino, 1988; Spiess and Florian, 1989; Wu *et al.*, 1994; Cominetti and Correa, 2001; and Cepeda *et al.*, 2006); as well as the schedule-based approach (*e.g.* Tong, 1986; Hickman and Bernstein, 1997; Florian, 1999; Tong and Wong, 1999; Nuzzolo *et al.*, 2001; and Poon *et al.*, 2004).[1] A wealth of researches have been conducted, and provided insights into both methods. Among the earliest comprehensive reviews on the relevant modelling methods were contributed by Bouzaïene-Ayari *et al.* (1998). According to their findings, the function that underlies the route choice models could be summarised in the following three aspects: (a) characteristics of the supply on transit networks and services; (b) information about the supply that passengers could have before and during their journeys; and (c) passengers' responses towards current situations given related travel information. Later, Nuzzolo *et al.* (2003) and Nuzzolo and Crisalli (2004) paid special attentions to the schedule-based transit assignment models and particularly elaborate the differences of the adaptability of schedule-based models to services with low and high frequencies; and the frequency-based models were reviewed in more detail by Schmöcker (2006) and Teklu (2008a). Furthermore, Nuzzolo and Crisalli (2009) extended the predecessor models to a broader scope, taking into account multi-modal transportation networks of both transit and freight services. More recently, Liu *et al.* (2010) inspected plenty of studies on passengers' route choice behaviours, ranging from the conventional deterministic models to various dynamic ones, given *e.g.* the effect of real-time information.

In the context of the above[2], the rest of this chapter is arranged as follows. The basic concepts and definitions of the route-choice modules are described in **Section 2.2**, which lays the foundation for the subsequent sections. **Section 2.3** elaborates in greater detail on passengers' choice behaviour at different stages of their journeys, and also the behavioural assumptions that underlie the module

[1] The frequency-based approaches are also known as headway-based, line-based models, *etc.*; and the schedule-based ones are often referred to as timetable-based, run-based, *etc.*

[2] Note that some of transit assignment models focus particularly on the bus network, and some others are based only on the underground railway network. In practice, though, the terms 'stop' and 'station' could be used interchangeably, the former is often referred to in the bus system while the latter is often referred to in the cases of underground networks. In this thesis, we are investigating only the latter cases, i.e. the underground system; and the two terms will also be exchangeable in the following texts where the term 'station' will be more frequently used. Additionally, it must be noted that a 'station' (and a 'stop') is explicitly distinguished from a 'platform' in this thesis.

building processes. **Section 2.4** discusses the interaction between the route choices of passengers and their journey time variability, as well as the issue about the route-choice data. On the strength of the discussions of various concerns related to the choice behaviours, **Section 2.5** points out the matter of our interest and concludes this chapter.

2.2 The foundation of modelling route choices

2.2.1 Transit network and alternative routes

In order to learn about passengers' route choice on a given transit network, it would entail a mathematical imitation of the choices as well as the passengers' decision-making process starting from their origin stations, with all onward 'journey segments' in sequence, to planned destinations. Consider a passenger is travelling on an underground system. Since the scope of this thesis is confined to the level of transit network rather than the practical O-D[3], in general, a single journey of the passenger between any given pair of O-D stations can be segmented into a series of such journey segments as follows:

- Access: starting from a ticket gate[4] or a ticket hall at the origin station and walking/moving[5] towards a platform for a transit line;
- Waiting: waiting on the platform for departure from the origin platform, until climbing aboard a train;
- Traveling: riding in the train from the current (origin station) platform to another (at the destination station), and getting off-board; and
- Egress: leaving from the destination platform and moving to a gateline, and exit from the destination station.

When passengers have to transfer from one line to another between different platforms, additional journey segments shall then be involved in: (*see next page*)

[3] On the level of the practical O-D, the network of interest may extend to travellers' actual origins and destinations, such as homes, offices and shopping centres.

[4] The location of any ticket gate within a station may also be referred as a 'gateline'.

[5] In this thesis, the terms 'walking' and 'moving' are exchangeable.

- Interchange walking: leaving from the current platform (for a line of a certain direction) and moving to another (for a line of a certain direction), in order to transfer from one line to another;[6]
- Interchange waiting: waiting on the platform for departure from the interchange station, until getting aboard a train on a connecting line.
- Onward travelling: riding in the connecting train from the current platform to another at the destination station, and getting off-board.

Each of these journey segments is associated with a travel cost (or disutility). A passenger's cost of a journey could generally be regarded as a sum – or rather a weighted sum – of the costs for all the journey segments, which is hereinafter termed the journey cost. However, different assumptions made by modellers on the specification of the cost function would bring about different travel cost for each journey segment and hence the journey cost for a travel route.

For modelling purpose, a transit network is described by nodes and directed arcs, with simulated passenger flows being transmitted via the different functional arcs between the nodes that act as decision-making points. On this basis, a sub-network that defines a station is usually taken for the focal issue (*cf.* Bouzaïene-Ayari *et al.*, 2001; and Billi *et al.*, 2004). At each station, passengers will need to choose one of 'attractive lines' and travel to the next stop. The definition of the attractive lines was given by De Cea and Fernández (1993), and it indicates the fact that not all transit lines available at a station/platform would be taken into account by passengers, as they might simply ignore the lines that could conceivably lead to a relatively disadvantageous route. In practice, the attractive lines, which passengers may face and choose from at each of the interchange stations, build different possible routes connecting to their destination station. Given the passengers' perceptions to the journey cost, the passenger flow sourced from an origin station may then split up among the attractive lines and hence among the alternative routes. It must be pointed out that, in effect, the true set of alternatives for route choices cannot be determined

[6] In the case of cross-platform interchange (*i.e.* interchange between lines at an island platform), this journey segment could be ignored, or integrated into the subsequent journey segment as 'interchange waiting'. In this thesis, we assume that each of platforms at a station is served by only one transit line.

accurately in that the reasonableness of any of the alternatives may not be verified (*cf.* Guo, 2008, pp.262-263).

With regard to the schedule-based models, every move of trains and passengers in the transit network is marked with a time-stamp. Thus, these entities can be located, described, and differentiated from each other in both the temporal and the spatial dimension. Representation of the transit network is thereby adapted from the line-based spatial-only graph (*i.e.* without time dimension), which is used for frequency-based models, into a run-based spatiotemporal graph that can show each of a series of runs as scheduled. Therefore, the characteristics of each service run can be taken into account and modelled separately.

2.2.2 Journey cost

The core belief that underlies the outcomes of any route choice models is that a traveller always chooses a 'cost-efficient' route to complete his/her journey. That is, for each passenger, the journey cost of his/her chosen route is supposed to be the minimum or the optimum, in comparison with other alternative routes. A key issue is to properly specify how the cost should be calculated. Such journey cost can be analysed either based on every single route (especially in early models, such as Dial, 1967; Fearnside and Draper, 1971; and le Clercq, 1972), or in the context of a hyperpath (*e.g.* Nguyen and Pallottino, 1988; as well as Spiess and Florian, 1989). In the light of the definition by Nguyen and Pallottino (1988), a hyperpath consists of a set of routes considered simultaneously by a passenger, with each being referred to as an 'elementary path' or an attractive route. It involves a set of sequential decisions of the passenger choosing from among attractive lines at an origin (and every intermediate stop), in order to start (and continue) his/her journey. Taken in this sense, the journey cost of the passenger is effectively treated (by modellers) as a probabilistic cost over a set of attractive routes. On the same basis, Spiess and Florian (1989) termed the series of decisions a strategy whereby the passenger can reach his/her destination subject to route choice probabilities. As multiple transit lines exist, more than one hyperpath can be available and utilised, and so different strategies can be applied by passengers based on their own considerations.

In some cases, the term 'cost' can be merely regarded as the total travel time through the journey, namely, the weighted sum of observations or estimates for the travel time for every journey segment of a route. While in other cases, it can be dealt with as a generalised cost in a synthetic manner, which takes into account not merely travel time but also other stochastic attributes and uncertainties up to the complexities of modelling perspectives of analysts (*e.g.* Szeto *et al.*, 2011; and Szeto *et al.*, 2013). They may include reliability of transit services, crowdedness, discomfort, value of time, seat availability, as well as passengers' perceptions to these issues and so on.

2.2.3 Fundamental behavioural assumptions

As a matter of fact, passengers may not be able to know exactly the true journey cost of each alternative route (or any hyperpath). Instead, they may estimate it, given their own preferred route-choice sets and thus can make trade-off choice decisions. In this context, another major issue with respect to modelling the route choices is a (mathematical) representation of the passengers' decision-making processes, which would have to rely on related behavioural assumptions.

In the real world, the travellers' route choices are essentially the outcomes of their reacting to supply of a transit network. The network supply could relate to attributes of the network as well as the transit services – basically, layouts of the stations, transit lines, operation schemes (*e.g.* timetables), service capacities, as well as provision of both offline and real-time information on the services. By the force of the interplay over time between the travel demand and the supply, the passenger flows merge and split at the start of every journey segment (as defined in **Section 2.2.1**). As a consequence, all available routes of the transit network are loaded with the traffic. Such process could also be referred to the construction of a hyperpath/hyperpaths as well as strategy/strategies (as described in **Section 2.2.2**), which are typically considered by most of the existing, especially the frequency-based, transit assignment models to deal with the passenger traffic distribution.

Moreover, consider that passengers are travelling on an underground network with a high frequency of trains. Under this circumstance, intuitively, it does not seem to concern the majority of the travellers that whether there would be a train available as soon as they arrive at a platform. In other words, a short wait would be supposed to be acceptable for most passengers. In addition, it is commonly assumed that an individual passenger's arrival at a station or a platform is independent of each other; and it is also supposed to be irrespective of any vehicle's arrivals. These two assumptions in many transit assignment models bring about a uniformly random passengers' arrival rate (*e.g.* Spiess and Florian, 1989); and this underlies the classic assumption that passengers always choose to board the firstly arriving vehicle that belongs to their attractive lines set, given a Poisson process of the transit vehicles' arrival. In contrast, when the line service frequency is relatively low, those passengers would be more likely to plan in advance for their access as well as possible interchanges. This is in order to minimise the waiting cost such as the waiting time for a specific train/run of an attractive line.

In general, a transit map may often serve as the most important (or even sole) source of information about the transit network. Practically all travellers would use it as a reference to make route-choice decisions. In that situation, the transit map would tend to have the utmost impact on the passengers' travel strategies (*cf.* Guo, 2011), especially when there is no additional information provided to those who are unfamiliar with the network. Nevertheless, those experienced or frequent travellers, *e.g.* commuters, may have rather fixed route choices among all the alternatives, based on their prior knowledge about the transit system. They may have already made a decision on which route to choose before they arrive at the origin station. On the other hand, real-time information during the course of the passengers' journeys may also influence their choice decisions (*cf.* Hickman and Wilson, 1995). For instance, if the information about waiting times for the next trains of all attractive lines is available prior to the passengers' heading to the platform for a preferred line, some of the them may reckon that their predetermined routes would become less or no longer satisfactory, and thus potentially turn to alternative attractive routes (*cf.* Gentile *et al.*, 2005; and Cats *et al.*, 2011).

In the frequency-based models without considering the common lines problem, the headway can be treated as the time interval between two trains in a row that serve for the same line, with the mean being the average waiting time for that line. While the common lines problem is included, trains on different lines arrive at a platform alternately according to their respective scheduled headways. In this regard, the average inter-arrival time between runs is shorter due to the joint services, namely, a passenger's average waiting time for boarding (an attractive line) is dependent on a combined service frequency of all the attractive lines. What is more, the exponential distribution has been the most common assumption prescribed for the headways; whereas Bouzaïene-Ayari *et al.* (2001) argues it may not be appropriate in the case of reliable service regularity, as extremely irregular headways are not frequently encountered which however might be the case for exponential distribution. The Erlang distribution was later proposed and used for approximating the headways (*e.g.* Bouzaïene-Ayari *et al.*, 2001). For the common lines problem, conventionally, the probability of anyone boarding an attractive line is calculated as the proportion of its service frequency among all alternatives. This implies that the more frequent a line service is, the higher probability that a vehicle of the transit line would be firstly arriving, and the greater chance it could obtain of being chosen. Each alternative route can be assigned a probability of being chosen, even though illogical ones are never used that have zero probabilities.

2.3 Route choice behaviours

Each journey segment has its own service capacity, and offers limited ability to accommodate and manage the flows of passenger traffic. The passengers may often experience congestion (or even overcrowding) when walking within the stations, waiting on the platforms as well as travelling in the trains, especially at rush hour when passenger-traffic reaches a peak. It arises since the network supply of the corresponding journey segment is not able to meet the extra travel demand during a given operational period. Such a traffic situation could be very typical of rush hours, such as the morning and evening peaks, which is in stark contrast to off-peak times dealing mostly with a normal (or even free) flow of passenger traffic. Unlike bus systems where bunched services may be available

In addition, any planned engineering work would cause delay or cancellation of trains; and particularly, unexpected emergencies would also hinders the system from releasing the surges of incoming passenger flows. These incidents may have a major impact on passengers' journey cost, and hence their travel behaviours. On the basis of the foundation laid by the previous section, this section provides deeper insight about the passengers' possible route-choice behaviours throughout the passengers' journeys.

2.3.1 Moving through passageways

The service capacity for passengers moving within any underground station shall involve in all types of pedestrian facilities inside the station. Generally, such facilities include ticket halls (or concourses), level/ramp passages, pedestrian conveyors, escalators, lifts and staircases.[7] A number of different types of such passageways together construct a pedestrian pathway for passengers' access from gatelines to platforms, transfer between platforms, as well as egress from platforms to gatelines (*cf.* **Section 2.2.1**). Thus, the measure of the capacity of the pedestrian service would largely depend on the attributes of these passageways, including *e.g.* total numbers, lengths, rises/runs and layout, which are closely related to the pedestrian passenger flows.

Hankin and Wright (1958) were among the first to carry out experiments concerning the within-station pedestrian traffic flows of passengers. They investigated the relationships between the pedestrian speed, flow of passengers and the capacity of passageways (including both level passages and staircases) for the LU stations. According to their studies, the pedestrian flow within a station was measured by the number of passengers per foot width per minute, while the speed was calculated as the time of their movement over a certain length; and both were based on a given pre-measured area. It was also illustrated in their analysis results that crowdedness would slow down passengers' walking speeds. Daly *et al.* (1991) illustrated the findings on the relationships between flow and walking time for each passageways within station that the speed-flow relationship was similar to that of the road traffic conditions. In addition, they all presented their experiment results about walking speeds on different types of passageways, on the conditions of free passenger flow and when the facility capacity was reached, *etc*. Those important conclusions drawn from their experiments were also confirmed in relevant studies, conducted by *e.g.* Harris (1991); Cheung (1998) and Lam and Cheung (2000). Furthermore, Lam and

[7] We use the term 'passageway' as a generic term. Note that in some other studies, *e.g.* Daly *et al.* (1991), the platforms and intersection area of different passageways were also were also examined, however, which are not considered as the passageways in the current context.

Cheung (2000)[8] derived and calibrated the travel time function for each type of passageways with the data collected on the Hong Kong metro system, and compared the average speeds with the findings on the LU.

Moreover, not only is it the disutility of crowdedness and walking speed/time that may be considered by passengers, but they may also have different tastes in walking distance. At the origin station, the pathways with shorter walking distance to the platforms for attractive lines might usually be more preferable to passengers, especially those who are commuters, older or disabled people. This factor may potentially dominate their route choices, in the absence of real-time information around the gateline area (*i.e.* at the start of access). Note, however, that some of the travellers with limited walking ability may need assistance of lifts (and might also tend to avoid the crowd). Such facilities may or may not necessarily be on the shorter (or the shortest) pathways.

Besides, the interchange (including the platform-to-platform walking and waiting on the platform) is considered particularly sensitive to passengers, as it might be deemed to cause an 'interruption' in one's single journey. Regarding the journey cost specification, usually, the extra disutility would be associated with both the interchange walking and waiting, which can be termed a 'transfer penalty' (*e.g.* Guo, 2008). Moreover, at different stations, passengers suffer different levels of transfer penalties. Surveys and behavioural modelling are the two main methods to understand the transfer behaviours.

2.3.2 Waiting and failures of boarding

As for the passenger traffic gathered on the platforms, whether and when the passengers would be able to board a train is another one of the key issues for the formulation of a route-choice module for any transit assignment models. This is particularly significant for modelling the rush-hour traffic, due to the fact that the limited loading capacity of trains/carriages imposes restrictions on extra boarding demand. A portion of the passengers waiting on platforms may fail to get aboard (after one or several attempts). Such boarding failure(s) prolongs a

[8] The average walking speeds about the LU presented in this study will be later used as reference materials in **Chapter 4**.

passenger's waiting (hence their waiting time), and might significantly increase the possibility of a longer total journey time.

A common assumption for passengers' boarding is that all the passengers would choose to board the firstly arriving train among attractive lines, if additional information (*e.g.* remaining waiting time for a certain service) is not available. In uncongested situations, all the wait-to-board passengers are assumed to be always able to get aboard a train on an attractive line, since the capacity of carriages is treated as unlimited. However, with respect to models for congested transit networks, the train/carriage capacity should be strictly constrained, and that the situation that passengers may fail to board is explicitly considered. Under such circumstances, there could be continuously accumulated volume of traffic as those fail-to-board passengers would be still waiting, which aggravates the crowding on the platform and thereby affects the service that follows at a subsequent scheduled time interval. The increasingly intensive congestion may maintain during peak period.

In reality, the fail-to-board passengers could be generally classified into two groups. One group includes those who do intend but are not able to climb aboard the train, due to limited standing space in trains; whereas the other group contains passengers who actually decline or are not willing to board. Basically, these two groups could be referred to two situations, respectively, as follows: (a) the train capacity has been completely fulfilled and the carriages cannot accommodate all the wait-to-board passengers; and (b) at the same time, some of the wait-to-board passengers are sensitive to congestion in a train (and/or chances of having a seat) and hence give up the chance to board, despite availability of standing room. Consequently, at least a headway is added to the waiting time that each of those fail-to-board passengers spend in both situations.

Besides, it is arguable that every wait-to-board passenger on the platform may have the same chance of boarding. This statement could be reasonable, but mainly in uncongested conditions, as passengers are more likely to mingle and those who arrive later could wait by any carriage door. Nonetheless, the first-come-first-serve rule should be more appropriate given that passengers queue by each door of the carriages, especially in congested situation.

For the waiting time specification, early models, such as De Cea and Fernández (1993), considered it to be monotonically increasing as the passenger volume increases, which was then specified as a congestion cost function relating to the notion of effective frequency (*e.g.* Wu *et al.*, 1994; and Cominetti and Correa, 2001). An effective frequency was used to characterise an attractive line or common lines. That is, if the passengers' chances of encountering a full train rise, the effective frequency of the relevant service should decrease, and thus the waiting time for that train shall become longer. However, the congestion cost function does not actually restrain train capacities from being overloaded by excess travel demand. Later models then (*e.g.* Lam *et al.*, 1999; Nguyen *et al.*, 2001; Lam *et al.*, 2002; Hamdouch *et al.*, 2004; Yin *et al.*, 2004; and Hamdouch and Lawphongpanich, 2008; Teklu, 2008b) specified explicit constraints to impose restrictions on the excess passenger-traffic flows being assigned onto any route sections with limited capacities.

Moreover, special attention to the probability of failing-to-board that affects the search for the shortest hyperpath was paid by researchers such as Kurauchi *et al.* (2003), Schmöcker *et al.* (2008) as well as Schmöcker and Bell (2009). The choice set of lines considered by passengers who fail to board may change in different time intervals, and it depends only on the current condition, which is known as Markov property and also discussed by Teklu *et al.* (2007). Fail-to-board passengers who keep waiting on the same platform obey with the Markov properties. Whether or not they would be able to board a train that has currently arrived is not related to where they started their journey or how long they have been waiting. The boarding and alighting demand at the current platform are necessary, which also requires the knowledge of the traffic volumes at the upstream stops each associates with a timestamp. Consequently, the waiting time at a given platform depends on the variations of traffic volume over time or time intervals, and the passengers in the train and that on the platform would practically have a longer the waiting time.

Another issue that may also impede passengers' boarding is seat availability, which may influences passengers' travel strategies and can be taken into account only in less- or un-congested circumstances; whereas, this is not the case when the network is suffering from high congestion during periods of peak demand of rush-hour traffic. Because in highly congested conditions whether a passenger could be seated on-board would not be the main concern. Instead, whether there is a chance for passengers to get aboard would be valued, given that the vehicle still has capacity of extra boarding demand and that the on-board crowding does not outstrip passengers' tolerance limits to the congestion.

2.3.3 Travelling and on-board crowdedness

If a chosen line remains crowded for several trains (*i.e.* runs), some fail-to-board passengers would rather keep waiting for a following train, notwithstanding an extra waiting time. A less crowded train/line may be more attractive to some passengers, even though it tends to give rise to a longer total journey time compared to its alternatives (*cf.* Leurent, 2010). That is to say, the passengers' perceptions to their on-board travel (or their perceived on-board travel cost) may not be as bad as the actual travel cost. In this regard, the passengers' aversion to congestion or overcrowding is involved in modelling their choice decisions.

Furthermore, passengers who stand and those who are seated on board may experience different levels of travel discomforts (*cf.* Tian *et al.*, 2007; Sumalee *et al.*, 2009; Leurent, 2010; Hamdouch *et al.*, 2011; and Schmöcker *et al.*, 2011). As such, the fact that some passengers are sensitive to seats would also lead to different specifications of the on-board travelling cost, thus affecting the formation of passengers' travel strategies. While a train is crowded, the discomfort level is assumed to be much higher for the standing-on-board passengers compared to the seated ones. It may be assumed that passengers being seated would be less influenced by the on-board crowdedness. In other words, they would be likely to have similar level of discomfort as being travelling under less congested (or even uncongested) conditions. On the other hand, the degrees of the seat-sensitive passengers' incentives of pursuing vacant seats would differ, which can hardly be quantified. Before passengers board a train, the key influencing factors may involve the total journey distance as well as the seat occupancy. For the passengers who have been standing and travelling on-board, the elapsed time of standing and the remaining distance for their journeys would be likely to become more predominant.

What is more, suppose that a train is approaching or has already arrived at a platform at either an origin or an interchange station. Passengers who have been waiting on the platform will start boarding as soon as the on-board passengers who intend to alight are all cleared. The wait-to-board passengers may estimate how much boarding capacity there could be available by then. Those who are seat-sensitive may also consider whether there is a chance of being seated and/or even calculate the chance of obtaining a seat at a subsequent station. Thus, their decisions will be made as to whether to board or still keep waiting on the current platform for next coming trains. At the same time, the standing-on-board passengers would also decide whether to alight and transfer at the current station, or travel to any of the following alternative interchange stations. In practice, decisions on whether to board the arriving train of an attractive line, keep waiting for the next run of the same line, or transfer to any of alternative services (or even transport modes), would largely be dependent upon what information (especially, the real-time information) and where/when such information would be provided in the passenger's decision-making process (*cf.* Nökel and Wekeck, 2009).

2.4 Discussions

2.4.1 Route choice and journey time variability

Evidently, from the above, the passengers do not necessarily make a journey by the shortest/fastest routes or with least interchanges in order to obtain a maximum savings on their journey times. In some situations, an alternative route may be more attractive and preferred by different individuals for various reasons. Still, as also mentioned above, the passengers' journey cost may be just referred to their total journey time; and in practice, the journey time variability is often considered to weigh up the reliability of the transit service. It can also exert effect on passengers' travel strategies based on their different perceptions to the system performance.

Unlike car traffic on road networks, the underground trains run on fixed tracks and are each associated with a timetable. Ideally, timestamps of arrival and departure of trains at a platform are strictly scheduled. Passengers' on-board travel time could be expected as ascertainable, conditional on the presence of punctuality of the trains running on a passenger's chosen path. Depending on information of the passenger's access and egress, his/her journey time could thus be well predictable, provided the absence of any incidents. However, for the most part, this may not be the case in practice, given varied attributes of the transit network as well as uncertainties, which potentially affect passengers' journey times, such as over-crowding, delay of transit services. While the level of service degrades in view of their comfort and/or preferences, their choice behaviour would be subject to a higher degree of riskiness of having an uncertain journey time. In particular, passenger-traffic congestion occurs frequently with surging travel demand, not only on-board, but also on platforms and passageways within transit stations, especially during rush hours. It can have significant impacts on service regularity as well as reliability, which in turn influence passengers' travel behaviour. Also insufficient service capacity (*e.g.* vehicle capacity or seat availability) may cause passengers' boarding failures, thereby delaying their journey times. Moreover, when a train breaks down and/or is suspended at a certain platform (say due to train system fault), an on-board passenger could possibly choose to keep waiting, interchange to any alternative line serving the same station/platform, or even egress and go for any other modes. Nevertheless, if there were not any alternative service available, the passenger would have to wait until the fault is cleared, or transfer to the next coming train.

On station-to-station level, every passenger has his/her own expected journey time from the origin to the destination, and the range of this expectation and itself depends on various travel information the passenger could obtain. Meanwhile, they may value much on the reliability of the services between which they are going to choose, in association with the variations of journey times that they may experience on their chosen paths. In high frequency service, a delay of a few seconds may result in a series of delays of the runs that follow, which in turn leads to reallocation of passenger distributions.

What's more, travel patterns vary potentially due to the travellers' responses towards the reliability of the transit service, especially as for those commuters who could gain experience of the network performance in terms of day-to-day variations in their travel times. Therefore, a good understanding of such different travel patterns under various backgrounds is essential to the efficient public transport planning, operations and travel demand management.

The journey time variability is measured based on many factors, such as individuals' preferred choices of departure times in view of their desired arrival times and the deviations between expected and actual journey times. The actual journey time experienced by a passenger may be very much different from that was expected or desired, with the average difference being concerned by both the passenger as well as transport operators. **T**ransport **f**or **L**ondon (TfL) defines the excess journey time as "the average time added to journeys by delays, crowding and queuing, over and above the nominal scheduled journey time" (Transport for London, 2010). It could be also drawn to a wider extent on considering that the passenger completes his/her journey faster than expected or desired. In the latter case, a redundant amount of time is unnecessarily budgeted, which is supposed to be minimised. And this is also the excess journey time defined in the former case. In practice, this extra budgeted time is observed from the departure time actually chosen by passengers who may allocate a considerable amount of 'buffer' time in order to flatten the journey time variability caused by any uncertainties. It was defined by Uniman (2009) as 'reliability buffer time', namely, the difference between the observed travel times of the 95th-percentile and the median for an given O-D over certain period of time under normal conditions (Uniman *et al.*, 2010).

2.4.2 Data for route choices

To gain an understanding of passengers' route choice behaviour, the data for their actual choices is vital. At an aggregate level, the passengers' average route choices – or rather, average proportions of the entire passenger-traffic flowing over the multi-route O-D – among the alternative routes could be estimated. We can gain this knowledge via random sampling of a group of individuals, from whose actual route choices a statistical result could be generalised to the overall passenger population. Usually, such data is acquired through surveys, such as online and paper-based questionnaires, as well as interviews. Besides, we may also draw support from mobile technology, such as **g**lobal **p**ositioning **s**ystem (GPS) devices and take full advantage of a range of applications developed for smartphones. Such tracking techniques are expected to effectively save on the time and cost of the traditional survey approaches, as well as to improve the scale and accuracy of the raw data.

In practice, however, all the above-mentioned methods can still be quite expensive and time-consuming to attain sizable, representative data samples. For one thing, many of the travellers who receive the information requests might be unavailable for participation in the surveys. For another, most people may express concerns about issues such as their privacy and the security of the real-time personal data residing online, hence less willingness to be involved with passive monitoring programmes using any of the 'privacy-surrendering' means. Moreover, some technical restrictions would just directly prevent passengers from using the mobile/wireless devices. For instance, the absence of mobile coverage on the LU network renders the passengers being unable to use their mobile phones.[9] Under these circumstances, the sample size of the data collected may tend to be limited. Otherwise, acquisition of an adequate data sample might necessitate a high cost of carrying out numerous repeated surveys.

With the widespread implementations of the AFC systems in the past decade, the ever-increasing popularity of smart cards among public-transport users enables a wealth of individual travel data to be conveniently available (*cf.* Pelletier *et al.*, 2011), which is also referred to as smart-card data. This has drastically reduced the need and expenditure for conducting the manual surveys, but also extended our ability to gather miscellaneous travel information of passengers (*cf.* Bagchi and White, 2005).

When travelling by underground rail services, smartcard users are required to touch their smart cards on card readers at the start and end stations respectively of their journeys, in order for fares to be properly deducted. In addition, when a passenger enters and exists a station by using a smart card, a card reader at a ticket gate processes the information of locations and timestamps at which the passenger's entry and exit occurs. Therewith the smart-card system holds vast quantities of journey records for all its anonymous users travelling within the system, such as ticket types and fare purchases, as well as total amounts of entries/exits at each station. What is more, a sufficiently large sample of those smart-card users' journey times can also be easily obtained from calculating the differences of their time-stamped 'touch-in-entry' (at the origin station) and 'touch-out-exit' (at the destination station).

[9] By far, mobile phone connectivity is not available on the LU network. Although the Wireless Fidelity (Wi-Fi) signal has been provided lately to about half of all the LU stations, it is not or has only been conditionally free for use.

So far, there have been many studies exploring the potentials of using the smart-card data. Pelletier *et al.* (2011) conducted a comprehensive overview of various aspects of its applications to public transport systems. And meanwhile, specific examples have been demonstrated on a number of different AFC systems all around the world, such as the Chicago Card and Chicago Card Plus (replaced by Ventra Card since July 2014) in Chicago (*e.g.* Utsunomiya *et al.*, 2006; and Zhao *et al.*, 2007); the EZ-Link card in Singapore (*e.g.* Chakirov and Erath, 2011; and Lee *et al.*, 2012); the Oyster smart-card in London (*e.g.* Chan, 2007; Zhao *et al.*, 2007; Wilson *et al.*, 2009; Uniman *et al.*, 2010; and Kurauchi *et al.*, 2012); the Passe-Partout PLUS in Gatineau (*e.g.* Morency *et al.*, 2007; and Trépanier *et al.*, 2009); and the 'Tarjeta Bip!' in Santiago (*e.g.* Munizaga and Palma, 2012); and the T-money Card in Seoul (*e.g.* Park *et al.*, 2008; and Jang, 2010); to name but a few. Among all the above-mentioned, focuses were centred mostly on estimation of the O-D travel demand matrix, metrics for transit service and journey time reliability, as well as interchange patterns of passengers transferring between different transit modes (*e.g.* between the underground and buses). In the cases of EZ-Link (*e.g.* Chakirov and Erath, 2011), as well as the T-money in Seoul (*e.g.* Jang, 2010), fare is charged on a distance basis. In that way, the transfer data for passengers using multiple modes is readily available. However, those researches investigated only the O-D pairs that are connected by a single/direct line or a single route.

An important issue that has been rarely addressed by the existing literatures is that the possible interchanges passengers may make during their journeys. This is mainly due to the unavailability of the data. The smart-card scheme for the underground system has practical limitations on data completeness. Generally, it does not allow for either tracking the passengers' movements during the course of their journeys, or recording their within-station interchanges, if any. Although the total of entries is counted at the gateline, the entry-flow splits towards different passageways. The passenger-traffic on any of the passageways is not available in the smart-card data. Likewise, the count of exits does not inform that from which lines the passengers travelled. Moreover, when a certain pair of O-D stations afford multiple alternative routes, a passenger's journey history recorded by the smart-card system would not give details on which specific route he/she has actually used for travelling between the O-D. In this situation, the smart-card data could not inform each individual passengers' route choice.

It is also arguable that such information could actually be gathered by drawing support from other technological means in addition to the smart-card scheme. In practice, however, it would entail levering in extra investment into the system, with respect to infrastructure, equipment, maintenance, as well as the delivery of those technologies. Even so, passengers' socio-demographic characteristics may still not be captured, but additional data from traditional surveys or other information systems should be necessary to supplement and boost both practical and theoretical research (*cf.* White *et al.*, 2010).

By and large, the shortage of sufficient and reliable data presents a major obstacle to further progress in studying the patterns of passenger-traffic flow as well as the passengers' travel behaviour.

2.5 Summary and conclusions

This chapter has presented various aspects of the behavioural processes of passengers' route choice on an underground system, with the emphases being placed on their choice behaviours at different journey segments.

As elaborated above, a route choice model simulates passengers' responses to different network attributes regarding the implementation of transit assignment models. Technically, it generates, at each decision point for passengers boarding and/or transferring, either deterministic or stochastic choices as to how the passengers would complete the rest of their journeys. The deterministic method is popular among frequency-based transit assignment models that are usually built on the formulation of the 'shortest hyperpath' (Nguyen and Pallottino, 1988) or, equivalently, the 'optimal strategy' (Spiess and Florian, 1989). Typical models consider passengers' choice probability of taking a line/train to be the ratio of its service frequency to the combined frequency of all viable alternatives. This is a fixed measure by which the passenger flow is apportioned on uncongested networks. As for more sophisticated stochastic cases[10], such as Lam *et al.* (1999), Nielsen (2000) and Sumalee *et al.* (2009), additional variables are taken into account such that the assignment of passenger-traffic has been further modelled with uncertainties, including level of discomfort, on-board crowding, seat availability and service reliability, *etc.*, which might influence passengers' perceptions in their route choices. However, these types of models are mostly looking at the route choices at an aggregate level.

It has been pointed out that different individual passengers have different perceptions as to how the transit system works, hence different sensitivities to the performance of the transit service, which in turn lead to their different route choices. It will be of importance and more interest to us to gain a better understanding of why one of alternative routes would be chosen by individual passengers and how they would react to changes of different attributes (about *e.g.* the walking, waiting, service reliability) of the transit network. As Nökel and Wekeck (2009) pointed out, there could hardly be a bundle of behavioural assumptions that perfectly represents the passengers' route choice behaviours. Therefore, a random utility model should be more suitable for the representation of the disutility that different passengers would have for different alternative routes. Micro-simulation approaches and/or discrete choice models[11] shall be necessary to accommodate such personal features that vary among individuals.

Although a variety of discrete choice models have been also studied by looking into contributing factors, to implement transit assignment models, the coefficients to those attributes were either simulated/calibrated (*e.g.* Nielsen, 2000) or estimated relying on survey data (*e.g.* Cats, 2011; and Cats *et al.*, 2011). Therefore, the main limitation of this technique is still the data availability. This thus leads to the crux of our problem, that is, data that shows each individual passenger's actual route choice is not easily accessible, which however may always be collected from surveys.

To overcome these issues, we explore new solutions in the following chapters. The fundamentals we have gained from the review of modelling passengers'

[10] Note that the modelling approaches based on the hyperpaths/strategies also serve as a way of describing stochastic route choices.

[11] The discrete choice modelling approaches are discussed in more detail in **Chapter 6**.

route choices on public transport, along with all the principles being presented, will greatly contribute to our understanding of how the choice-making process is like and how we may consider route choice models work. In addition, this would also provide guidance for us to derive the expected journey time of a travel route in the following chapter.

Chapter 3
Bayesian inference of probabilistic route choices

3.1 Introduction

As stated from **Section 2.4.2**, the modelling of passengers' route choice behaviour at the individual level would be largely subject to the availability of route-choice data gathered for each individual passenger. Rather than solely relying on the traditional survey methods for the collection of such data, this chapter offers a completely different standpoint of representing and learning the individuals' route choices for any given pair of multi-route O-D. That is, in this chapter, we aim to explore possibility of gaining knowledge about the route choices of passengers from interrogating data held in some databases[1] that we already have or are easily obtainable, especially the smart-card data.

The smart-card database for the underground system, as we mentioned earlier, is capable of supplying abundant data samples of individual passengers' journey times on a gateline-to-gateline basis and across all operating periods. In the case that a pair of O-D stations is served by a single transit line, where there could be one sensible route, the journey time information extracted from the smart-card data can be an ideal aid for examining the service performance of this O-D (*cf.* **Section 2.4.1**). As a significant measure of the level of service, the journey time variability, in turn, should then also characterise the only route between that O-D. With regard to other cases where alternative routes exist, the passengers' journey times would be directly affected by attributes of the particular route on which they choose to travel, such as the timetable, service delay and pedestrian facilities within stations. For the most part, such attributes shall differ between those alternative routes (*cf.* Train, 2009, p.21). In this regard, any sample data of passengers' journey times of a certain route would presumably exhibit a pattern,

[1] In addition to the smart-card data, some ancillary sources of information are also needed; and they are specified at appropriate stages in the subsequent sections of this chapter. The real data for practical use (on the LU system) is detailed in **Chapter 4**.

which differentiates that route from its alternatives. This point of view encourages us to contemplate potential use of the journey time data, in an attempt to understand the passengers' route choices, despite the absence of this route-choice information in the observed records.

Now consider a pair of O-D stations connected by more than one transit line, where we are fully aware of the existence of all the available alternative routes. Suppose that we have already managed to get hold of a huge data sample of individual passengers' journey times from smart-card records, with none detailing their travel routes. The following two research questions could be brought forward:

(I) Would we be able to find a way of relating a passenger's journey time from the available data to the 'unknown' route that has been actually used by the passenger, and/or what relationship might there potentially be between the two classes?

(II) Would it be possible for us to find out every individual's actual route choice according to his/her journey time (or by whatever means is appropriate)?

As stated in **Chapter 1**, the actual chosen route of each individual in this context could only be treated as being unobservable, but may still be known up to a choice probability. On this account, we shall have to take a theoretical consideration of the individual passengers' probabilistic route choices in order to address question (I) posed above. That is to say, everyone's route choice is turned into a probabilistic variable, which must therefore be investigated and learnt in a probability space. As such, we would seek only the choice probabilities that a passenger might have placed on each of the alternative routes.

As aforementioned, an observation that a passenger has spent a certain amount of journey time must be rooted in some attributes peculiar to his/her only chosen route. However, if we are given only the observed journey time, there can be multiple hypotheses in respect of the passenger's route-choice decision in the real world. Being inspired by the concept of Bayesian networks (Heckerman, 1997), we may anticipate, from a Bayesian perspective, a logical and causal connection between the passenger's journey time and his/her route choice in terms of a conditional choice probability. Such a probabilistic term is supposed to describe (and measure) how likely a passenger might have chosen any alternative route, on condition that we have already known his/her journey time. On this basis, we can then proceed to challenge whether or not it could afford an affirmative answer to question (I). Further, in response to question (II), an individual's route choice seems impossible to be explicitly identified in this probabilistic setting. Still, it would possibly be understood from statistical inference. An obvious initial inference to make might be that: a passenger must have chosen the route that is estimated as having the highest such choice probability among all the alternatives.

It is also worth noting that, with the choice probabilities for all the individual passengers, the average probability of any route being chosen can then be estimated accordingly within the passenger population. Nevertheless, to what extent we could draw such kind of conclusions, the focal issue will depend on whether and/or how we would be able to work out those conditional choice probabilities.

Overall, this chapter aims at building on Bayesian framework with an approach to finding out, on any given pair of multi-route O-D stations, each individual's probabilistic route choices, as well as passenger-flow distribution among the different alternative routes. The smart-card data records, from which samples of passengers' journey times, would serve as the prerequisite for the estimation of their route-choice probabilities.[2] Much of the work that had been accomplished by Fu (2012a) and Fu *et al.* (2014) paves the way for this whole chapter that contributes a refined, and greater, elaboration.

The rest of this chapter is arranged as follows. **Section 3.2** gives a detailed description of the probability space wherein the problem of passengers' probabilistic route choices is defined. In the subsequent sections, a possible solution to this problem is provided, with a probe into the finite mixture model. **Section 3.3** presents the formulation, data input as well as estimation method of the suggested model. In **Section 3.4**, a set of validation criteria are proposed in order to understand the model estimates in terms of the hidden variables of the route choices. Then, **Section 3.5** illuminates the use of the estimates of each individual's route-choice probabilities to infer the passenger-traffic distribution

[2] The passengers referred hereinafter are all assumed to be smart-card users.

among different alternative routes. A summary of limitations of this modelling approach is presented in **Section 3.6** for a conclusion of this chapter.

3.2 Problem description

The following notation listed below is used for facilitating a mathematical formulation of the probabilistic route-choice problem at issue.

Notation:

o	origin station of a give O-D pair
d	destination station of a give O-D pair
$\mathbb{N}_{\geq 2}$	set of the natural numbers that are greater than or equal to 2
r	travel route
N_R	total number of travel routes (connecting o to d); $N_R \in \mathbb{N}_{\geq 2}$
R	set of all alternative travel routes (connecting o to d)
$\mathbb{N}_{\geq 1}$	set of the natural numbers that are greater than or equal to 1
q	individual passenger (travelling from o to d)
N_Q	total number of passengers (travelling from o to d); $N_Q \in \mathbb{N}_{\geq 1}$
Q	statistical population of passengers (travelling from o to d)
R_q	personal route-choice set of q
$\langle q, r \rangle$	possible outcome that q has chosen r to make a single journey
Φ_q	set of all possibilities of q
C_q	set of all elementary events within the sigma-field given Φ_q
$choice_{qr}$	event that q chose r to make a single journey
$\Pr(\cdot)$	probability measure
δ_q	journey time of q
δ_q^{OBS}	journey time **obs**ervation ($\texttt{OBS}$) of q
$\langle q, r, \delta_q^{\text{OBS}} \rangle$	possible outcome that q has chosen r to make a single journey, with a journey time of δ_q^{OBS}
$\Phi_q^{(\delta)}$	set of all possible route choices of q, given δ_q^{OBS}

(*Continued*)

Notation: (*Continued.*)

$C_q^{(\delta)}$	set of all elementary events of route choices of q, given δ_q^{OBS}
$choice_{qr}^{(\delta)}$	event that q chose r to make a single journey and spent a journey time of δ_q^{OBS}
$\boldsymbol{\delta}_q$	elementary event that q spent a journey time of δ_q^{OBS}
$\Phi^{(\delta)}$	set of all possible route choices of Q , given δ_q^{OBS} $\forall q$
$C^{(\delta)}$	set of all elementary events of route choices of Q, given δ_q^{OBS} $\forall q$

Now, let us take a look at a simple underground network, which is outlined in **Figure 3.1** below.

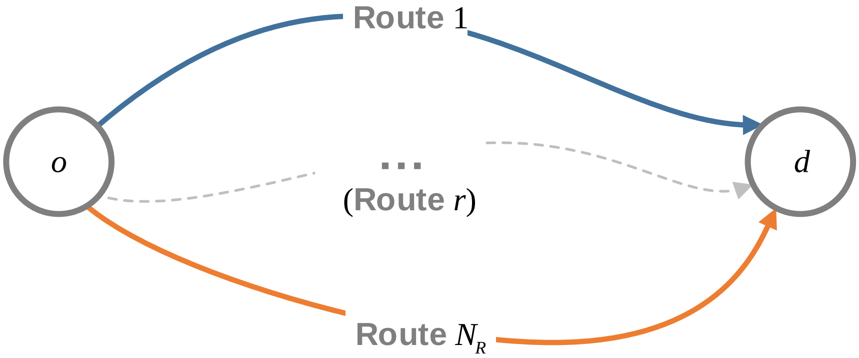


Figure 3.1 A single O-D network with multiple travel routes.

Basically, our focus here is only on a single pair of underground stations of origin and destination, denoted by o and d , respectively. As can be seen from the sketch above, there are supposed to be a total of N_R travel routes, where $N_R \in \mathbb{N}_{\geq 2}$, with $\mathbb{N}_{\geq 2}$ denoting a set of the natural numbers that are greater than or equal to 2. All these routes are deemed rational and numbered arbitrarily from 1 to N_R, which collectively form a finite set of alternatives available for every passenger travelling from o to d. We let R denote this universal route-choice set and define it as $R := \{r: \ r = 1, \ldots, N_R\}$, with r denoting a travel route.

Suppose that the overall passenger traffic (*i.e.* the travel demand) between this o-d amounts to N_Q in a certain time period, where N_Q is a positive integer (*i.e.* $N_Q \in \mathbb{N}_{\geq 1}$). We use the symbol q to denote an individual passenger, and number all the individuals arbitrarily from 1 through N_Q. Denote by Q the statistical population of passengers; and we define Q to consist of all the N_Q individuals by setting $Q := \{q:\ q = 1, \ldots, N_Q\}$.

It is highly likely that prior to commencing a journey every passenger, say q, would customise R and have his/her own personal route-choice set (*cf.* Ben-Akiva and Boccara, 1995). Let R_q denote the customised route-choice set of q. Apparently, it can be any of the non-empty subsets of all those alternative routes. Namely, $R_q \in 2^R \setminus \{\varnothing\}\ \ \forall q \in Q$ and $|R_q| \le N_R$, where 2^R is the power set of R. As such, R_q may also refer to a set of hyperpaths, from which passenger q chooses the optimal (*cf.* Schmöcker *et al.*, 2013). Yet, such individualised choice set can hardly be fully understood or predicted by anyone except passenger q himself/herself. Hence, within the scope of this chapter (and also the thesis), we postulate that

$$R_q \equiv R. \tag{3-1}$$

It is presumed by this identity that every individual would be taking into account the full set of available alternatives whilst making his/her route-choice decision.

Looking at a real underground network (*e.g.* the LU), mostly, there are actually limited sensible routes for each O-D pair. In view of this fact, the presumption of identity (3-1) would plausibly be the case, especially for non-commuters (*e.g.* tourists) who are not familiar with the transit services on the network. However, a commuter passenger may regularly take the route that he/she is accustomed to, and would barely use other alternatives unless necessary (for instance, as disruptions occur on the frequently-used route). In this regard, there would be a risk that identity (3-1) might be inappropriate, because in reality R_q might be merely a unit set, particularly if q is a frequent traveller between the given O-D. To some extent, such risk could be diminished as we contextualise passengers' route choices probabilistically.

Since every passenger's actual route choice is not known to us, we may say that a passenger might choose any one of the alternative routes. As a presupposition, identity (3-1) then allows us to enumerate a set of all possible outcomes of the route-choice decision made by the passenger. For each q, we let Φ_q denote the set of all his/her possible route choices; and it is defined as

$$\Phi_q := \{\langle q, r\rangle :\ r \in R\}, \tag{3-2}$$

where the 2-tuple, $\langle q, r\rangle$, is defined to be a possible outcome that passenger q has chosen route r from R. Clearly, $|\Phi_q| = N_R$. This equality implies that any of the N_R alternative routes might have been actually chosen by q.

In view of definition (3-2), a sigma-field over Φ_q could simply be defined up to 2^{Φ_q}, which includes all events that might potentially be drawn to our attention.[3] With regard to each of the individual passengers, of particular interest to us in the practice is a set of elementary events, which we represent by a symbol, C_q.[4] It is actually a subset of 2^{Φ_q}, defined as

$$C_q := \left\{choice_{qr} : \ r \in R\right\}, \tag{3-3}$$

where we let $choice_{qr}$ denote an elementary event corresponding to a possible outcome, that is, $choice_{qr} := \{\langle q, r\rangle\}$ $\forall q \in Q$, $\forall r \in R$. As such, the occurrence of $choice_{qr}$ should be described by a probability function, which we represent by $\Pr(choice_{qr})$; and of course, $0 \le \Pr(choice_{qr}) \le 1$.

Since a passenger chooses only one route, the following condition must hold:

$$\Pr(\bigcup_{C_q} choice_{qr}) = \sum_{r \in R} \Pr(choice_{qr}) = 1. \tag{3-4}$$

Moreover, we use the symbol δ_q to represent the journey time of q travelling from o to d, with δ_q^{OBS} denoting the corresponding real-valued observation. Assuming that q has made only this one single journey, as urged by question (I) posed in **Section 3.1** (*see* p.32), we shall further consider a 3-tuple, which is expressed in the form of $\langle\langle q, r\rangle, \delta_q^{\text{OBS}}\rangle$ or $\langle q, r, \delta_q^{\text{OBS}}\rangle$.[5] This is thus defined to be a possible outcome that passenger q has chosen route r, and that he/she has spent a journey time of δ_q^{OBS} to complete the journey.

[3] The definition of the sigma-field, 2^{Φ_q}, at this point, is to ensure the completeness of the definition and generality of the probability space under discussion. It does not affect the following descriptions in this thesis.

[4] Besides C_q, in some cases, researchers may also be interested in joint probabilities of two or more (elementary) events occurring at the same time, such as when a certain number of passengers, as experiment participants, are travelling together between a given O-D pair.

[5] $\langle\langle q, r\rangle, \delta_q^{\text{OBS}}\rangle = \langle q, r, \delta_q^{\text{OBS}}\rangle$.

Let $\langle q, r, \delta_q^{\text{OBS}} \rangle$ substitute for $\langle q, r \rangle$ of Φ_q. In this way, we obtain a parallel set of all possible outcomes of the route choice made by q, which we represent by the symbol $\Phi_q^{(\delta)}$. It is defined as follows:

$$\Phi_q^{(\delta)} := \left\{ \langle q, r, \delta_q^{\text{OBS}} \rangle : \; r \in R \right\}. \tag{3-5}$$

Likewise, a parallel event set is formed as well, denoted by $C_q^{(\delta)}$, concerning q's route choices with an actual observation of his/her journey time. We define it by setting

$$C_q^{(\delta)} := \left\{ choice_{qr}^{(\delta)} : \; r \in R \right\}, \tag{3-6}$$

where $choice_{qr}^{(\delta)}$ is defined to be an elementary event corresponding to a single possible outcome included in $\Phi_q^{(\delta)}$, that is, $choice_{qr}^{(\delta)} := \{\langle q, r, \delta_q^{\text{OBS}} \rangle\}$.

Given $C_q^{(\delta)}$, we also have

$$\sum_{r \in R} \Pr(choice_{qr}^{(\delta)}) = 1. \tag{3-7}$$

It is noted that $\Pr(choice_{qr}^{(\delta)})$ is in essence a conditional probability function, because passenger q's journey time, δ_q^{OBS}, has been already known. Still, his/her actual route choice is not observable. In this sense, we may only speculate on the passenger's route choice in the event of his/her journey time being observed. To elucidate this point, we may as well consider the observation of passenger q's journey time as an independent event, which we represent by $\boldsymbol{\delta}_q = \{\delta_q^{\text{OBS}}\}$. From this, we acquire

$$\Pr(choice_{qr}^{(\delta)}) = \Pr(choice_{qr} \mid \boldsymbol{\delta}_q). \tag{3-8}$$

As such, equation (3-7) is adapted straightforwardly as follows:

$$\sum_{r \in R} \Pr(choice_{qr} \mid \boldsymbol{\delta}_q) = 1; \tag{3-9}$$

and this new term, $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q)$, should be interpreted as the probability that passenger q might have chosen route r, given the evidence that his/her journey time is δ_q^{OBS}. It may serve as an acceptable answer to question (I) posed in the previous section thus.

In Bayesian statistics, $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q)$ is termed as a posterior probability of passenger q's route choice, in that it would only be learnt after taking into account his/her journey time observation. We would expect to work out this conditional choice probability for all alternative routes within R. Furthermore, a certain route $r^* \in R$ could be deemed to be the most probable (rather than the actual) choice of q if the following statements would be true:

$$\begin{cases} \arg\max\limits_{r \in R} \Pr(choice_{qr} \mid \boldsymbol{\delta}_q) \neq \varnothing\,; \\ \forall choice_{qr}^{(\delta)} \in C_q^{(\delta)} :\ \exists\, r^* \in \arg\max\limits_{r \in R} \Pr(choice_{qr} \mid \boldsymbol{\delta}_q)\,. \end{cases} \tag{3-10}$$

But to approach an answer to question (II) posed in **Section 3.1** (*see* p.32), it would fundamentally depend on whether the conditions (3-10) stated above could be met. It must be noted, however, that r^* may not be the actual route choice – even if $\Pr(choice_{qr^*} \mid \boldsymbol{\delta}_q) = \max\limits_{r \in R} \Pr(choice_{qr} \mid \boldsymbol{\delta}_q)$.

In conformity with Bayes' theorem (*see* Laplace, 1995, pp.135-142), the following formula is fully acknowledged:

$$\Pr(choice_{qr} \mid \boldsymbol{\delta}_q) = \frac{\Pr(choice_{qr})\Pr(\boldsymbol{\delta}_q \mid choice_{qr})}{\Pr(\boldsymbol{\delta}_q)}\,, \tag{3-11}$$

which certainly ensures that equation (3-9) holds true, in that $\Pr(\boldsymbol{\delta}_q)$, the denominator on the right-hand side of formula (3-11), remains the same for every alternative route. In addition, this term indicates that the probability that the journey time of q is δ_q^{OBS}, irrespective of occurrence of any other events. According to the law of total probability (*see* Zwillinger and Kokoska, 1999, p.31), $\Pr(\boldsymbol{\delta}_q)$ can be factored as

$$\Pr(\boldsymbol{\delta}_q) = \sum_{r \in R} \Pr(choice_{qr})\Pr(\boldsymbol{\delta}_q \mid choice_{qr})\,. \tag{3-12}$$

That is, it is also equivalent to the sum of the corresponding numerator over all routes.

As regards the numerator, $\Pr(choice_{qr})$ is termed a prior probability in this context. As mentioned earlier, this term may be interpreted as the probability that q might have chosen r. From the perspective of discrete choice modelling, it may be perceived as the personal propensity of q to choose r from R_q. In order to learn or predict such a preference, we commonly resort to the methods of discrete choice analysis, which, however, require data of the actual or stated route choice having been made by q (*cf.* **Section 2.5**). In this respect, we have made it quite plain that when such data is not available, the discrete choice models would not be manageable. Besides, $\Pr(\boldsymbol{\delta}_q \mid choice_{qr})$ is correspondingly termed a likelihood function, which indicates the likelihood that $\boldsymbol{\delta}_q$ would take place, on condition that the event, $choice_{qr}$, has already occurred. Since $\Pr(\boldsymbol{\delta}_q)$ $\forall q \in Q$ is positive, the posterior probability, $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q)$, should be directly proportional to the product of the prior probability and the likelihood function below:

$$\Pr(choice_{qr} \mid \boldsymbol{\delta}_q) \propto \Pr(choice_{qr})\Pr(\boldsymbol{\delta}_q \mid choice_{qr}).$$

It is clear that if there exists a route r, in which case the product – the numerator of the fraction in formula (3-11) – can be maximised, it also maximises the posterior probability of our interest. However, neither $\Pr(choice_{qr})$ nor $\Pr(\boldsymbol{\delta}_q \mid choice_{qr})$ is understandable per se in light of information of only one individual. On this account, they would have to be learnt from the frequentist view based on data at an aggregate level.

As $\Phi_q^{(\delta)}$ gathers all possible route choices of q, a sample space of all such possibilities for the population, Q, on the given network of o-d can be formulated upon $\bigcup_{q \in Q} \Phi_q^{(\delta)}$. We represent this sample space by $\Phi^{(\delta)}$, which is defined as

$$\Phi^{(\delta)} := \left\{ \langle q, r, \delta_q^{\text{OBS}} \rangle :\ q \in Q,\ r \in R \right\}. \tag{3-13}$$

This is based on an assumption that each passenger has completed only one single journey.[6] In reality, different passengers might have chosen the same route and happened to have the same journey time. It should be noted that $\langle q, r, \delta_q^{\text{OBS}} \rangle$ $\forall q \in Q$ under consideration is actually different from one passenger to another, in that each observation of δ_q^{OBS} is peculiar to q and all individuals within Q are assumed to be independent of one another.

We let $C^{(\delta)}$ denote a set of events for the pair of o-d. It is defined accordingly as

$$C^{(\delta)} := \left\{ choice_{qr}^{(\delta)} :\ q \in Q,\ r \in R \right\}. \tag{3-14}$$

[6] It should be pointed out on this occasion that the practical data of observed journey time may be (unbalanced) panel data (*e.g.* the Oyster data), where one passenger may make a number of journeys between the same O-D at different periods. However, the mixture model (that will be described in **Section 3.3**) cannot deal with such panel characteristics. In that case, we can only assume that the route-choice decisions made by the same individual are independent over time; and every journey record is associated with a virtually 'different' individual.

Within the range of $C^{(\delta)}$, in practice, the prior probability $\Pr(choice_{qr})\ \forall r \in R$ indicates an average probability that any individual passenger (drawn randomly from the whole passenger population Q) might have used route r, regardless of his/her journey times. In other words, $\Pr(choice_{qr})$ could be perceived, and hence measured, as the proportion of passengers who have actually chosen r. As such, this can be understood as the relative frequency of counts of the passengers (or journeys) on route r in the context of frequentist statistics. Meanwhile, the likelihood function $\Pr(\boldsymbol{\delta}_q \mid choice_{qr})\ \ \forall q \in Q$ should express a probability that the observed journey time of q would have been δ_q^{OBS}, given the fact that he/she actually chose r. Since every individual who chose r is assumed to be identical, $\Pr(\boldsymbol{\delta}_q \mid choice_{qr})$ essentially becomes a probability distribution of the journey time distribution of the r-th route.

Now that the problem of passenger's route choices has been being surveyed in a probabilistic context, the risk of identity (3-1) being a false statement would substantially diminishes, and that should be defused by differences in choice probabilities among available alternative routes. In general, we would expect that the passenger's choice probability, measured between 0 and 1, shall approximate 1 for the chosen route, while those for other alternatives included in R must be approaching, but not necessarily, 0 (*cf.* **Section 6.3**).

Based on the probability space specified above, we are thus driven towards looking at the problem of passengers' probabilistic route choices in terms of the conditional probability distribution of their journey times. Within the scope of Bayesian framework, we therefore introduce, in the next section, another formulation of the problem and more specifically, a mixture distribution of the passengers' journey times.

3.3 Finite mixture model for journey time distribution

In this section, we follow up the example network of *o-d* depicted in **Figure 3.1** (*see* p.35), and formulate the probabilistic route-choice problem from another angle, in order to explore a possible solution to this problem.

Considering the availability of the N_R alternative routes, the whole passenger population Q is presumably composed of N_R subpopulations, each of which aggregates all the passengers in Q who have chosen one of the alternative routes. We let Q_r denote the subpopulation of route r, and use δ_r to represent a random variable of journey time of passengers travelling through the r-th route. It is plausible that individual journeys completed by the passengers from Q_r must collectively yield a certain distribution of δ_r for each alternative route; and the mean values (and/or medians) for all those journey time variables $\delta_1,\ldots,\delta_{N_R}$ would be likely to be statistically different from one route to another (though this presupposition would not necessarily be the case if the alternative routes are practically similar).

Moreover, all the individual journeys based on $Q_1,\ldots,Q_{N_R}$, in the aggregate, would also contribute a distribution of journey times of the heterogeneous population as a whole. In fulfilment of definition of the mixture distribution, by reference to McLachlan and Peel (2000, pp.6-8) as well as Frühwirth-Schnatter (2006, pp.1-23), such a journey time distribution can be considered as a mixture of the journey time distributions of $\delta_1,\ldots,\delta_{N_R}$, each being termed a component distribution of the mixture. More specifically, in our case, this is in essence a finite mixture distribution for a collection of a finite number of the journey time variables. Therewith it would also show, albeit not necessarily, the presence of heteroscedasticity among the N_R component distributions for the varied subpopulations $Q_1,\ldots,Q_{N_R}$.

It is also noteworthy at this point that we may actually redefine the subpopulation as well as the corresponding variables, whereby we use a mixture distribution to describe other statistical events for a given O-D. For instance, in the context that passengers choose from among a set of hyperpaths (*cf.* **Section 2.2**), a group of passengers travelling on the same hyperpath could then be referred to as a subpopulation. In this case, we shall consider passengers' choices of alternative hyperpaths, instead of a single route described above. Similarly, we may also distinguish different classes of travellers, such as slower and faster walkers. Then a component distribution should correspond to a probability distribution of journey times of a specific passenger class.

In the scope of this thesis, we consider only the general case introduced at the beginning of this section.

3.3.1 Model formulation

Now we take a step further to inspect the posterior probability, $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q)$, in the setting of the mixture distribution. In line with our specific target set in the previous section, the notation below is used to set the stage for the formulation of a finite mixture distribution of passengers' journey times.

Notation:

Q_r	subpopulation of all passengers who chose route r
δ_r	journey time of r between o and d (*referring to* **Figure 3.1**)
δ	journey time of travelling from o to d
$m(\delta)$	probability density function[7] of a mixture distribution of δ [8]
$c_r(\delta)$	probability density function of probability distribution of δ_r, also referred to as component distribution associated with r
ω_r	mixture weight placed on $c_r(\delta)$
$\boldsymbol{\omega}$	N_R-dimensional vector of all ω_r for $m(\delta)$
$\boldsymbol{\theta}_r$	vector of the distribution parameter(s) of $c_r(\delta)$
$\boldsymbol{\Theta}$	N_R-dimensional vector of all $\boldsymbol{\theta}_r$ for $m(\delta)$

According to the common definition, a mixture distribution or, equivalently, a probabilistic mixture model (*cf.* McLachlan and Peel, 2000, p.6), is generally specified to be a weighted sum of **p**robability **d**ensity **f**unction**s** (PDFs) of all the relevant component distributions. In that sense, the mixture weight[9] that is placed on each of the components should indicate an average probability that any given value (or any observed value at random) within the whole statistical population may be sourced from that component distribution. For practical

[7] In this thesis, the probability density function may also be treated as a probability mass function whereby the probability of the journey time taking any given value could be figured out.

[8] Note that δ of the function, $m(\delta)$, indicates a vector of the random variables, *i.e.* $\delta = (\delta_1 \ldots \delta_{N_R})$.

[9] In different literatures, it is also called mixing/mixture probabilities or proportions, *etc*. In this thesis, it is referred to as 'mixture weight', to avert any confusion.

applications, the most commonly used mixture is the finite mixture with the components all (being assumed to be) having the same distributional form, and hence the same estimator(s) for parameter(s). In the scope of this thesis, we only examine this class of mixture distributions, which is referred to as the standard mixture model, and later applied to data of passengers' journey times.

We use the symbol δ to represent passengers' journey time for travelling between $o\text{-}d$, and treat it as a random variable. Then we let $m(\delta)$ denote its PDF. It shall be a mixture of N_R components, each of which can be represented by $c_r(\delta)$ as the PDF of δ_r . Further, denote by $\boldsymbol{\omega}$ a random vector of the mixture weights, that is, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{N_R})$ with $\omega_r \; \forall r \in R$ being a random variable of the mixture weight placed on the r-th component PDF $c_r(\delta)$. Now a finite mixture model of passengers' journey time could be represented in the form as follows:

$$m(\delta \,|\, \boldsymbol{\omega}) = \sum_{r \in R} \omega_r c_r(\delta), \tag{3-15}$$

where $0 \le \omega_r \le 1 \;\; \forall r \in R$, and

$$\sum_{r \in R} \omega_r = 1. \tag{3-16}$$

It is noticeable that there appears to be a formal resemblance between the mixture PDF specified by formula (3-15) and the total probability presented as formula (3-12). In fact, there is a close correspondence in nature between the two formulas. Based on the premises stated in **Section 3.2** that all passengers share the same route-choice set and that they choose their own travel routes independently, the passengers are deemed identical individuals. In that sense, any samples of δ_q^{OBS} drawn randomly from the passenger population are independent, and identically distributed. At the aggregate level, this assumption allows the term $\Pr(\boldsymbol{\delta}_q)$ to generalise the probability distribution of all the passengers' journey times. As such, $\Pr(\boldsymbol{\delta}_q)$ does correspond to the mixture PDF $m(\delta)$.

Moreover, $c_r(\delta)$ is specific to route r , and it gives information about how likely it is that a certain journey time would have been experienced by any passenger who have actually chose the r-th route. Thus, this PDF is, in effect, a general form of the likelihood function $\Pr(\boldsymbol{\delta}_q \,|\, choice_{qr})$ being presented as formula (3-12).

Additionally, in our case, the mixture weight ω_r should, as explained above, refer to the probability that route r would be chosen by any individual within the whole passenger population Q. Therefore, for each individual, there is general equivalence between ω_r and $choice_{qr}$, as well as that between $\boldsymbol{\omega}$ and C_q in accordance with definition (3-3). For each alternative route, $\Pr(choice_{qr})$ also corresponds with the probability distribution of ω_r, again based on the same underlying assumption. Besides, due to the constraint specified by equation (3-16), values taken by $\boldsymbol{\omega}$ should depend only on $N_R - 1$ of all the mixture weights.

It must be pointed out that in the common specification of the mixture model, ω_r is usually perceived as a real-valued quantity. As a matter of fact, it shall appear as a probability function. At this stage, we may also suppose that $\Pr(choice_{qr})$ refer to a known quantity. Therefore, we could have

$$\Pr(\boldsymbol{\delta}_q \mid choice_{qr}) = c_r(\delta = \delta_q^{\text{OBS}}); \tag{3-17}$$

while for all $q \in Q$,

$$\Pr(choice_{qr}) = \omega_r, \tag{3-18}$$

In line with formula (3-11), the posterior probability of passenger q choosing route r, given his/her journey time δ_q^{OBS}, could be calculated by the formula as follows:

$$\Pr(choice_{qr} \mid \boldsymbol{\delta}_q) = \frac{\omega_r c_r(\delta = \delta_q^{\text{OBS}})}{m(\delta = \delta_q^{\text{OBS}} \mid \boldsymbol{\omega})}. \tag{3-19}$$

Now if we could solve both the components and their mixture weights, each individual's probabilistic route choices would be learnt in terms of the route-choice probabilities that are contingent upon observing his/her journey time. A set of the choice probabilities for all alternative routes would then provide a feasible, complete answer to question (I) posted in **Section 3.1** (*see* p.32) and would also lay a foundation for inferring, rather than determining, each passenger's route choice.

To figure out formula (3-19), we shall consider further a parametric equivalent of formula (3-15). Let $\boldsymbol{\Theta}$ denote a random vector of parameters for the mixture component distributions, *i.e.* $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_R})$, with $\boldsymbol{\theta}_r \ \forall r \in R$ being specifically for $c_r(\delta)$. In this way, formula (3-15) is adapted to:

$$m(\delta \mid \boldsymbol{\omega}, \boldsymbol{\Theta}) = \sum_{r \in R} \omega_r c_r(\delta \mid \boldsymbol{\theta}_r). \tag{3-20}$$

Note that $\boldsymbol{\theta}_r$ is also a random vector with its dimension depending upon the total number of parameters of $c_r(\delta)$. It could be either a unit vector (in the case that $c_r(\delta)$ corresponds to a probability distribution with only one parameter) or a multi-dimensional case otherwise.

3.3.2 Incomplete data

In order for the model (3-20) to fit in a specific case, data is a matter of vital importance for learning its parameters. We use the following notation below for a mathematical representation of our available data as well as the posterior probabilities of individual's route choices estimator upon the data.

Notation:

n	sample size of a given data set; $n \in \mathbb{N}_{\geq 1}$ and $n \leq N_Q$
$\Delta^{\texttt{DES}}$	set of **des**ired ($\texttt{DES}$) data, which includes both of passengers' route choices and their journey times
$r^{(q)}$	categorical variable of component-label, indicating the route choice of passenger q; $r^{(q)} \in R$
Δ	set of all journey time observations for *o*-*d* (*see* **Figure 3.1**)
$\hat{\omega}_r$	estimate of mixture weight ω_r
$\hat{\boldsymbol{\omega}}$	estimate of vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{N_R})$
$\hat{\boldsymbol{\Theta}}$	estimate of vector $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_R})$
$\hat{\boldsymbol{\theta}}_r$	estimate of vector $\boldsymbol{\theta}_r$
$\pi(\cdot)$	posterior probability (density) function for passengers' route choices given their journey times
$\pi_{qr}^{\texttt{MIX}}$	posterior probability that q chose route r (given $\delta_q^{\texttt{OBS}}$), estimated from a **mix**ture ($\texttt{MIX}$) model
$\boldsymbol{\Pi}_{\Delta}^{\texttt{MIX}}$	$n \times N_R$ matrix that enumerates all $\pi_{qr}^{\texttt{MIX}}$ estimated from a mixture model on Δ

Consider a random sample of n passengers, where $n \in \mathbb{N}_{\geq 1}$ and $n \leq N_Q$. Ideally, we would expect to get a set of data that shows every individual's journey time as well as his/her actual chosen route. We can represent the desired data set by $\Delta^{\texttt{DES}} = \{\langle \delta_q^{\texttt{OBS}}, r^{(q)} \rangle : q = 1, \ldots, n\}$, with $r^{(q)} \in R$ being a route-label for q, where each data point, $\langle \delta_q^{\text{OBS}}, r^{(q)} \rangle$, would correspond to a piece of record showing that q has actually used the $r^{(q)}$-th route, and his/her journey time is δ_q^{OBS}. As per the individual route-labels, the full data sample can be directly divided into N_R sub-data sets, whereby we would be able to derive an estimate of ω_r $\forall r \in R$. Also, for each of the alternative routes, $\boldsymbol{\theta}_r$ could be learnt as well based on the sub-data set associated with route r.

In our case, however, since we are using the smart-card data, it only provides us with a data set of the passengers' journey times. We may represent this data set by $\Delta = \{\delta_q^{\text{OBS}} : q = 1, \ldots, n\}$. It is hereby referred to as incomplete data due to its lack of the information on the route-labels $\{r^{(1)}, \ldots, r^{(n)}\}$, compared to Δ^{DES}. In this regard, we perceive Δ to be a sample of journey time observations each being attached with a hidden route-label (still denoted by $r^{(q)}$ $\forall q = 1, \ldots, n$). As such, $r^{(q)}$ turns into a random variable that follows a categorical distribution. It is noted that this distribution also corresponds to that of ω_r on the premise that all sampled individuals are identical to each other. On this basis, we use a function, denoted by $\pi(r^{(q)} \mid \delta)$, to report each individual's probabilistic route choices, or rather, probabilities of his/her route choice, conditional on his/her journey time. More specifically,

$$\pi(r^{(q)} = r \mid \delta, \boldsymbol{\omega}, \boldsymbol{\Theta}) = \frac{\omega_r c_r(\delta \mid \boldsymbol{\theta}_r)}{m(\delta \mid \boldsymbol{\omega}, \boldsymbol{\Theta})}. \tag{3-21}$$

As stated in **Section 3.2**, we are expecting to estimate a set of such posterior probabilities of every individual facing all alternative routes. For convenience, we use π_{qr}^{MIX} to represent the estimate of $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q)$ $\forall q = 1, \ldots, n$, $\forall r \in R$, where the superscript '$\texttt{MIX}$' stands for '**mix**ture model' and it indicates that the mixture distribution per se is actually a naive Bayes model[10] (*cf.* Lowd and Domingos, 2005). Therefore, we could have

$$\pi_{qr}^{\text{MIX}} = \frac{\hat{\omega}_r c_r(\delta = \delta_q^{\text{OBS}} \mid \hat{\boldsymbol{\theta}}_r)}{m(\delta = \delta_q^{\text{OBS}} \mid \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\Theta}})}, \tag{3-22}$$

(*see next page*)

[10] The mixture model here will be used as a basis for further updating of the estimator of each individual passenger's posterior probabilities of route choices. The superscript used here also serves as an identifier that will distinguish the posterior probability estimates of a mixture model in this chapter from the updated ones in **Chapter 5**.

where $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \ldots, \hat{\omega}_{N_R})$ and $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{N_R})$, with $\hat{\omega}_r$ and $\hat{\boldsymbol{\theta}}_r$ $\forall r \in R$ being the parameter estimates relating to $c_r(\delta)$. Note that

$$\sum_{r \in R} \pi_{qr}^{\texttt{MIX}} = 1, \tag{3-23}$$

Based on the dataset, Δ, the set of posterior probability estimates can then be enumerated by $\boldsymbol{\Pi}_{\Delta}^{\texttt{MIX}}$ in the form of a $n \times N_R$ matrix:

$$\boldsymbol{\Pi}_{\Delta}^{\texttt{MIX}} = \begin{pmatrix} \pi_{11}^{\texttt{MIX}} & \cdots & \pi_{1N_R}^{\texttt{MIX}} \\ \pi_{21}^{\texttt{MIX}} & \cdots & \pi_{2N_R}^{\texttt{MIX}} \\ \vdots & \ddots & \vdots \\ \pi_{n1}^{\texttt{MIX}} & \cdots & \pi_{nN_R}^{\texttt{MIX}} \end{pmatrix}, \tag{3-24}$$

Also, $\boldsymbol{\Pi}_{\Delta}^{\texttt{MIX}}$ would actually serve as the probability measure defined on $C^{(\delta)}$ that has been defined in **Section 3.2** (*see* definition (3-14), p.40). To gain knowledge of $\boldsymbol{\Pi}_{\Delta}^{\texttt{MIX}}$, our goal now is to seek the estimates, $\hat{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\Theta}}$.

3.3.3 Model estimation

In this section, the notation listed below is used for a description of the general estimation procedure of the mixture model parameters.

Notation:

$\ell(\cdot)$	likelihood function
Δ_r^{KMS}	set of journey time observations, which is produced by *K*-means (`KMS`) clustering and labelled r
η_r^{KMS}	median (or centroid-value) of Δ_r^{KMS}
$a(\delta_q^{\text{OBS}}) = r$	function that relates journey time observation δ_q^{OBS} to Δ_r^{KMS}
$\kappa(\cdot)$	objective function to be minimised for *K*-means clustering
σ_r^{KMS}	standard deviation of set Δ_r^{KMS}
ω_r^{KMS}	proportion of sub-dataset Δ_r^{KMS} in data set Δ
μ_r	subpopulation mean of Q_r
σ_r	subpopulation standard deviation of Q_r
$\boldsymbol{\mu}$	N_R-dimensional vector containing all subpopulation means μ_r
$\boldsymbol{\sigma}$	N_R-dimensional vector containing all subpopulation standard deviations σ_r

The **E**xpectation-**M**aximisation (EM) algorithm introduced by Dempster *et al.* (1977) can be employed to estimate the parameters, $\boldsymbol{\omega}$ and $\boldsymbol{\Theta}$, of the mixture model as specified by formula (3-20). In practice, as elucidated by Redner and Walker (1984, p.197), this algorithm effectively provides an iterative procedure that searches for the most likely – or rather the optimal – values of the unknown distribution parameters with respect to a data sample. The acquisition of the estimates is predicated on the maximisation of its likelihood or log-likelihood function of which the value shall increase at every iteration.

In our case, we let $\ell(\boldsymbol{\omega}, \boldsymbol{\Theta}; \Delta)$ denote the likelihood function of the combined set of $\boldsymbol{\omega}$ and $\boldsymbol{\Theta}$, given the data set, Δ. The corresponding log-likelihood function of the mixture journey time distribution can be represented in the form (3-25) as follows:

$$\log \ell(\boldsymbol{\omega}, \boldsymbol{\Theta}; \Delta) = \sum_{q=1}^{n} \log\left(\sum_{r=1}^{N_R} \omega_r c_r(\delta = \delta_q^{\mathrm{OBS}} \mid \boldsymbol{\theta}_r) \right). \tag{3-25}$$

Note that the iteration of the estimation may stop at either a local or the global maximum of log-likelihood function above.

Generally, the EM algorithm handles the data sample in accord with the following steps (*cf.* McLachlan and Peel, 2000, pp.48-50):

(i) Initialise both $\boldsymbol{\omega}$ and $\boldsymbol{\Theta}$, and label the initial values as $\boldsymbol{\omega}^{(0)}$ and $\boldsymbol{\Theta}^{(0)}$, respectively, which will be entered into the next step.

To be more specific, we shall be considering that $\boldsymbol{\omega}^{(\mathrm{E})} = \boldsymbol{\omega}^{(0)}$ and $\boldsymbol{\Theta}^{(\mathrm{E})} = \boldsymbol{\Theta}^{(0)}$, where the superscript '(E)' on the symbols signifies that the values are used for step (ii) – referred to as 'Expectation' (or commonly, 'E-step').

(ii) For the 'Expectation': calculate $\boldsymbol{\Pi}_{\Delta}^{\mathrm{MNB}}$ according to formula (3-22), with the data of Δ, given that $\hat{\boldsymbol{\omega}} = \boldsymbol{\omega}^{(\mathrm{E})}$ and $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(\mathrm{E})}$. The calculation result thus yields a conditional distribution of $r^{(q)}$.

On this basis, the 'Expectation' function of the log-likelihood, which is formulated as (3-26) below, is computed; and it will be maximised in step (iii) – referred to as 'Maximisation' (or commonly, 'M-step'):

$$\sum_{q=1}^{n} \sum_{r=1}^{N_R} \pi_{qr}^{\mathrm{MNB}} \left[\log \omega_r + \log c_r(\delta \mid \boldsymbol{\theta}_r) \right]. \tag{3-26}$$

(iii) For ‘Maximisation’: find optimal values of the parameters, labelled $(\boldsymbol{\omega}^{(\mathrm{M})},\boldsymbol{\Theta}^{(\mathrm{M})})$, which maximise (or increase the current value of) the ‘Expectation’ function (3-26); and then let $(\boldsymbol{\omega}^{(\mathrm{E})},\boldsymbol{\Theta}^{(\mathrm{E})})$ be updated with $(\boldsymbol{\omega}^{(\mathrm{M})},\boldsymbol{\Theta}^{(\mathrm{M})})$.

(iv) Repeat (ii) and (iii) until the improvement on the value of function (3-26) is no more than a pre-specified small constant – referred to as a threshold.

It should be noted that to gain $(\boldsymbol{\omega}^{(\mathrm{M})},\boldsymbol{\Theta}^{(\mathrm{M})})$ at step (iii) is in fact to search for optimal values of $(\hat{\omega}_1,\ldots,\hat{\omega}_{N_R},\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_{N_R})$. For this purpose, we shall need to take the derivatives of function (3-25) with respect to each of the parameters, and set them to equal 0, respectively. By doing so, we would have $\boldsymbol{\omega}^{(\mathrm{M})}=(\omega_1^{(\mathrm{M})},\ldots,\omega_{N_R}^{(\mathrm{M})})$ with

$$\omega_r^{(\mathrm{M})}=\frac{\sum_{q=1}^{n}\pi_{qr}^{\mathrm{MNB}}}{n}. \tag{3-27}$$

On the other hand, however, $\boldsymbol{\Theta}^{(\mathrm{M})}=(\boldsymbol{\theta}_1^{(\mathrm{M})},\ldots,\boldsymbol{\theta}_{N_R}^{(\mathrm{M})})$ would be derived on the understanding that the distributional form of δ_r is available.

Besides, it has been demonstrated by Seidel *et al.* (2000) that specifications for the initialiser of the model parameters for step (i), as well as the stopping criteria regarding step (iv), might exert some influence on the model estimation. A decent set of initial values as well as a threshold that terminates the iterations could play a sensitive role in securing credible, practical estimates via the general EM algorithm described above.

There are a number of studies (*e.g.* McLachlan, 1988; Melnykov and Melnykov, 2012; as well as Blömer and Bujna, 2013) having been devoted to the efforts to test different initialising strategies. In most cases, the K-means[11] clustering method (*cf.* Forgy, 1965; and MacQueen, 1967) is well-qualified to afford an acceptable starting point. The symbol ‘K’ refers literally to the total number of clusters into which a data set shall be categorised. In our case, K is equal to the size of route-choice set, *i.e.* $K=N_R$; and all the journey time observations, $\delta_1^{\mathrm{OBS}},\ldots,\delta_n^{\mathrm{OBS}}$, should be divided into K subsets. As for the term ‘means’, it may refer to a vector of K centroid-values, which we represent by $(\eta_1^{\mathrm{KMS}},\ldots,\eta_{N_R}^{\mathrm{KMS}})$,

[11] It is also referred to as the ‘*k*-means’ in many literatures, *e.g.* MacQueen (1967).

with η_r^{KMS} denoting the centroid-value of a sub-dataset denoted by Δ_r^{KMS}. In this way, the superscript, '`KMS`', indicates that both the centroid-value and the sub-dataset are generated based on the ***K*-means** (`KMS`).

In general, we use the *K*-means clustering method to include δ_q^{OBS} (from Δ) into Δ_r^{KMS} by minimising the total 'distances' from δ_q^{OBS} $\forall q = 1,\ldots,n$ to η_r^{KMS} over all the N_R sub-datasets. In this thesis, we measure this 'distance' by the absolute difference between each $\delta_q^{\mathrm{OBS}}$and the median based on the corresponding sub-dataset. Within Δ_r^{KMS}, each of journey time observations is supposed to be tightly close to η_r^{KMS}, and is as far from the observations of other sub-datasets as possible. The objective function to be minimised is:

$$\kappa(a(\delta), \eta_{a(\delta)}^{\mathrm{KMS}}) = \sum_{q=1}^{n} \left| \delta - \eta_{a(\delta)}^{\mathrm{KMS}} \right|, \tag{3-28}$$

where $a(\delta)$ acts as a classifier, namely, a function that iteratively (re-)labels an observed journey time as belonging to one of the sub-datasets, $\Delta_1^{\mathrm{KMS}},\ldots,\Delta_{N_R}^{\mathrm{KMS}}$, until function (3-28) reaches a local (or, though not necessarily, a global) minimum. More specifically, given that $a(\delta = \delta_q^{\mathrm{OBS}}) = r$, the observation, δ_q^{OBS}, is classified as an element in Δ_r^{KMS}. Hereby the method of *K*-means clustering may provide a rudimentary (but sensible) partition of the sample data into N_R mutually exclusive sub-datasets. So, with Δ_r^{KMS} $\forall r \in R$, in addition to η_r^{KMS}, initial values for parameters such as the standard deviation (denoted by σ_r^{KMS} $\forall r \in R$) could also be obtained in the light of specific mixture models. Moreover, an initial value for ω_r (denoted by ω_r^{KMS}) could be gained by calculating the proportion of all observations clustered in Δ_r^{KMS} among all those included in Δ. As such, this method is similar to the EM algorithm but confined to deterministic clustering with the data (*cf*. Bishop, 2006, pp.443-444).

Equally important to setting initial values is the selection of the threshold value. This would mainly be related to the speed of convergence of the algorithm, and determine whether the iteration should proceed or stop. As Karlis and Xekalaki (2003) pointed out, a smaller value of the threshold affords a more demanding stopping condition of the iterative computation, and hopefully would be more likely to make for the global maximum likelihood; but it may also cause a slower convergence of the estimates, or even worse, a failure of convergence when a predefined maximum number of the iterations for estimation has been reached. As a matter of fact, it must be noted that the distribution type of the components and hence the mixture is not known to us in our case. We should avoid blindly pursuing the global maximum of the log-likelihood function, because doing so may potentially lead to the problem of overfitting (*see* Guyon and Elisseeff, 2003; and Guyon *et al.*, 2010). That is to say, a model being estimated might be almost perfectly fit for a data sample, yet the estimates of the model parameters might not be explicable. In that sense, we may test a range of threshold values for model estimation, so as to locate the optimal values that practically imitate the actuality given the data available, regardless of whether the estimated results refer to a local or the global optimisation.

Now that for each of the alternative routes, we know nothing about its journey time distribution $c_r(\delta)$, an immediate thought (in most practical applications) is to assume that each of the route-specific journey time variables δ_r may be following some common statistical distribution, such as a Gaussian distribution (also known as normal distribution) or a log-normal distribution. Suppose, for example, that $c_r(\delta \mid \boldsymbol{\theta}_r)$ is a PDF of a Gaussian distribution that we could represent by $\mathcal{N}(\mu_r, \sigma_r)$. That is to say, $\delta_r \sim \mathcal{N}(\mu_r, \sigma_r)$ $\forall r \in R$, with μ_r and σ_r denoting, respectively, the mean and the standard deviation of the sub-population, Q_r. In that case, for each $r \in R$, δ_r is a Gaussian (or normal) random variable; and $\boldsymbol{\theta}_r$ corresponds to a vector, (μ_r, σ_r). The mixture distribution thus formed is a Gaussian mixture distribution, with its PDF being parameterised by $\boldsymbol{\omega}$ and $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{N_R})$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{N_R})$. Accordingly, we could adapt function (3-20) for a Gaussian mixture as follows:

$$m(\delta \mid \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{r \in R} \omega_r c_r(\delta \mid \mu_r, \sigma_r). \tag{3-29}$$

The adaptation to any other probability distributions, *e.g.* the aforementioned log-normal distribution, would do likewise with their own specific distribution parameters.

3.4 Inference of passenger traffic distribution

By applying the EM algorithm to cluster an available data set of journey time observations, $\Delta = \{\delta_q^{\text{OBS}} : q = 1, \ldots, n\}$ (*referring to* **Section 3.3.2**), $\boldsymbol{\Pi}_{\Delta}^{\text{MIX}}$ could be acquired through the acquisition of $(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\Theta}})$. As explained in **Section 3.3.1**, $\hat{\omega}_r$ reflects the proportion of passenger-traffic on route r. Besides this aggregate measure, we are also interested in trying to infer each individual's real route choice from $\mathbf{\Pi}_{\Delta}^{\mathtt{MIX}}$ (*see* question (II) posed in **Section 3.1**, p.32), and attempting to find out route-specific sub-datasets of the smart-card data.

The notation used to illuminate the methods of inference is listed below.

Notation:

$\zeta(\cdot)$	assignment function used for the *naive* inference of each passenger's route choice, based on mixture model
$\Delta_r^{\mathtt{INF}_0}$	set of journey time data of passengers who chose r, based on the *naive* **inf**erence ($\mathtt{INF}_0$)
$n_r^{\mathtt{INF}_0}$	number of passengers using r, based on the *naive* inference
$\omega_r^{\mathtt{INF}_0}$	proportion of passengers using route r, according to the *naive* inference
Λ_q	random variable for passenger q, which follows the standard uniform variable; $\Lambda_q \sim \boldsymbol{\mathcal{U}}(0,1)$
$\mathbf{\Lambda}$	vector of all Λ_q, for a given data set of journey times
λ_q	generated (real-valued) number of Λ_q
$\xi(\cdot)$	assignment function used for the *effective* inference of each passenger's route choice, based on mixture model
$\Delta_r^{\mathtt{INF}}$	set of journey time data of passengers using r, based on the *effective* **inf**erence ($\mathtt{INF}$) from a mixture model
$n_r^{\mathtt{INF}}$	number of passengers using r, based on the *effective* inference
$\omega_r^{\mathtt{INF}}$	proportion of passengers using route r, according to the *effective* inference

3.4.1 *Naive* inference

Recall the initial assumption that has been made on each individual's possible route choices from **Section 3.2** (*see* the conditions formulated by (3-10), p.39). If $\pi(r^{(q)} = i \mid \delta_q^{\mathrm{OBS}}) \geq \pi(r^{(q)} = j \mid \delta_q^{\mathrm{OBS}})\ \ \forall i, j \in R$, then it might be relatively more likely that q chose the i-th route. Following this logic, the simplest inference could be drawn that route i is the actual route choice made by q. Accordingly, we define an assignment function $\zeta(\cdot)$ by setting

$$\zeta(q) = \arg\max_{r \in R} \pi_{qr}^{\mathrm{MNB}}. \tag{3-30}$$

The function, $\zeta(q)$, as defined by equation (3-30), labels passenger q (hence his/her journey time observation δ_q^{OBS}) as being from route r, given the highest π_{qr}^{MNB} among all $r \in R$. By doing so, a rough estimate of every individual's actual route choice could be learnt; and a total of N_R sub-datasets of Δ are also sorted out accordingly, with each being related to one of the alternative routes. We let $\Delta_r^{\text{INF}_0}$ denote the r-th sub-dataset according to such inference, and it could be expressed as follows:

$$\Delta_r^{\text{INF}_0} = \left\{\delta_q^{\text{OBS}} : \zeta(q) = \{r\}\right\}. \tag{3-31}$$

On this basis, the sum of passengers in Δ, who chose route r, should then be equal to the size of $\Delta_r^{\text{INF}_0}$, which we represent by $n_r^{\text{INF}_0}$. That is,

$$n_r^{\text{INF}_0} = \left|\Delta_r^{\text{INF}_0}\right|. \tag{3-32}$$

Furthermore, denote by $\omega_r^{\text{INF}_0}$ the proportion of passenger-traffic shared by route r to the entire passenger traffic. It could thus be estimated as follows:

$$\omega_r^{\text{INF}_0} = \frac{n_r^{\text{INF}_0}}{n}. \tag{3-33}$$

Note that in fact this is a marginal inference (*cf.* Leonard *et al.*, 1989) such the estimates derived from it will be referred to here as a naive Bayesian inference. This may be more advisable in a situation that the observations are entirely distinguishable or the true sub-datasets are mostly non-overlapping.

3.4.2 *Effective* inference

It should again be noted that any route r (or say the r^* presented in condition (3-10), *see* **Section 3.2**, p.39), albeit with the highest posterior choice probability among all the alternatives, may or may not be the actual choice of the passenger q. On that account, we further allow for such uncertainty for each individual.

In addition to the *naive* inference above, some unknown/random factor shall be taken into consideration for the comparisons of π_{qr}^{MNB} $\forall r \in R$; while still, we give priority to r of which the posterior estimate is relatively higher. Thus, we draw support from the order statistics of $\pi_{q1}^{\text{MNB}},\ldots,\pi_{qN_R}^{\text{MNB}}$, which we represent by $\pi_{q(1)}^{\text{MNB}},\ldots,\pi_{q(N_R)}^{\text{MNB}}$, with $\pi_{q(r)}^{\text{MNB}}$ $\forall r \in R$ being the r-th smallest estimate of π_{qr}^{MNB}. What is more is that we bring in a n-dimensional random vector, denoted by $\mathbf{\Lambda}$, where

$\mathbf{\Lambda} = (\Lambda_1, \ldots, \Lambda_n)$, with Λ_q following the standard uniform distribution – namely, $\Lambda_q \sim \boldsymbol{\mathcal{U}}(0,1)\ \ \forall q = 1, \ldots, n$. For each q within the data sample, a random number – or rather, a pseudo-random number, denoted by λ_q, is generated from the unit interval, wherein all the values, denoted by $\lambda_1, \ldots, \lambda_n$, are equally likely, and $0 \le \lambda_q \le 1$ (*cf.* Riley and Goucher, 2009, pp.131-132).

We hereby define another assignment function, which we represent by $\xi(q)$, and it is expressed as follows:

i. if $\lambda_q \le \pi^{\mathrm{MNB}}_{q(N_R)}$, then $\xi(q) = \{r:\ r = N_R\}$, which corresponds to $\zeta(q)$ (*see* function (3-30));

ii. otherwise, for $j = 0, \ldots, N_R - 1$, if $\sum_{i=N_R-(j-1)}^{N_R} \pi^{\mathrm{MNB}}_{q(i)} < \lambda_q \le \sum_{i=N_R-j}^{N_R} \pi^{\mathrm{MNB}}_{q(i)}$, then $\xi(q) = \left\{r:\ r = N_R - j\right\}$.

Similar to the *naive* inference (*cf.* formula (3-31), *see previous page*), we now use Δ^{INF}_r to represent the sub-data set relating to route r based on $\xi(q)$; and so then

$$\Delta^{\mathrm{INF}}_r = \left\{\delta^{\mathrm{OBS}}_q :\ \xi(q) = \{r\}\right\}. \tag{3-34}$$

Likewise, we let n^{INF}_r and ω^{INF}_r denote, respectively, the total number of passengers in Δ who chose route r and the estimated proportion of passenger-traffic shared by r. In contrast to formulas (3-32) and (3-33) (*see previous page*), we have

$$n^{\mathrm{INF}}_r = \left|\Delta^{\mathrm{INF}}_r\right|, \tag{3-35}$$

and hence

$$\omega^{\mathrm{INF}}_r = \frac{n^{\mathrm{INF}}_r}{n}, \tag{3-36}$$

which is believed to afford a more robust estimate than $\omega^{\mathrm{INF}_0}_r$ at the aggregate level.

3.5 Interpretation and validation of mixture model estimates

So far, all the estimates derived from the mixture model have actually been based on the posterior probability distribution of the hidden route-labels, $r^{(q)}\ \ \forall q$. According to the sub-dataset of journey time data that are inferred for each route r, the real sub-dataset might be learnt from its corresponding component distribution $c_r(\delta)$, with some essential statistical features, *e.g.* the mean journey time. Notwithstanding this, however, there is a lack of evidence of a one-to-one correspondence between an estimated component of the mixture and an alternative route in the real world. This factual circumstance could not be immediately determined. In other words, it is not yet known which one of the components (labelled r) mirrors the journey time distribution of which route in reality, nor is it confirmed if the estimates per se are credible. As such, the implementation of the finite mixture models is a process of an 'unsupervised classification' (also known as 'unsupervised learning'), which detects and attempts to reveal latent categories of the observational data (*cf.* Bousquet *et al.*, 2004, pp.77-112). Besides, as stated by James *et al.* (2013, p.374), the estimated results are in fact difficult to evaluate – not just because of the independence of the data, but also because of unavailability of a benchmark for validation. To tackle these issues, as rules of thumb, it would be necessary to

(a) identify comparable features between the estimated components from the mixture model and the routes in the real world; and

(b) ponder how to make a judgement about those comparable features.

In this section, we propound some general criteria for the assessment of applicability of the mixture model in our case.

3.5.1 Expectation of journey time for a given pair of O-D

Following the two principles outlined above, the first thought upon comparable features is the mean journey time. This is because, on the one hand, the mixture model under discussion is examining the probability distributions of journey time between a given pair of O-D, wherein the mean value plays an essential role. We would also expect there to be differences among the mean journey times of the different routes, whereby the component distributions of the alternatives could be distinguished from one another in terms of their central locations.

On the other hand, in practice, it would be possible for us to calculate an expected journey time for each alternative route completely independently of the mixture model. In that regard, the following notation (*see next page*) is employed to formulate the computation of the route-specific journey time.

Notation:

$T^{\mathtt{ENT}}$	time-stamp at which passengers pass through a ticket gate to enter o, referred to as '**ent**ry time' $(\mathtt{ENT})$
l'	transit line for the first leg of a single journey
$t^{\mathtt{ACC}}_{l',o}$	**acc**ess $(\mathtt{ACC})$ walking time from a gateline to the l'-platform at o
$T^{\mathtt{ARR}}_{l',o}$	time of passengers' **arr**ival $(\mathtt{ARR})$ on a l'-platform at o
$t^{\mathtt{WFD}}_{l',o}$	**w**aiting time to board a l'-train **f**or **d**eparture $(\mathtt{WFD})$ from the l'-platform at o
$T^{\mathtt{DEP}}_{l',o}$	time of passengers' **dep**arture $(\mathtt{DEP})$ from the l'-platform at o
$T^{\mathtt{dep}}_{l',o}$	time of **dep**arture $(\mathtt{dep})$ of a l'-train from the l'-platform at o
s	interchange station between o and d
$t^{\mathtt{OBT}}_{l',[o,s]}$	**o**n-**b**oard **t**ravel $(\mathtt{OBT})$ time in a l'-train running from o to s
$t^{\mathtt{run}}_{l',[o,s]}$	**run**ning $(\mathtt{run})$ time of a l'-train, from o to s
$T^{\mathtt{ARR}}_{l',s}$	time of passengers' **arr**ival $(\mathtt{ARR})$ on the l'-platform at s
l''	transit line for the second leg of a single journey
$t^{\mathtt{TIC}}_{[l',l''],s}$	walking time to **t**ransfer from the l'-platform to the l''-platform at **i**nter**c**hange $(\mathtt{TIC})$ station s
$T^{\mathtt{ARR}}_{l'',d}$	time of passengers' **arr**ival $(\mathtt{ARR})$ on the l''-platform at d
$T^{\mathtt{DEP}}_{l'',s}$	time of passengers' **dep**arture $(\mathtt{DEP})$ from the l''-platform at s
$t^{\mathtt{OBT}}_{l'',[s,d]}$	**o**n-**b**oard **t**ravel $(\mathtt{OBT})$ time in a l''-train running from s to d
$T^{\mathtt{dep}}_{l'',s}$	time of **dep**arture $(\mathtt{dep})$ of l''-train from the l''-platform at s
$t^{\mathtt{run}}_{l'',[s,d]}$	**run**ning $(\mathtt{run})$ time of a l''-train, from s to d
$t^{\mathtt{WIC}}_{l'',s}$	**w**aiting time to board a l''-train for departure from the l''-platform at **i**nter**c**hange $(\mathtt{WIC})$ station s
$t^{\mathtt{EGR}}_{l'',d}$	**egr**ess $(\mathtt{EGR})$ time from the l''-platform to a gateline at d
$T^{\mathtt{EXT}}$	time-stamp at which passengers pass through a ticket gate to exit from d, referred to as '**exit** time' $(\mathtt{EXT})$
h	label of travel route, referred to as 'route-label'
$t_h(\phi,\psi)$	journey time of passengers travelling by h, given that he/she boards the ϕ-th arriving train at o (and, if h involves interchange, the ψ-th arriving train at s)
$y^{\mathtt{PSG}}$	walking speed on level/ramp **p**a**s**sa**g**eways $(\mathtt{PSG})$
$y^{\mathtt{UPS}}$	walking speed of going **ups**tairs $(\mathtt{UPS})$

(*Continued*)

Notation: (*Continued.*)

$y^{\mathtt{DNS}}$	walking speed of going **d**ow**ns**tairs ($\mathtt{DNS}$)
$y^{\mathtt{ESC}}$	**esc**alators/lifts ($\mathtt{ESC}$) speed
$\mathbf{y}$	vector that contains passengers' walking/moving speeds on each type of pathways
u	underground station (representing o, d or s)
$\tau_{hu}^{\mathtt{WLK}}$	expected **walk**ing ($\mathtt{WLK}$) time at u along h
$x_{uh}^{\mathtt{PSG}}$	total length of level/ramp **pas**sa**g**eways ($\mathtt{PSG}$) at u on h
$x_{uh}^{\mathtt{UPS}}$	total run of **s**tairs for going to **up**per ($\mathtt{UPS}$) levels at u on h
$x_{uh}^{\mathtt{DNS}}$	total run of **s**tairs for going **d**ow**n** ($\mathtt{DNS}$) to lower levels at u on h
$x_{uh}^{\mathtt{ESC}}$	total run of **esc**alators/lifts ($\mathtt{ESC}$) at u on h
$\mathbf{x}_{uh}$	vector that contains reciprocals of distances for each type of pathways at u on h;
ϕ	number of attempts to successfully board a train at o
ψ	number of attempts to successfully board a train at s
$t_h^{\mathtt{REF}}$	expected average journey time of travelling by h, serving as a **ref**erence ($\mathtt{REF}$) value for interpreting estimates from a mixture model
v_h	indicator that equals one if h is a direct service, and zero if it is an indirect service
$u(v_h)$	function that indicates whether a station on h is s or d
$l(v_h)$	function that indicates whether a transit line on h is l' or l''
$\hat{\sigma}_h$	estimate of a sample standard deviation of journey time of h
$\hat{\sigma}_h^{\mathtt{SEM}}$	estimate of a standard error of the mean journey time of h
$t_*(\cdot)$	Student's t-value with certain degrees of freedom and a given probability level $*$

We let $T^{\mathtt{ENT}}$ denote the time-stamp at which passengers pass through a ticket gate of the origin station o. This information is easily obtainable from smart-card data. Denote by l' a transit line the passengers decide to take; and further to this, let $t_{l',o}^{\mathtt{ACC}}$ and $T_{l',o}^{\mathtt{ARR}}$ denote, respectively, the passengers' walking time from the gateline to a platform for l' and their arrival time on that platform. Then we can have: (*see next page*)

$$T_{l',o}^{\mathrm{ARR}} = T^{\mathrm{ENT}} + t_{l',o}^{\mathrm{ACC}}.$$

Furthermore, we let $t_{l',o}^{\mathrm{WFD}}$ and $T_{l',o}^{\mathrm{DEP}}$ denote, respectively, the passengers' waiting time to board a train of l' and their departure time with l' from o. Apparently,

$$t_{l',o}^{\mathrm{WFD}} = T_{l',o}^{\mathrm{DEP}} - T_{l',o}^{\mathrm{ARR}}; \tag{3-37}$$

Assuming that all trains on the underground network are running on schedule, thus, in accordance with the timetable, $T_{l',o}^{\mathrm{DEP}}$ is equal to the scheduled departure time of the train, which we represent by $T_{l',o}^{\mathrm{dep}}$. In that way,

$$T_{l',o}^{\mathrm{DEP}} = T_{l',o}^{\mathrm{dep}};$$

and

$$t_{l',o}^{\mathrm{WFD}} = T_{l',o}^{\mathrm{dep}} - T_{l',o}^{\mathrm{ARR}}.$$

If line l' serves an indirect route, in which case a transfer is necessary at an interchange station. We use the letter s to represent the interchange station, and let $t_{l',[o,s]}^{\mathrm{OBT}}$ denote the expected on-board travel time on l' between the platforms of o and s. Based on the assumption above, $t_{l',[o,s]}^{\mathrm{OBT}}$ would be equivalent to the corresponding train's scheduled running time, which we represent by $t_{l',[o,s]}^{\mathrm{run}}$. That is to say,

$$t_{l',[o,s]}^{\mathrm{OBT}} = t_{l',[o,s]}^{\mathrm{run}}. \tag{3-38}$$

Thus, the time of passengers' arrival at s, denoted by $T_{l',s}^{\mathrm{ARR}}$ accordingly, is expected to be calculated as follows:

$$T_{l',s}^{\mathrm{ARR}} = T_{l',o}^{\mathrm{DEP}} + t_{l',[o,s]}^{\mathrm{OBT}},$$

and also,

$$T_{l',s}^{\mathrm{ARR}} = T_{l',o}^{\mathrm{dep}} + t_{l',[o,s]}^{\mathrm{run}}.$$

Suppose the passengers need to transfer at s from l' to a connecting line, denoted by l'', which links s to the destination station d. We then use $t_{[l',l''],s}^{\mathrm{TIC}}$ to represent their transfer/walking time from the platform for l' (at o) to another for l'' at s. With this information, the time of their arrival on l''-platform, which is denoted by $T_{l'',s}^{\mathrm{ARR}}$, is expected to be calculated as follows:

$$T_{l'',s}^{\mathrm{ARR}} = T_{l',s}^{\mathrm{ARR}} + t_{[l',l''],s}^{\mathrm{TIC}};$$

Similar to the first journey leg, the expected departure time of the passengers from s, denoted by $T_{l'',s}^{\mathrm{DEP}}$, and their on-board travel time between platforms of s and d, denoted by $t_{l'',[s,d]}^{\mathrm{OBT}}$, are both assumed to be in line with the service timetable. Namely,

$$T_{l'',s}^{\mathrm{DEP}} = T_{l'',s}^{\mathrm{dep}};$$

and

$$t_{l'',[s,d]}^{\mathrm{OBT}} = t_{l'',[s,d]}^{\mathrm{run}}. \tag{3-39}$$

where $T_{l'',s}^{\mathrm{dep}}$ and $t_{l'',[s,d]}^{\mathrm{run}}$ denote correspondingly the scheduled departure time and the running time of the train, respectively. On this basis, the passengers' waiting time to board a train of l'' for departure from s, which we represent by $t_{l'',s}^{\mathrm{WIC}}$, is expected to be

$$t_{l'',s}^{\mathrm{WIC}} = T_{l'',s}^{\mathrm{dep}} - T_{l'',s}^{\mathrm{ARR}}; \tag{3-40}$$

and the time of the passengers' arrival on the platform of the destination d, denoted by $T_{l'',d}^{\mathrm{ARR}}$, can be calculated as follows:

$$T_{l'',d}^{\mathrm{ARR}} = T_{l'',d}^{\mathrm{DEP}} + t_{l'',[s,d]}^{\mathrm{run}}.$$

Moreover, we let $t_{l'',d}^{\mathrm{EGR}}$ denote the passengers' egress/walking time from the platform for l'' to a gateline at the destination d. So the time-stamp of their exit, denoted by T^{EXT}, is expected to be

$$T^{\mathrm{EXT}} = T_{l'',d}^{\mathrm{ARR}} + t_{l'',d}^{\mathrm{EGR}}.$$

Given the fact that the above derivation process is independent from the mixture model where the letter, r, has been serving as a component-label (*i.e.* a hidden route-label), we use another letter, h, to represent each of the alternative routes in the real world. That is, h acts as a real-world counterpart of r.

The expected journey time of route h, which we represent by t_h is identified by (o,d,s,l',l''). Thus it is straightforwardly calculated as follows:

$$t_h = t_{l',o}^{\mathrm{ACC}} + t_{l',o}^{\mathrm{WFD}} + t_{l',[o,s]}^{\mathrm{OBT}} + t_{[l',l''],s}^{\mathrm{TIC}} + t_{l'',s}^{\mathrm{WIC}} + t_{l'',[s,d]}^{\mathrm{OBT}} + t_{l'',d}^{\mathrm{EGR}}; \tag{3-41}$$

and if l' connects o and d directly, formula (3-41) would then become

$$t_h = t_{l',o}^{\mathrm{ACC}} + t_{l',o}^{\mathrm{WFD}} + t_{l',[o,d]}^{\mathrm{OBT}} + t_{l'',d}^{\mathrm{EGR}}. \tag{3-42}$$

Note that information about $t^{\mathtt{ACC}}_{l',o}$, $t^{\mathtt{TIC}}_{[l',l''],s}$ and $t^{\mathtt{EGR}}_{l'',d}$ could be obtained from either field surveys or approximate calculation based on existing research findings.

Here we describe a simple method for gaining a practical approximation of their expected values. Let $\mathbf{y}=(y^{\mathtt{PSG}},y^{\mathtt{UPS}},y^{\mathtt{DNS}},y^{\mathtt{ESC}})$ denote a vector of speeds of passengers' walking on the different types of passageways at any station (*cf.* **Section 2.3.1**), where $y^{\mathtt{PSG}}$, $y^{\mathtt{UPS}}$, $y^{\mathtt{DNS}}$ and $y^{\mathtt{ESC}}$ denote the walking speeds of passengers moving along level/ramp passages ($\mathtt{PSG}$), climbing upstairs ($\mathtt{UPS}$), downstairs ($\mathtt{DNS}$), and on the stationary escalators/lifts ($\mathtt{ESC}$), respectively. Then we could represent the layout of each station in terms of the type and length of its passageways. We use u to represent an underground station. It represents any station of the origins, destinations or interchanges. For each u, let $x^{\mathtt{PSG}}_{uh}$, $x^{\mathtt{UPS}}_{uh}$, $x^{\mathtt{DNS}}_{uh}$ and $x^{\mathtt{ESC}}_{uh}$, denote the total lengths measured for each type of passageways of h. Note that the measurement for the stringer lengths of a stairway/escalator may depend on both the angle and height (*see e.g.* Davis and Dutta, 2002); or the total run or the total rise may be measured instead. Thus, a simple linear expression could be considered to relate all the above-mentioned factors to passengers' average walking time along route h at station u. Let this average be denoted by $\tau^{\mathtt{WLK}}_{uh}$, with superscript '$\mathtt{WLK}$' being short for '**walk**ing time'. It is specified as follows:

$$\tau^{\mathtt{WLK}}_{uh}=\mathbf{y}\cdot\mathbf{x}_{uh}, \tag{3-43}$$

where $\mathbf{x}_{uh}=(1/x^{\mathtt{PSG}}_{uh},1/x^{\mathtt{UPS}}_{uh},1/x^{\mathtt{DNS}}_{uh},1/x^{\mathtt{ESC}}_{uh})$. It must also be noted that $\mathbf{y}$ may vary between different periods of a day, and should be non-linearly related to the pedestrian flows in different passageways. For practical purpose, we might only take consideration of the average values of the speeds for each type of passageways (*see e.g.* Daly *et al.*, 1991; as well as Lam and Cheung, 2000) during given a specific period, such as morning peak, off-peak, evening peak.

Besides, in uncongested conditions, as has been mentioned in **Section 2.3.2**, every passenger is assumed to be able to board the first arriving train of a line they choose when he/she arrives at the platform. Nonetheless, when the train arrives with carriages being almost fully loaded or overcrowded, there would be barely room available for extra boarding demand. In that situation, some passengers may fail, or reject, to board but would rather wait for the next coming trains; but also their waiting time spent on the platform would increase by, say a headway according to the timetable. This may happen at either the origin or the interchange station, or at both of the stations, especially at the morning/evening rush hour. As a result of boarding failure, an increase in the total journey time of individual passengers is foreseeable in light of their whereabouts (*e.g.* origin station or interchange) and the number of attempts-to-board they make. In view of this fact, we use $t^{\texttt{WFD}}_{l',o,\phi}$ to represent passengers' waiting time to successfully board a train at the ϕ-th attempt at an origin station; and denote by $t^{\texttt{ICW}}_{l'',s,\psi}$ the waiting time for boarding a train on a connecting line at the ψ-th attempt. Based on formulas (3-41) and (3-42), the journey time of passengers travelling by h could be generally specified by

$$t_h(\phi,\psi) = t^{\texttt{ACC}}_{l',o} + t^{\texttt{WFD}}_{l',o,\phi} + t^{\texttt{OBT}}_{l',[o,u(v_h)]} + v_h \cdot \left(t^{\texttt{TIC}}_{[l',l''],s} + t^{\texttt{WIC}}_{l'',s,\psi} + t^{\texttt{OBT}}_{l'',[s,d]}\right) + t^{\texttt{EGR}}_{l(v_h),d}, \quad (3\text{-}44)$$

where

$$v_h = \begin{cases} 1, & \text{if } h \text{ is an indirect route;} \\ 0, & \text{if } h \text{ is a direct route;} \end{cases}$$

$$u(v_h) = \begin{cases} s, & \text{if } v_h = 1; \\ d, & \text{if } v_h = 0; \end{cases}$$

and

$$l(v_h) = \begin{cases} l', & \text{if } v_h = 0; \\ l'', & \text{if } v_h = 1. \end{cases}$$

To present a general picture of the average journey time of each alternative route, we may consider the following four straightforward cases as follows.

I. For direct services:

   i. passengers get aboard the firstly arriving train after they arrive at the platform and depart from the origin station, *i.e.* $\phi = 1$; and

   ii. passengers get on board a train at the second attempt at the origin station, *i.e.* $\phi = 2$;

II. For indirect services:

   i. passengers can always get aboard the firstly arriving train after they arrive at the platforms of both the origin and the interchange stations, *i.e.* $\phi = 1$ and $\psi = 1$;

   ii. passengers get aboard a train at the second attempt at the origin station, and board the firstly arriving train at the interchange station, *i.e.* $\phi = 2$ and $\psi = 1$; (*see next page*)

iii. passengers board the firstly arriving train at the origin station after they arrive at a platform at the origin station, and get aboard at the second attempt at the interchange station, *i.e.* $\phi = 1$ and $\psi = 2$; and

iv. passengers get aboard at the second attempt at both the origin and the interchange station, *i.e.* $\phi = 2$ and $\psi = 2$.

Note that the expected journey time of h should be represented by a weighted average of $t_h(\phi,\psi)$ considering the four circumstances stated above (or even more complex situations). However, those weights for each circumstance are not known. In a simplistic way of calculation, we consider in this thesis only an average of $t_h(\phi,\psi)$ under the different circumstances specified above, that is,

$$t_h^{\text{REF}} = \frac{1}{4}\sum_{\phi=1}^{2}\sum_{\psi=1}^{2} t_h(\phi,\psi) . \tag{3-45}$$

It is hereinafter referred to as the (presumptive) expected journey time of travelling by h, and considered a prime indicator manifesting differences of journey time between alternative routes. This t_h^{REF}, together with $t_h(\phi,\psi)$, will all be used as reference values for matching a component-label (associated with the mixture model) to a route-label.

What is more, a **c**onfidence **i**nterval (CI) for t_h^{REF} at, say, the 95% **c**onfidence **l**evel (CL) would be further needed, so as to provide a reference range allowing for inherent errors in the specification and calculation of t_h^{REF}. In this regard, firstly, we may perceive each of $t_h(\phi,\psi)\ \forall \phi,\psi$ as an observation of journey time; and they together form a small sample of four observations. As such, t_h^{REF} actually serves as the sample mean, and we could estimate the corresponding sample standard deviation (denoted by $\hat{\sigma}_h$) as

$$\hat{\sigma}_h = \sqrt{\frac{1}{3}\sum_{\phi=1}^{2}\sum_{\psi=1}^{2}\left[t_h(\phi,\psi) - t_h^{\text{REF}}\right]^2} . \tag{3-46}$$

Secondly, we may also perceive each $t_h(\phi,\psi)$ itself as a sample mean of an arbitrary sample of journey times on route h, from which a standard error of the mean (denoted by $\hat{\sigma}_h^{\text{SEM}}$) could be estimated as follows:

$$\hat{\sigma}_h^{\text{SEM}} = \frac{\hat{\sigma}_h}{\sqrt{4}} . \tag{3-47}$$

For the estimation of a CI for t_h^{REF} at a certain CL, it would depend on the size of the available sample size as well as the distribution of the journey time of route h. For example, suppose the journey time distribution of h is Gaussian, where the variance or standard deviation of the population is unknown. We shall then consider t_h^{REF} to be a variable of sample mean for the Gaussian subpopulation of passengers who chose h. As such, $(t_h^{\text{REF}} - \mu)/\hat{\sigma}_h^{\text{SEM}}$ is following a Student's $\mathscr{t}$-distribution[12] with three degrees of freedom, *i.e.* $(t_h^{\text{REF}} - \mu)/\hat{\sigma}_h^{\text{SEM}} \sim \mathscr{t}(3)$; on this basis, the true mean of the Gaussian subpopulation would be likely to be within the range of $t_h^{\text{REF}} - \hat{\sigma}_h^{\text{SEM}} \mathscr{t}_{0.025}(3)$ to $t_h^{\text{REF}} + \hat{\sigma}_h^{\text{SEM}} \mathscr{t}_{0.025}(3)$, which refers to the CI at the 95% CL (*cf.* Johnson and Bhattacharyya, 2009, pp.351-358). Note that the CI for other types of distributions, such as log-normal distribution (*e.g.* Parkin *et al.*, 1990) would be different. In this thesis, we assume that the CI derived from Gaussian population could suffice to provide the reference range for most cases.

A set of criteria are proposed in the following subsection; and details will be explained with specific case studies in the next chapter.

3.5.2 General principles

The following notation is used to in this section to clarify the proposed some general principles for the interpretation and validation of the estimation results obtained from the mixture model.

Notation:

$\hat{\mu}_r$	estimate of subpopulation mean μ_r
$\hat{\sigma}_r$	estimate of subpopulation standard deviation σ_r
GOF	indicator of **g**oodness-**o**f-**f**it between observed and simulated journey time data
Δ^{SIM}	**sim**ulated (`SIM`) data set of passengers' journey times, which is generated from a mixture model (being estimated)
$\delta_{\mathscr{g}}^{\text{SIM}}$	**sim**ulated journey time, with subscript $\mathscr{g}$ being its index; $\mathscr{g} \in \mathbb{N}_{\geq 1}$

[12] We use the symbol, $\mathscr{t}$, to represent this distribution in order to avoid confusion with the variables represented by using the letter, t.

To have a preliminary review of the model rationality, as a rule of thumb, the first is to compare the estimate of mean journey time for each mixture component and t_h^{REF} for each alternative route. Ideally, as mentioned above, it would be expected that t_h^{REF} $\forall h$ would be distinguishable from one another. For each r, we let $\hat{\mu}_r$ and $\hat{\sigma}_r$ denote the (real-valued) estimates of the mean and standard deviation of journey time, respectively. Ideally, it would also be expected that $\hat{\mu}_1,\ldots,\hat{\mu}_{N_R}$ are distinguishable from one another, and so would be $t_1^{\mathrm{REF}},\ldots,t_{N_R}^{\mathrm{REF}}$. Note that differences among derived values from either way may not necessarily be clearly identifiable. That is to say, in some cases, we may obtain, for example, t_i^{REF} and t_j^{REF} ($\forall i, j \in R$ and $i \neq j$), which are fairly close or nearly the same. This might be because in actual fact the attributes of the two routes i and j are similar in almost every aspect, such as service timetable and common passageways. In this situation, we would then expect that $\hat{\sigma}_r$ $\forall r \in R$ would differ. Otherwise, the estimates of the mixture model might imply that the passenger-traffic are approximately equally distributed among the alternative routes, in the light of $\hat{\omega}_r$.

Further to the above consideration, $\hat{\mu}_r$ would also be expected to approximate a certain t_h^{REF} among all the alternative routes, whereby we may pre-match the component-label r to a route labelled h. For any of the pre-matched pair, there is bound to be a difference in the estimates. Yet the extent to which the difference might be acceptable should be on a case-by-case basis. A CI for t_h^{REF} at the 95% CL could be estimated and used further to provide a possible range of the mean journey time. As $\hat{\sigma}_h^{\mathrm{SEM}}$ is necessary for the calculation (*cf.* formulas (3-46) and (3-47)), how reliable the $\hat{\sigma}_h^{\mathrm{SEM}}$ is, is arguable (*cf.* Nagele, 2003). Meanwhile, we should also check $\hat{\mu}_r$ with $t_h(\phi,\psi)$ separately under each of the four specified circumstances. Extreme cases, *e.g.* $\phi,\ \psi > 2$, may also be taken into account, especially for rush-hour traffic, as this could be a case such that a fairly large difference between a pre-matched pair of component-label, r, and a route-label, h, would be interpreted. Otherwise, we shall consider that the model is not suitable.

Furthermore, we should look at the proportions of passenger-flow between the *effective* inference ω_r^{INF} from the model and the actual usage of r. In this respect, information about the latter may be based on earlier surveys. Likewise, in an ideal situation, it would be expected that both results would roughly equal each other. This comparison may largely be affected by the accuracy of the latter, the model per se notwithstanding, especially when the survey sample is quite small. If there appear to be considerable gaps between the sets of values, a larger survey sample should be required, and/or $\omega_r^{\mathrm{INF}_0}$ (based on the *naive* inference), could be further checked. Otherwise, we shall test some other parametric distribution as the components, or abandon the test model.

A final issue that needs to be addressed is selection from among all acceptable models. Suppose several different mixture models are tested and all of them can meet the criteria set out above. In that case, we shall not arbitrarily reject any of the test models, but should choose the one that provides a relatively better fit to the sample data of journey times. With regard to the selection of a Bayesian model, usually, penalised-likelihood information criteria are used as a reference (Dziak *et al.*, 2012), such as **A**kaike's **i**nformation **c**riterion (AIC) as well as **B**ayesian **i**nformation **c**riterion (BIC). Both AIC and BIC evaluate a model's goodness of fit; and for each of the two criterion, the lower value is yielded for a fitted model, the better the model should be. Their pros and cons have been discussed in a wide range of studies (*e.g.* Kuha, 2004). However, there may be potential for concern about inconsistency of data scale. In some circumstances, given the limitation of the software package for estimating the mixture models,[13] we may change the scale of the original data. For example, we may fit a Gaussian distribution to natural logarithms of log-normal data, as the logarithm of a log-normal variable is normally distributed (Mood *et al.*, 1974, pp.540-541). In that case, the scales of AIC/BIC differ between mixture models fitted for different scaled data set, and thus cannot be compared between the different test models.

On that account, we use the measure of normalised root mean square error as an indicator of **g**oodness **o**f **f**it of a mixture model, which we represent by gof. It measures normalised differences between the sample data set (still denoted by $\Delta = \{\delta_q^{\mathrm{OBS}} :\ q = 1,\ldots,n\}$; *see* **Section 3.3.2**) and a set of simulated data of the same size (Farmer and Sidorowich, 1987). We use Δ^{SIM} to represent the simulated data set, which should be generated from the test mixture model. In this case, it can be further expressed in the form of a set, $\Delta^{\mathrm{SIM}} = \{\delta_{\mathscr{q}}^{\mathrm{SIM}} :\ \mathscr{q} = 1,\ldots,n\}$, where $\delta_{\mathscr{q}}^{\mathrm{SIM}}$ denotes a single simulated data point, with $\mathscr{q}$ being its index. Let both Δ

[13] For example, for our case studies in **Chapter 4**, the software package is available only for Gaussian mixture model.

and Δ^{SIM} be treated as n-dimensional vectors, gof is computed as follows (*cf.* Chan and Cannon, 2002):

$$gof = \frac{\left\| \Delta - \Delta^{\text{SIM}} \right\|}{\left\| \Delta - \frac{1}{n}\sum_{q=1}^{n} \delta_q^{\text{OBS}} \right\|}. \tag{3-48}$$

It must be noted that, for computation of the numerator of formula (3-48), both the sampled journey time data and the simulated data should be sorted in an ascending order. We use the order statistics for both the data sets, which we could represent by $\delta_{(1)}^{\text{OBS}},\ldots,\delta_{(n)}^{\text{OBS}}$ (for the sampled data) and $\delta_{(1)}^{\text{SIM}},\ldots,\delta_{(n)}^{\text{SIM}}$ (for the simulated data), respectively, with the subscript, (q), indicating that the statistic is the q-th smallest value. In that way, formula (3-48) is equivalent to formula (3-49) as follows:

$$gof = \frac{\sqrt{\sum_{q=1}^{n}(\delta_{(q)}^{\text{OBS}} - \delta_{(q)}^{\text{SIM}})^2}}{\sqrt{\sum_{q=1}^{n}(\delta_{q}^{\text{OBS}} - \frac{1}{n}\sum_{q=1}^{n}\delta_{q}^{\text{OBS}})^2}} \tag{3-49}$$

Then we shall need to compare the values of gof between all the test models; and the mixture model with a lower gof should be considered a better fit. Since Δ^{SIM} is randomly generated given a candidate model, this comparison would need to be repeated a number of times, from which the model having the higher rate of gaining a better fit (*i.e.* a lower value of gof) would be preferable.

3.6 Summary and conclusions

Relying on Bayes' theorem, this chapter has formulated and discussed a probabilistic framework for the use of the finite mixture model to obtain passengers' route choices between a pair of O-D stations on the underground system. It has also proposed a set of complementary approaches to evaluate the model applicability in practical use. The model allows for each individual's route choice being learnt up to their choice probabilities for all alternative routes. It attempts to seek out passengers' route-choice probabilities in a situation where the passenger's actual route choice is not known. Such choice probabilities are, in essence, posterior probabilities estimated on condition that the passengers' journey time is observed (*cf.* **Appendix A**). The estimates are fundamentally reliant on observational data of the passengers' journey times being modelled by finite mixture distributions. Proportions of the passenger traffic flowing on different alternative routes could also be estimated accordingly, given the O-D travel demand. The inferences of traffic distributions are validated by comparing them to survey findings, which in turn provides corroborative evidence supporting the estimates of the individuals' probabilistic route choices.

In practice, there are several issues of applying the mixture model for practical use. Firstly, it should be noted that one aspect of this general mathematical problem is determining the optimal number of mixture components to fit. In that case, the number of the components is treated as a variable and shall be estimated together with the model parameters. However, in the scope of this thesis, we will consider only the situation that this number is a given constant. Also note that the specification of the model (including the number of alternative routes) must ensure that the estimated components would be explicable. In other words, whether the estimated mixture and the components are meaningful will depend on whether we are able to interpret them as being a mixture or components. For practical application, we could either refer to the existing data (*e.g.* survey data) or draw support from a choice-set generation model, in order to identify possible alternative routes hence an appropriate number of the mixture components.

Secondly, the distributional form of the component PDFs will have to be pre-specified for any mixture model. This prior knowledge can be of significant importance for the application of the mixture model, particularly about the types of journey time distributions, and the passenger-traffic proportion, of each alternative route. Still, characteristics of the journey time distribution for each of the alternative routes are not really known, unless a subset of route-specific journey time observations is available. In practical application, a range of different standard mixture distributions can be considered. Also, note that the journey times of different alternative routes on the same O-D may be following different types of statistical distribution, and more advanced mixture models can be studied for future research.

Thirdly, all the alternative routes (hence the universal route-choice set) must be identified. This thesis has only assumed every individual passenger would take into account the same choice set, *i.e.* the universal set, when starting a journey at the origin station. However, as has been also discussed in **Section 3.2**, different passengers travelling between the same O-D, might have their own different perceived route-choice sets. This refers to two aspects. For one thing, the choice sets may differ among those passengers, because different people might take into account different alternative routes and carry out different choice tasks. For another, the alternatives encompassed in a choice set may not be equally perceived by an individual in terms of their own preferences and different attributes of the alternative routes. Such attributes involve a variety of factors influencing passengers' travel decisions, including systematic variables (*e.g.* service frequency, walking distance for interchange), individual perceptions to over-crowding and seat availability, provision of real-time information, and other uncertainties, *etc*. Take a two-alternative case, for instance, while one route may be more likely to be used due to *e.g.* shorter travel time, the other is supposed to have relatively less chance of being selected. Therefore, the probability that an alternative route is chosen will also depend on to which choice set it belongs. Not all the alternatives would be simultaneously preferred by every passenger, especially when more alternatives become available. Then a challenge will be on how to explicitly specify or identify each individual's perceived choice set.

Fourthly, as emphasised in previous sections, the journey time variability, hence journey time distributions, over longer periods is uncertain. In this respect, a simulation-based transit assignment model may be useful for the estimation of the distribution of journey times.

Lastly, the context of the application of the mixture model has been confined in this thesis to the underground system only. Nevertheless, this method could be easily adapted to other transport systems, such as a road network and even a multi-modal transpiration system, provided that the essential data is accessible (*cf.* **Section 4.2**; *see also* **Figure 1.1**, p.5). For example, to estimate drivers' route choices on a detector/monitor-equipped road network, a mixture distribution of travel times could be specified for a given pair of monitoring points. This would rely largely on the corresponding data of the timestamps at which each car passes the monitors. In a multi-modal context, some travellers may transfer from one mode to another, or even use more modes to make a single trip between a given pair of O-D points. If information about their interchanges between the different modes is not available, we may also make use of the mixture model to estimate the ridership share on each mode based on the modelling of the O-D travel times.

Chapter 4
Application of mixture models of route choices: case studies on the London Underground

4.1 Introduction

This chapter presents a set of case studies on the London Underground (LU, also commonly known as the Tube) system. The LU provides us with a massive and well-developed transit network under the management of Transport for London (TfL)[1]; and the full network is shown in **Appendix B**, which is abridged from the original standard Tube map (Transport for London, 2013b).

In accordance with the zonal fare scheme[2] adopted by TfL, the rail network (including the LU, DLR, LO and National Rail services; *see also* Footnote 1) is carved out into eleven fare zones. A central zone (Zone-1) is based in the centre of London, which is surrounded by five concentric ring-zones (ranging from Zone-2 to Zone-6) radiating outward one by one. In addition, there exist three ancillary zones (Zone-7 to Zone-9) that are positioned to the northwest central London, plus 2 further zones (Watford Junction and Grays). Unlike buses and trams where flat fare regime is adopted, fare charged on using the rail services mostly depends on how many zones a passenger travels through.

In this context, the scope of the case studies was focused only within Zone-1 of the LU network. This area is illustrated in **Figure 4.1** (*see next page*), within which flat fare applies.

[1] As an integrated part of the local government, TfL undertakes responsibility for the transport system and its services across the Greater London area; and takes charge of the management and operations for multifarious public transport services. It involves a variety of transport modes including London Buses (Bus), LU, Dockland Light Railway (DLR), London Overground (LO), London Tramlink (Tram), London River Services (LRS), and some other customised services such as Dial-a-Ride particularly for disabled people, *etc*.

[2] Service fare on the LU is calculated on the basis of a zonal system, that is, a passenger would be charged a certain amount of fare according to the Zones in which his/her journey started and ended.

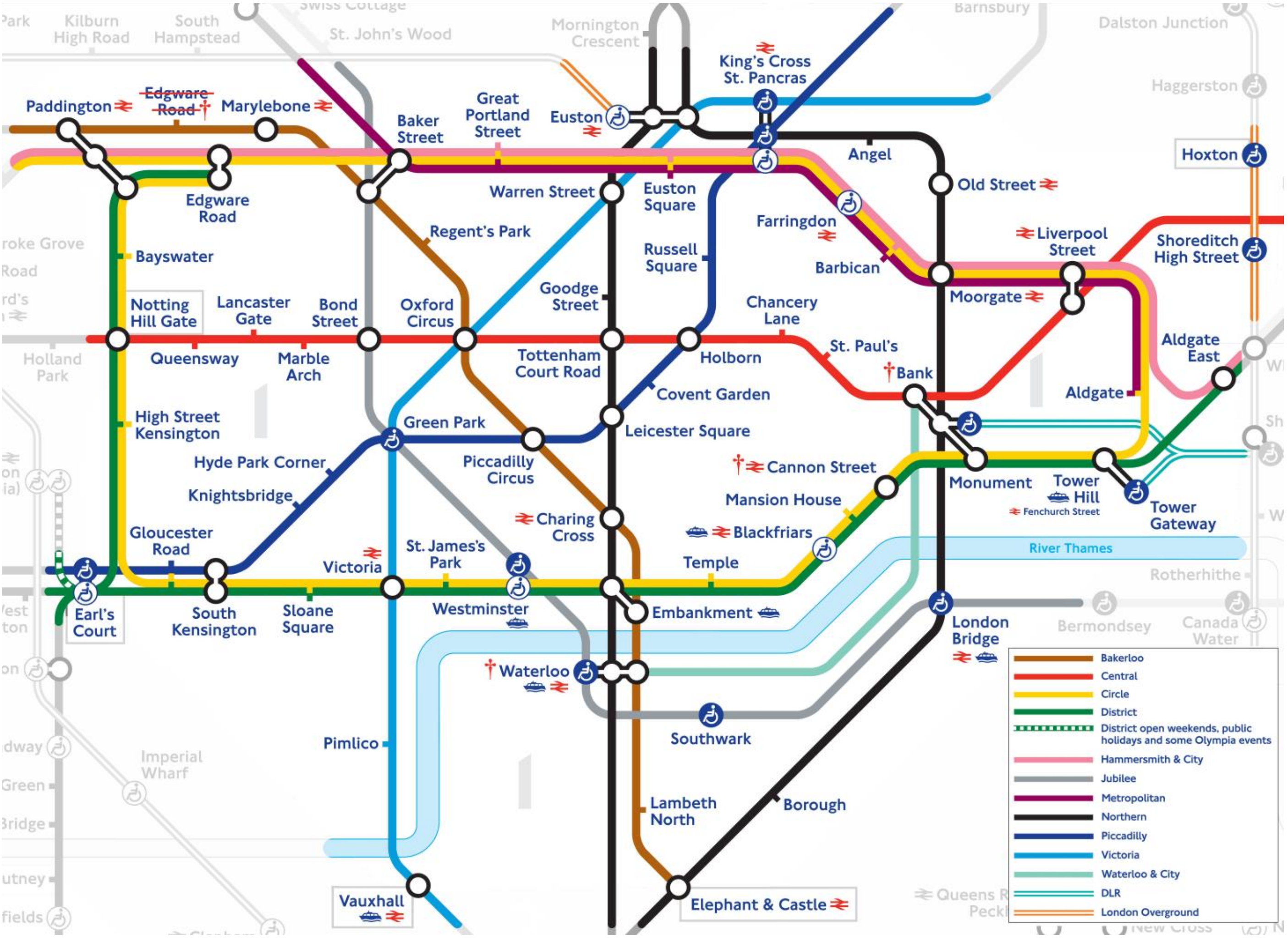


Figure 4.1 Tube map for Zone-1, abridged from the original standard Tube map (© 2013 Transport for London).

The main aims of conducting these case studies are (a) to demonstrate the application of the mixture models discussed in **Chapter 3**; as well as (b) to test the model applicability, namely, in what situation and to what extent the mixture models could be suitable for understanding passengers' route choices. In addition, a secondary objective of this chapter is to reveal what sorts of information the smart-card data may contain, as well as the role that it would potentially play for further research.

The rest of this chapter is arranged as follows. **Section 4.2** describes the data that will be used for applying the mixture models. In **Section 4.3**, considering a range of network-scales, case studies about the application of mixture models are looking at different O-D pairs with two or more alternative routes. A summary of findings is presented in **Section 4.4**.

4.2 Data description

4.2.1 The Oyster-card data

Individual passengers' journey times is without doubt the most important data for applying mixture models, particularly for estimating the model parameters. Passengers' journey records captured via Oyster smart-card system (hereafter referred to as Oyster) on the LU network is the only source of the journey time data used in this thesis.

The Oyster system implemented within the Greater London area and managed by TfL, is one of the most successful applications of AFC systems. Due to widespread usage of the smart-card, named as Oyster card, more than 80% of journeys across the TfL network are being paid via the Oyster (*cf.* Transport for London, 2012), taking advantage of its discounts in comparison to traditional paper tickets. Such a high market share sufficiently warrants the potential of the Oyster card data (hereafter referred to as Oyster data) for measuring various aspects of the quality of transit service, and being an effective data source for statistical analyses of exploring and revealing travel patterns on the TfL network. The Oyster data is collected automatically as the Oyster card being touched on a card reader. Miscellaneous information is then generated and appropriately stored in separate data subsets. They gather both aggregated statistical data (*e.g.*

the count of entries and exits at each station, and the number of journeys grouped by different transport modes, time periods, stations and ticket types) and, on the individual level, detailed (but anonymised) travel history of all Oyster card users (hereafter referred to as Oyster user) and the fare payment information. In addition, each of Oyster card users' journeys is presented seriatim as per transaction in the data set of travel history, which is mainly targeted in our following analysis.

Similar to many other AFC systems, the Oyster scheme records timestamps of every individual Oyster user's touch-in and a following touch-out, respectively, along with identities of the corresponding stations (*cf.* **Chapter 2**). This process corresponds to a single journey under the scope of this thesis, whereof the gateline-to-gateline journey time is referred to by Chan (2007) as **O**yster **j**ourney **t**ime, which we represent by OJT and treat as a random variable. Clearly, OJT is equivalent to the previously defined journey time variable, δ, that is discussed in the previous chapter. That is to say, for any O-D pair, which is connected by multiple alternative routes, it is believed that a sample of OJT observations (*i.e.* real-valued Oyster journey times) collected from all passengers (during a given period) would be following a mixture distribution. Still, we use T^{ENT} and T^{EXT} to denote, respectively, the timestamps of any Oyster user's entry and exit logged by the Oyster system, and treat them as random variables (*cf.* **Section 3.5.1**). Accordingly, OJT could be further represented as follows:

$$OJT = T^{\text{EXT}} - T^{\text{ENT}}, \tag{4-1}$$

which provides a straightforward calculation of the Oyster users' journey times. For convenience, we represent an observation of OJT by OJT^{OBS}, with the superscript 'OBS' being short for '**obs**erved value' (or '**obs**ervation'). Also, OJT^{OBS} of an Oyster user is equivalent to δ_q^{OBS}, which represents a journey time observation of a passenger labelled q (*cf.* **Chapter 3**). It must be noted that because of system constraints, all recorded timestamps of entry and exit are only accurate to minute, and so is the computed OJT^{OBS} (Chan, 2007). The Oyster system omits the time of seconds but rounds the timestamps to the nearest minute that is less than or equal to the actual clock time. This thus results in an error of up to 59 seconds (or -59 seconds) in the calculation of OJT^{OBS}. As such, this error could be regarded as purely random (*e.g.* a random variable following a uniform distribution over the interval of $[-59, 59]$ in seconds). In this thesis, however, only the computed OJT^{OBS} is taken into account. In line with the premise stated before (*see* **Section 3.2**), each count of the journey records in the Oyster data is associated with only one OJT^{OBS}.

As the main data source for this study, a processed data set that reveals the distributions of OJT^{OBS} was provided by TfL and filed in the format of a combination of origin, destination, date and time-band. Moreover, the data set was sorted into four time-bands, as specified in **Table 4.1** below.

Table 4.1 Time-bands set by TfL

Time-band	applying to an OJT^{OBS}, given T^{ENT} falling into the period:
AM Peak	between 07:00 and 10:00 (*a.m.*) on a weekday (Monday – Friday)
PM Peak	between 16:00 and 19:00 on a weekday (Monday – Friday)
Off-Peak	of any time during a weekday other than AM Peak and PM Peak
Weekend	of any time during Saturday and Sunday, and also bank holiday

The journey time distribution for each O-D is calculated in 99 percentiles each representing 1% of Oyster users travelling from the given origin to destination during a specified time period (date and time-band). It starts with the fastest 1% travellers followed by the second fastest and so on, and goes up to 99% in ascending order of recorded Oyster journey times of all complete Oyster journeys. The journey counts by time-band are given as well. Presumably, those OJT^{OBS} that are greater than 99th-percentile records for each time period are considered as outliers. By this, data for every combination of date and time-band shapes a cumulative distribution of OJT in that period; whereby the percent distribution is also calculated. However, those OJT^{OBS} whose values are not exceeding the upper outer fence, *i.e.* three times **i**nter**q**uartile **r**ange (IQR) more than the 75th quartile, would be considered to be valid entries (*cf.* Frigge *et al.*, 1989) for estimating the mixture models. Because, above that level, the data volatility is such that any observations are supposed to be uncorrelated with the provided transit services, hence treated as outliers. Also note that although OJT should be taken for a continuous variable, it can only take discrete values in minutes due to constraint of the Oyster system. Empirical distributions could be acquired from observed values that fall into minute-blocks.

In addition, supplementary information that is independent of the OJT^{OBS} data would also be indispensable to proceed to validate the results being estimated from the mixture models. Above all, knowledge of the route-specific average journey times (*i.e.* $t_h^{\mathrm{REF}} \quad \forall h \in R$), which is clearly not affected by the mixture model (*cf.* **Section 3.5.1**), would afford underlying evidence to relate a mixture component to an alternative route in real life. In this respect, several different sources have been also available to provide relevant information, presented in the following subsections.

4.2.2 Data for computation of route average journey time

A database established from the '**A**ccess, **E**gress and **I**nterchange' (AEI) survey on the LU system gives simple random samples of individual travellers' walking time on pre-determined pedestrian paths within the LU stations. From this, expected values of $t_{l',o}^{\mathrm{ACC}}$, $t_{l'',d}^{\mathrm{EGR}}$ and $t_{[l',l''],s}^{\mathrm{TIC}}$ (defined in **Section 3.5.1**) for each of the identified alternative route can be obtained. Note that there could possibly be several alternative passages for access, egress as well as interchange at some of the LU stations, though, only one pre-specified passage, as the mostly used pedestrian path, for each of station had been timed in the AEI survey. In that case, an online database, called **D**irect **E**nquiries[1] (DE), providing information about all available passages within each of the LU stations, will also be utilised to adjust the data of average walking times from the AEI survey (*see* formula (3-43), p.61).

The third data source sustaining the computation of $t_h(\phi,\psi)$ is the timetable of the LU lines services, which is available for all passengers. It provides the scheduled departure times, $T_{l',o}^{\mathrm{dep}}$ and $T_{l'',s}^{\mathrm{dep}}$, of each run of the transit lines, as well as their platform-to-platform running times, $t_{l',[o,s]}^{\mathrm{OBT}}$ and $t_{l'',[s,d]}^{\mathrm{OBT}}$. Therewith $t_{l',o}^{\mathrm{WFD}}$ and $t_{l'',s}^{\mathrm{WIC}}$ are also derived (from formulas (3-37) to (3-42), pp.59-60), whereby $t_h^{\mathrm{REF}} \quad \forall h \in R$ could be derived according to definition given by formula (3-44) (*see* p.62).

[1] Available online at http://www.directenquiries.com/londonunderground.aspx; last accessed on 30 September 2014.

4.2.3 Data from Rolling Origin and Destination Survey

We also cross-validate the empirical route-choice set of each of the O-D networks with the feedbacks on the alternative routes via **R**olling **O**rigin and **D**estination **S**urvey (RODS)[2] (up till 2010). The RODS is an annual project, having been conducted by TfL since 1998. From this database, we may learn information about each respondent's actual travel route, including the stations at which his/her journey started and ended, where he/she made an interchange, and also some basic socio-demographic data, including such as age and purpose of the journey (*cf*. Guo and Wilson, 2011).

According to TfL, this programme looks into the travel patterns on weekdays only, when the system is operating normally; and any undesirable actions (*e.g.* long-term closures) are not covered. Therefore, between a given O-D pair, what, or which, routes most passengers would commonly use for day-to-day commute could be learnt. What's more, the sum totals of the respondents choosing each of the alternative routes are counted as well, whereby proportions of passenger-traffic on each route can be roughly obtained from the relative frequency.

We let n_r^{ROD} and ω_r^{ROD} denote the count and the percentage of all respondents who made their journeys by using route r, respectively. It must be pointed out that n_r^{ROD} and ω_r^{ROD} are counted on the rolling twelve-year basis; and thus, they may not represent the true usage of r. Notwithstanding, ω_r^{ROD} would still serve as a comparatively good reference for us to assess the estimates from the mixture models in our case.

4.3 Case studies on the London Underground

This section elaborates on the manipulation of specific case studies with the data described above to demonstrate the application of the method elaborated in **Chapter 3**. Seven O-D pairs[3] within Zone-1 of the LU network were selected as

[2] Data descriptions available online at http://data.london.gov.uk/datastore/package/tfl-rolling-origin-and-destination-survey; last accessed on 30 September 2014.

[3] In total, there are seven cases of O-D pairs, only five of which will be presented in the current chapter, with the results of the rest two cases being exhibited in **Appendix C**.

typical examples for the demonstration, which will be investigated separately in the following sections.

The selection of these example networks is based primarily on the standard Tube map (*see* **Figure 4.1**, p.72), complying with conditions as follows.

- There are at least two travel routes available, which connect the stations of the origin and the destination.
- There must be no more than one interchange for each route between the O-D stations.
- All the alternative routes[4] (between the O-D stations) are empirically identified from the standard Tube map, and also checked with the RODS data for AM Peak (*i.e.* between 7:00 *a.m.* and 10:00 *a.m.* on weekdays, *see* **Table 4.1**, p.75).
- There is a relatively high volume of passenger traffic on the network, especially during the AM Peak, which thus makes for a sufficiently large data sample.

For each case study, the network will be illustrated in a map-view, which is tailored in the scope of the standard Tube map with essential elements being retained. Only the transit lines as well as intermediate stations pertinent to the O-D will maintain their original appearance on the Tube map, with the rest of the network being presented in monochrome.[5]

To apply the proposed method, a few prerequisites will have to be met. First of all, the type of journey time distribution of each alternative route, and hence the distributional form of the mixture model, need to be pre-specified. In this section, for each case study, we will look at two standard mixture distributions for a comparative test: a **G**aussian **m**ixture (GM), *i.e.* $\delta_r \sim \mathcal{N}(\mu_r, \sigma_r) \;\; \forall r \in R$, as well as a **l**og-**n**ormal **m**ixture (LNM), *i.e.* $\delta_r \sim \log \mathcal{N}(\mu_r, \sigma_r) \;\; \forall r \in R$; and compare the estimates for the two models. As stated by Marron and Wand (1992), the family

[4] Any travel route that does not involve interchange will be hereinafter referred to as a direct route/service; otherwise, an indirect route/service.

[5] It should be noted that confluences of passenger flows of those monochrome and coloured lines certainly will, in reality, affect the traffic and (hence) the (average) journey time of the O-D. However, this effect would not be relevant in that the traffic and journey times will both be statistically analysed given probability distributions, and that the transit lines and stations in monochrome could be left out of account.

of GM has a great flexibility such that it usually provides close approximation to arbitrary probability distributions in various contexts (*cf.* **Section 3.3.3**). Yet it should also be noted that the shape of a univariate Gaussian distribution is symmetric to its mean (and its median), whereas in reality a journey time distribution seems more often to be positively skewed (*see* Fu *et al.*, 2012a). On this account, the log-normal distribution may potentially be more suitable. For all this, both GM and LNM will be estimated for each of the selected single O-D pairs, by taking advantage of the Oyster data.

Within the scope of this thesis, the issues concerning the threshold value for estimating the mixture models are not addressed (*cf.* **Section 3.3.3**). For each of the case studies being conducted in this thesis, several threshold values were tested in the model estimation; and only the sensible and explicable results are presented in the following subsections for a demonstration of the application of the mixture model.

The available data of OJT^{OBS} for the model estimation were collected during the period from 27th June 2011 (Monday) to 30th March 2012 (Friday), spanning over 193 weekdays, which do supply each case with an adequate sample. It must be pointed out that since the logarithmic journey times should then follow a GM distribution, the two data sets, *i.e.* the raw data and its logarithms, were both fitted by GM model, respectively, given the same sample of journey times.

A summary of basic information about all the seven pairs of O-D is reported in **Table 4.2** (*see next page*), where n_0 and n denote the sample size before and after the extreme outlying values being excluded, respectively. Besides, the data sample of each of the case studies is for 193 weekdays, except that the ***Case-6*** and ***Case-7*** contain journey records data for 192 days and 162 days, respectively. The databases of AEI, RODS and the published timetable are all ready to use.

Table 4.2 An introductory summary of the LU case studies

EB, WB, NB and SB are short for **e**ast**b**ound, **w**est**b**ound, **n**orth**b**ound and **s**outh**b**ound, respectively.

Case	**Origin** o	**Destination** d	**Interchange**			**Journey time**	**RODS result**		**Sample size**
-	(LU station)	(LU station)	Line, l'	LU station, s	Connecting line, l''	t_h^{REF} (minute)	ω^{ROD} (%)	n^{ROD}	n / n_0
1	Victoria	Holborn	**Victoria** (NB)	Oxford Circus	**Central** (EB)	19.6	71.3	526	24,760 / 25,122
			Victoria (NB)	Green Park	**Piccadilly** (EB)	23.0	28.7		
2	Euston	St. James's Park	**Victoria** (SB)	Victoria	**Circle** **District** (WB)	18.8	42.8	437	22,379 / 22,968
			Northern (SB)	Embankment	**Circle** **District** (WB)	22.1	57.2		
3	Victoria	Liverpool Street	**Victoria** (NB)	Oxford Circus	**Central** (EB)	25.0	48.1	557	36,262 / 36,668
			Circle (EB)	–	–	33.3	51.9		
4	Angel	Waterloo	**Northern** (SB)	Bank/Monument	**Waterloo & City** (SB)	25.2	42.9	77	14,419 / 14,637
			Northern (SB)	London Bridge	**Jubilee** (WB)	26.9	13.0		
			Northern (NB)	Euston	**Northern** (SB)	29.8	44.2		
5	Liverpool Street	Green Park	**Central** (WB)	Oxford Circus	**Victoria** (SB)	21.5	71.9	196	17,102 / 17,423
			Central (WB)	Holborn	**Piccadilly** (WB)	26.3	17.9		
			Central (WB)	Bond Street	**Jubilee** (SB)	27.1	10.2		
6	Euston	South Kensington	**Victoria** (SB)	Victoria	**Circle** **District** (WB)	22.4	57.4	209	8,116 / 8,277 (192 days)
			Victoria (SB)	Green Park	**Piccadilly** (WB)	26.2	21.1		
			Northern (SB)	Leicester Square	**Piccadilly** (WB)	28.4	21.1		
			Northern (SB)	Embankment	**Circle** **District** (WB)	29.7	0.4		
7	Victoria	Waterloo	**Circle** **District** (EB)	Embankment	**Bakerloo** (SB)	20.9	15.3 [1]	386	7,935 / 8,140 (162 days)
			Circle **District** (EB)	Embankment	**Northern** (SB)	18.1			
			Circle **District** (EB)	Westminster	**Jubilee** (SB)	15.4	48.2		
			Victoria (NB)	Green Park	**Jubilee** (SB)	16.4	36.5		

[1] In ***Case-7***, according to the RODS result, 15.3% of all the respondents chose to transfer at **Embankment**, without detailing which connecting lines were chosen.

4.3.1 Two alternative routes (*Case-1* – *Case-3*)

This section examines three pairs of O-D stations selected from the LU network. Each of the O-D pairs is connected by two alternative routes, and has its own distinct characteristics. For the first two cases, which are code-named '***Case-1***' and '***Case-2***', respectively, only indirect routes are available. For the third case code-named '***Case-3***', one of its two alternative routes actually offers a direct service, whereby an interchange between transit lines might not be necessary during passengers' travel. Additionally, all passengers in ***Case-1*** may have an only line option for the first journey leg but must choose between alternative transfer stations, while those in ***Case-2*** should make a choice between lines at their origin station. ***Case-3***, by contrast, presents a 'dilemma' for the passengers: whether to choose a direct or an indirect route.

4.3.1.1 *Case-1*: Victoria – Holborn

The abridged Tube map illustrated in **Figure 4.2** below shows the single O-D network, **Victoria – Holborn**, for our first case study. Both of the O-D stations are highlighted with red-shaded circles, ●.

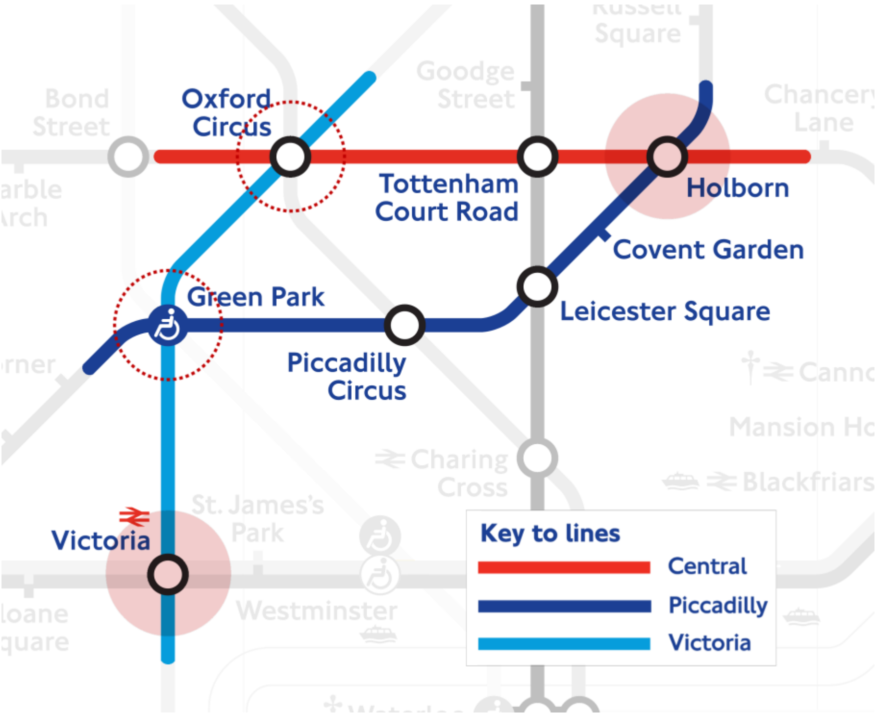
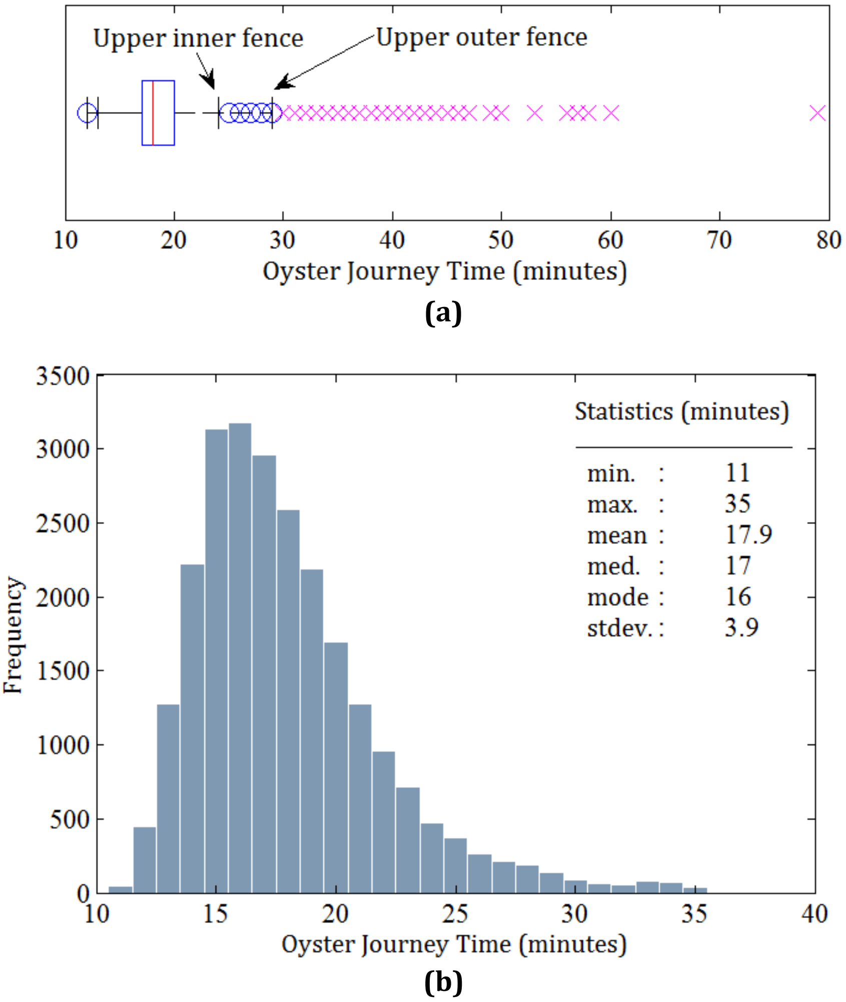


Figure 4.2 The LU network connecting the O-D pair: **Victoria – Holborn**.

In this case, all passengers heading from **Victoria** to **Holborn** would start their journeys by taking a northbound Victoria line train. Then they shall choose to change onto an eastbound train of either the Piccadilly line at **Green Park** or the Central line at **Oxford Circus**. The two alternative interchange stations are each being marked with a red-dotted circle ◌. We could represent the former route by $h=1$ and the latter by $h=2$. This route-choice set was initially identified based only on the Tube map. It also corresponds with the evidence from the RODS about this O-D pair. Moreover, on the rolling basis, the survey result suggested, as shown in **Table 4.2** (*see* p.80), that more than 70% of the passengers travelling on this O-D might choose to make an interchange at **Oxford Circus**, *i.e.* to use the route labelled $h=2$.



(a)

(b)

Figure 4.3 Summary of OJT^{OBS} data for **Victoria – Holborn**:

(a) a box-and-whisker plot of the raw data $(n_0 = 25{,}122)$; and
(b) a histogram of the valid data $(n = 24{,}760)$.

Throughout the observation period from 27th June 2011 (Monday) to 30th March 2012 (Friday), there were 25,122 journeys in total recorded by the Oyster within the time-band of AM Peak (*i.e.* between 07:00 *a.m.* and 10:00 *a.m.*). In this context, we would test whether a mixture model could deliver the same or similar results from the sample set of OJT^{OBS} obtained from the Oyster system.

A graphical summary of the sample data in this case is presented in **Figure 4.3** (*see previous page*). **Figure 4.3(a)** provides a box-and-whisker plot of the entire data set. The red bar on the rectangle 'box' (bordered in blue) marks the median of the sample. The left and right edges of the blue box indicate the 25th and the 75th quartiles, respectively, which are also referred to as the lower and upper fences of the data (*cf.* **Section 4.2.1**); and the box width (of the horizontal side) showing the IQR. The bar located to the left side of the box marks the lower inner fence (*i.e.* 1.5 times IQR less than the 25th quartiles), within which the minimum journey time being observed falls.[1] As the 'whisker' extends to the right of the box, upper boundaries of both the inner and outer fence are marked [2] (*cf.* Freeman *et al.*, 2008, p.41). The magenta crosses, ×, which are beyond the upper outer fence, stand for extreme outliers; but those were all excluded for subsequent analyses. As stated in **Section 4.2.1**, we regarded the data (displayed as blue circles, ○) lying between the upper inner and outer fences to be, albeit suspicious, within the acceptable range of valid data. Finally, 24,760 of OJT^{OBS} were statistically covered by the upper outer fence (with 24,028 inside the upper lower fence).

The frequency distribution of the valid data is shown in **Figure 4.3(b)** (*see previous page*). Given the existence of two alternative routes as described above, the histogram shall resemble a two-component mixture distribution, which ideally would exhibit bimodality; whereas here it appears only a unimodal profile. As such, this might generally imply two possibilities. One is that the two presumptive components might largely overlap, suggesting further, perhaps, that the passengers had similar perception on both routes. In that situation, we

[1] Since the smallest value of journey time observations in a data set is considered valid, the inner lower fence could be ignored for the case studies in this thesis. It is presented in the box-and-whisker plots for demonstration purpose only.

[2] This is slightly different from the standard or conventional representation of box-and-whisker plots where the whisker normally ends at the upper inner fence.

may anticipate that the measures of the central tendency of the two components would be similar. Another possibility is that if in fact there was a difference of centrality between the two components, one of the alternatives shall be weighted less while the other must be given a much higher mixture weight. Bearing the conjecture in mind, we conducted parallel testing of GM and LNM models on the same data set of OJT^{OBS}.

For the estimation of the two mixture models, initial values of all the model parameters were also estimated but from K-means clustering method (described in **Section 3.3.3**) with the same sample. The initial estimates are presented in **Table 4.3** below.

Table 4.3 Parameter estimates of GM and LNM models based on OJT^{OBS} data for **Victoria – Holborn**

The initial values and the model parameters were estimated using the K-means clustering and the EM algorithm, respectively. $n = 24,760$.

	GM		**LNM**	
Component-label	$r=1$	$r=2$	$r=1$	$r=2$
Initial values				
η_r^{KMS} (minute)	16.0	21.0	15.7	22.0
σ_r^{KMS} (minute)	1.7	3.4	1.7	3.1
ω_r^{KMS} (%)	64.1	35.9	64.1	35.9
Parameter estimates				
$\hat{\mu}_r$ (minute)	16.6	22.2	16.5	21.3
$\hat{\sigma}_r$ (minute)	2.3	4.5	2.4	4.4
$\hat{\omega}_r$ (%)	75.4	24.6	69.1	30.9

In addition, the presumptive component distributions in each model are labelled, respectively, by $r = 1$ (also ‘Route1’) and $r = 2$ (also ‘Route2’), representing the two alternative routes. They will hereafter be referred to as component-labels.[3]

[3] For illustrative purposes only, the component, whose value of η_r^{KMS} was relatively smaller, was labelled by a smaller real number, whereby the estimates will always be present in ascending order of η_r^{KMS} as the component-labels increases. This will also apply to all the subsequent case studies.

Note in the case of LNM model that we applied K-means clustering to the natural logarithms of OJT^{OBS} data, instead of the data set being originally derived. As a result, it numerically narrowed down the extent of dispersion of the data. The initial values for estimating LNM were thus more statistically centralised. Consequently, the medians (denoted by η_r^{KMS}) and standard deviations (denoted by σ_r^{KMS}) turned out to be slightly different from their counterparts for GM model. At this stage, the clustered sub-datasets for $r=1$ and $r=2$, were mutually exclusive (*cf.* **Section 3.3.3**), where the preliminary sub-dataset being clustered for $r=1$ encompassed all the relatively shorter journey times being around 16 minutes. That sub-dataset should contain a majority of the observations.

Given the initial values, the parameters for both GM and LNM distributions of OJT^{OBS} were then estimated, using the EM algorithm. The estimation results are also presented in **Table 4.3** (*see previous page*). The estimates from both models suggested that roughly 70%-75% of passengers might have actually chosen the quicker route, Route1; while the rest, about 25% to 30%, might have travelled between the O-D by using Route2. This profile showed a close similarity to the RODS results of this O-D. Moreover, compared to σ_r^{KMS}, the increases in $\hat{\sigma}_r \ \forall r$ largely reflect a partial overlap between the two component distributions.

Furthermore, the probabilities that any passenger might have chosen each of the alternative routes, conditional on his/her journey time, are illustrated in **Figure 4.4** (*see next page*), where the dotted and solid curves are related to GM and LNM, respectively. As can be seen from the graphs, if a passenger's journey time was about 20 to 21 minutes, both models would suggest that he/she might have similar or the same preference of both the alternative routes. Route1 had a higher probability of being chosen by faster passengers whose journey times were less than that critical value, while those who spent longer journey time in travelling on this O-D might be more likely to have chosen Route2. What is more, if anyone's journey time was longer than 26 minutes, given GM, or 30 minutes, given LNM, both the mixture models would simply make us believe that the journey time observation should be in no doubt from Route2, though this conjecture might not necessarily be the case in reality.

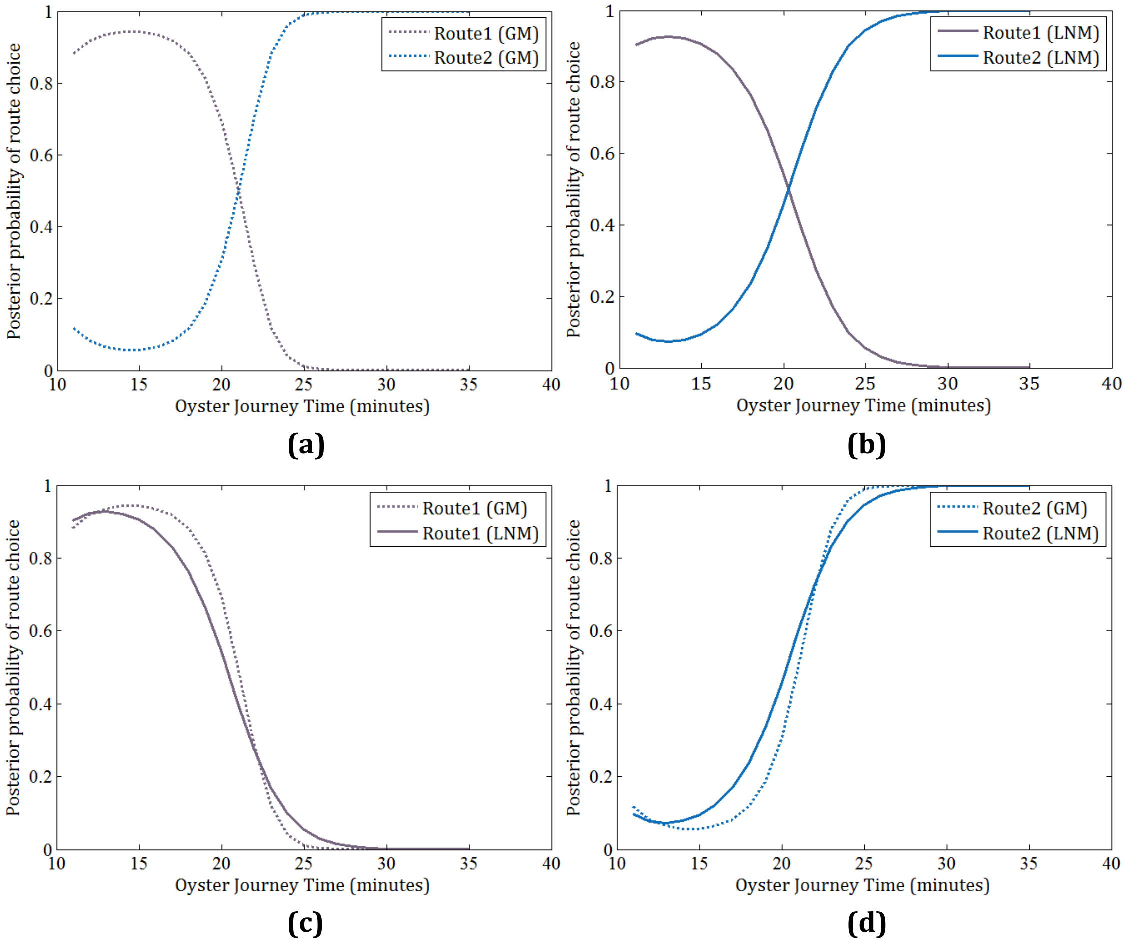


Figure 4.4 Posterior probabilities of route choices given OJT^{OBS} for **Victoria – Holborn** $(n = 24,760)$:

(a) for both routes, based on GM; **(b)** for both routes, based on LNM;
(c) for Route1, based on GM and LNM; and **(d)** for Route2, based on GM and LNM.

Table 4.4 Inferences of proportion of passenger traffic on each alternative route connecting **Victoria** to **Holborn** $(n = 24,760)$

	GM		**LNM**	
Component-label	$r = 1$	$r = 2$	$r = 1$	$r = 2$
$\hat{\omega}_r$ (%)	75.4	24.6	69.0	31.0
$n_r^{\mathrm{INF}_0}$	21,027	3,733	19,751	5,009
$\omega_r^{\mathrm{INF}_0}$ (%)	84.9	15.1	79.8	20.2
n_r^{INF}	18,693	6,067	17,082	7,678
ω_r^{INF} (%)	75.5	24.5	69.0	31.0

Based on the estimates of posterior route-choice probabilities of every individual passenger, both the *naive* and the *effective* inferences of passenger-traffic distributions between the two routes were also made. The results are presented in **Table 4.4** (*see previous page*), together with $\hat{\omega}_r$ for comparisons. As specified in **Section 3.4.1**, $n_r^{\mathrm{INF}_0}$ and $\omega_r^{\mathrm{INF}_0}$ represent the total number and proportion of passengers who chose the r-th route, respectively, from the *naive* inference. This was drawn based on the condition that a passenger might have most likely chosen the route assigned the highest posterior probability. In comparison, n_r^{INF} and ω_r^{INF} were calculated according to the *effective* inference (*see* **Section 3.4.2**). Each of the inferences demonstrate an aggregation of every sampled individual's probabilistic choices between the two alternative routes. The results in this case indicated that ω_r^{INF} was practically consistent with the estimates $\hat{\omega}_r$ from the mixture models for each component, or rather, for each alternative route. To this point, an issue remaining to be solved was to match the estimated components to the real routes. That is, we needed to understand that which specific routes in reality 'Route1' and 'Route2' shall represent.

Table 4.5 Expected journey times of simulated samples for each alternative route connecting Victoria to Holborn

	Calculated average travel time (minutes)	
l' – l''	Victoria – Central	Victoria – Piccadilly
s	Oxford Circus	Green Park
Journey segment		
$t_{l',o}^{\mathrm{ACC}}$	2.7	2.7
$t_{l',o,1}^{\mathrm{WFD}}$ / $t_{l',o,2}^{\mathrm{WFD}}$	0.8 / 2.8	0.8 / 2.8
$t_{l',[o,s]}^{\mathrm{OBT}}$	3.0	1.0
$t_{[l',l''],s}^{\mathrm{ICT}}$	3.3	3.7
$t_{l'',s,1}^{\mathrm{ICW}}$ / $t_{l'',s,2}^{\mathrm{ICW}}$	1.3 / 3.6	1.1 / 3.5
$t_{l'',[s,d]}^{\mathrm{OBT}}$	3.0	6.0
$t_{l'',d}^{\mathrm{EGR}}$	2.8	4.5
Route-label	$h=1$	$h=2$
Total average		
$t_h(1,1)$	16.9	19.9
$t_h(2,1)$	18.9	21.9
$t_h(1,2)$	19.3	23.3
$t_h(2,2)$	21.3	24.3
t_h^{REF}	19.6	23.0

Following the computation procedure of route-specific average journey time, as demonstrated in **Section 3.5.1**, t_h^{REF} for both of the alternative routes on this O-D were calculated using the AEI data, and adjusted with the stations layout data; and the results are presented in **Table 4.5** (*see previous page*). It showed that t_1^{REF} and t_2^{REF} were clearly distinguishable between the two routes.

Further to the calculation of t_h^{REF}, we shall then made a sequence of comparisons in order to find out what the route-labels mean. We compared between the estimated means (*see also* **Table 4.3**, p.84) and the average journey times for each alternative route, as well as between the estimated mixture weights (including proportions of passenger traffic; *see also* **Table 4.4**, p.86) and the RODS results. All the information for such comparisons is summarised in **Table 4.6** below.

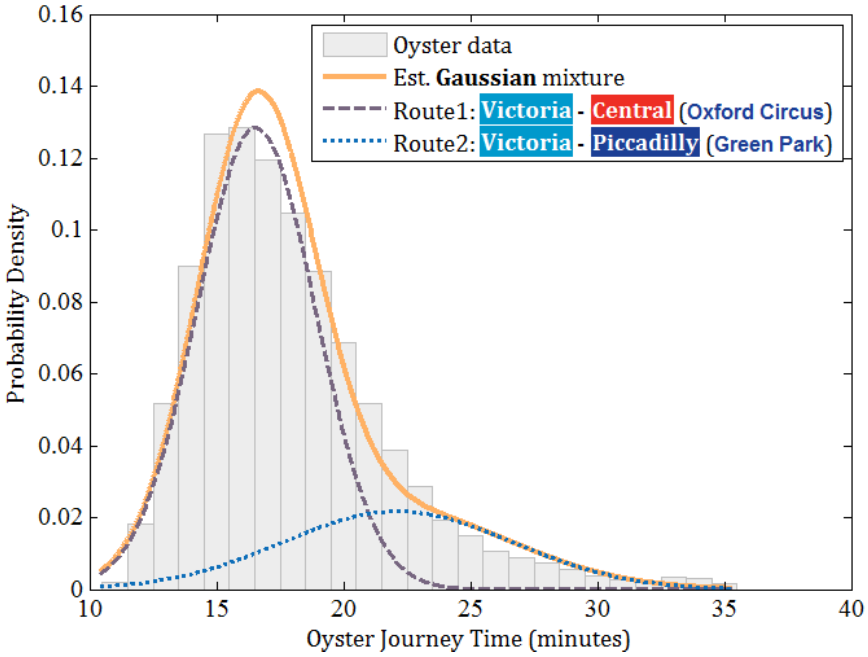
Table 4.6 Matching the estimated mixture components with the real-world routes for **Victoria – Holborn**

		r **matches** h	
Component-label r		$r=1$	$r=2$
Journey time (minutes)			
$\hat{\mu}_r$	GM	16.6	22.2
	LNM	16.5	21.3
t_h^{REF} $(\hat{\sigma}_h^{\mathrm{SEM}})$		19.1 (0.9)	22.1 (0.9)
CI for h	95% CL	[16.3, 21.9]	[19.3, 25.0]
Traffic distribution (%)			
$\hat{\omega}_r$	GM	75.4	24.6
	LNM	69.1	30.9
ω_h^{ROD} (n_h^{ROD})	AM Peak	71.3 (375)	28.7 (151)
	A weekday	66.2 (612)	33.8 (313)
Route-label h		$h=1$	$h=2$
		Victoria – **Central** **Oxford Circus**	**Victoria** – **Piccadilly** **Green Park**

Take the estimates from GM model for example. In line with **Table 4.5** (*see previous page*), t_1^{REF} and t_2^{REF} denotes, respectively, the calculated average journey times of the routes, "**Victoria** – **Central**, via **Oxford Circus** station" and

"Victoria – Piccadilly, via Green Park station". At first glance, it is noticeable that $t_1^{\mathrm{REF}} < t_2^{\mathrm{REF}}$, and also that $\hat{\mu}_1 < \hat{\mu}_2$. Additionally, $t_2^{\mathrm{REF}} \approx \hat{\mu}_2$. Such an outcome would largely imply that Route1 (*i.e.* $r = 1$) might correspond to the former route, which was labelled by $h = 1$; and similarly, Route2 (*i.e.* $r = 2$) could be regarded as the alternative labelled by $h = 2$. Although $\hat{\mu}_1 < t_1^{\mathrm{REF}}$, it still fell within the 95% CI of t_1^{REF}, given $t_h(\phi, \psi)$ $\forall \phi = 1, 2$ and $\forall \psi = 1, 2$ (*cf.* **Section 3.5.1**). If all the conjectures above were true, $r = 1$ must be equivalent to $h = 1$; and $r = 2$ must also be the same as $h = 2$. A strong supporting evidence to this supposition was that $\hat{\omega}_r$ $\forall r = 1, 2$ showed a close similarity to the corresponding RODS results, ω_h^{ROD} $\forall h = 1, 2$. According to the criteria laid down in **Section 3.5.2**, it could then be concluded in this case that Route1 was extremely likely the route, $h = 1$, and Route2 the other, $h = 2$. With regard to the estimates of LNM model, we could derive the same conclusion from **Table 4.6** (*see previous page*).

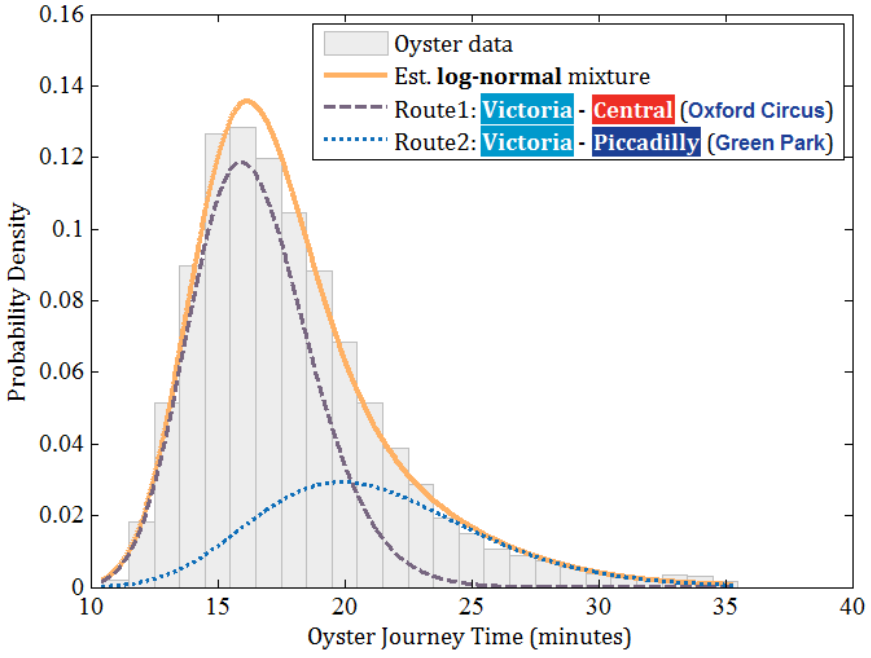
Based on all the results above, **Figure 4.5** below (*and also next page*) delineates a graphical view of the estimated the PDFs of the GM and LNM distributions as well as all the components.



(a)

Figure 4.5 Estimated mixture distributions, and weighted components thereof, of *OJT* for Victoria – Holborn ($n = 24,760$) :

(a) estimated GM model; and
(b) estimated LNM model (*see next page*).



(b)

Figure 4.5 (*Continued.*)

Evidently, both the GM and LNM models could be deemed to perform very well on this O-D; and they both were eligible in terms of the judging criterion for estimated parameters.

With the aid of *gof* , the indicator for goodness of fit calculated by formula (3-46) (*see* p.63), we compared the statistical performance of the two models by repeating the computation 1,000 times; and the results are presented in **Table 4.7** below.

Table 4.7 Goodness-of-fit test result for **Victoria – Holborn**

The calculation of *gof* was repeated 1,000 times for each model.

	GM	**LNM**
Rate of obtaining lower *gof* (%)	0.3	99.7
Average *gof*	0.109	0.096

In this case, LNM was deemed to be more suitable, due to its lower average *gof* and a far higher rate of gaining a lower value of *gof* .

4.3.1.2 *Case-2*: Euston – St. James's Park

For the second case study, code-named ***Case-2***, we also scrutinise an O-D pair with two indirect routes: **Euston – St. James's Park**. Its network is illustrated in **Figure 4.6** below.

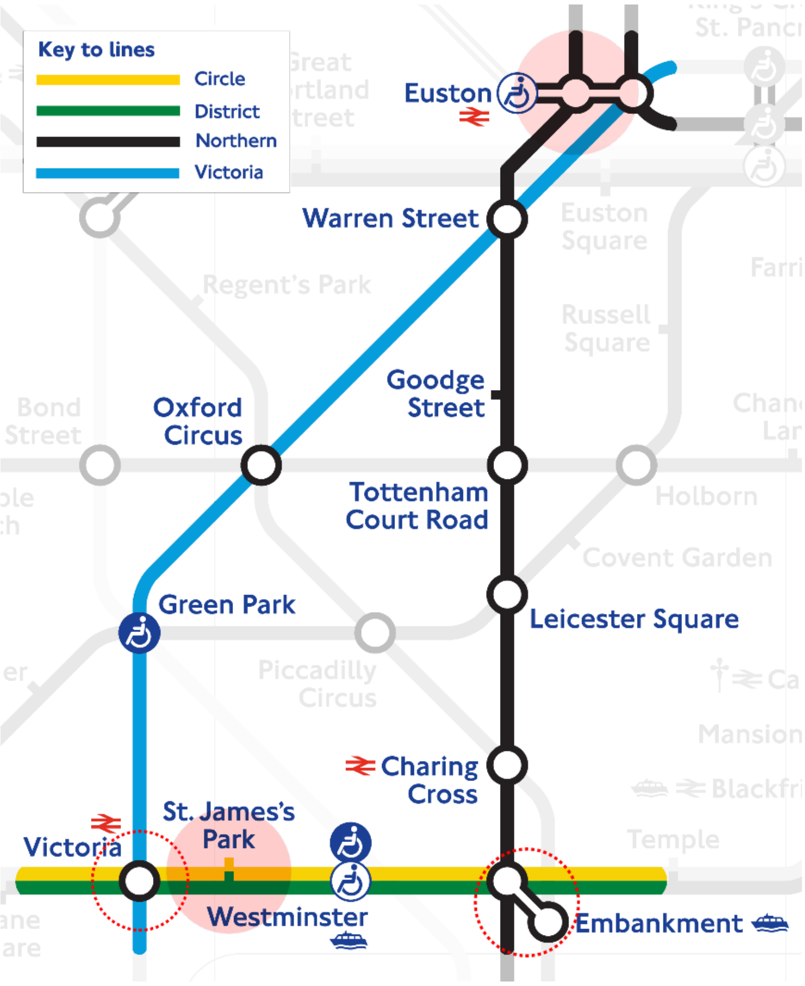
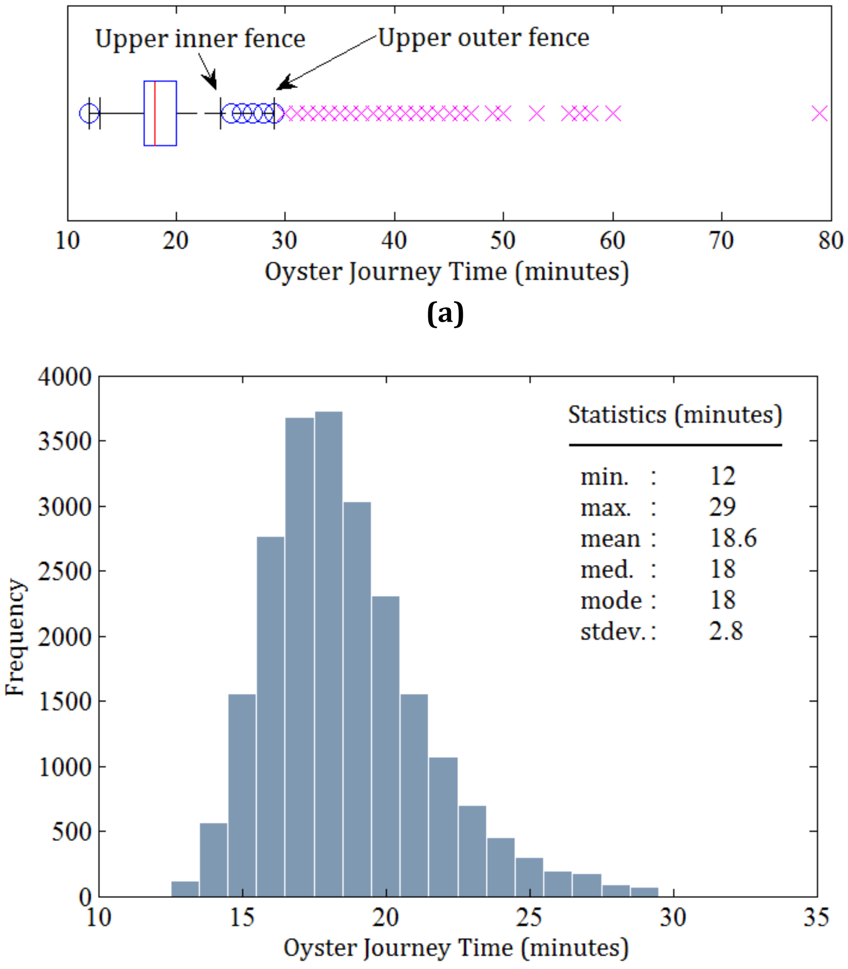


Figure 4.6 The LU network connecting the O-D pair: **Euston – St. James's Park**.

In contrast with ***Case-1***, all passengers travelling on this O-D must firstly choose between two different lines at **Euston**, the origin station. They will have to make a decision whether to take the **Victoria** line (southbound) or the **Northern** line (southbound) for their first journey leg. On the second journey leg, those who take the **Victoria** line will transfer at **Victoria** station to either the **District** line (eastbound) or the **Circle** line (eastbound).[4] In a similar way to ***Case-1***, we let this route be labelled $h = 1$. All the other passengers, who choose the **Northern** line at the origin and alight at **Embankment** station, will then have to transfer to a westbound train on either of the two common lines. This latter route was labelled $h = 2$. The data of OJT^{OBS} collected during the period of observation for this O-D is summarised in **Figure 4.7** below, with **Figure 4.7(a)** describing the original data set and **Figure 4.7(b)** depicting a histogram of all the valid data for use.



(a)

(b)

Figure 4.7 Summary of OJT^{OBS} data for **Euston – St. James's Park**:

(a) a box-and-whisker plot of the raw data $(n_0 = 22,968)$; and
(b) a histogram of the valid data $(n = 22,379)$.

[4] Within Zone-1 of the LU network, the operational routes of the **District** line and the **Circle** line are parallel and share the same platform at the stations they stop along the way.

What was similar to ***Case-1*** was that the shape of the histogram shown in **Figure 4.7(b)** (*see previous page*) still did not demonstrate distinct characteristics of a bimodal distribution, despite the availability of two alternative routes. Again, this might be due to either a substantial overlap between the journey time distributions of the two routes or a significant weighting disparity between the two in the mixture distribution (*cf.* **Section 4.3.1**, p.78). Notwithstanding such unimodality, we applied K-means clustering method to the valid OJT^{OBS} data to gain two sets of initial values for the estimation of GM and LNM models, respectively.

The estimation results of the initial values as well as the mixture model parameters are presented in **Table 4.8** below.

Table 4.8 Parameter estimates of GM and LNM models based on OJT^{OBS} data for **Euston – St. James's Park**

The initial values and the model parameters were estimated using the K-means clustering and the EM algorithm, respectively. $n = 22,379$.

	GM		LNM	
Component-label	$r=1$	$r=2$	$r=1$	$r=2$
Initial values				
η_r^{KMS} (minute)	17.0	20.0	17.0	20.1
σ_r^{KMS} (minute)	1.2	2.2	1.3	2.0
ω_r^{KMS} (%)	55.5	44.5	55.5	44.5
Parameter estimates				
$\hat{\mu}_r$ (minute)	17.6	21.2	17.8	22.3
$\hat{\sigma}_r$ (minute)	1.8	3.0	2.0	2.7
$\hat{\omega}_r$ (%)	72.4	27.6	82.8	17.2

We could see that the estimates of the component means (denoted by $\hat{\mu}_r$) did not differ very much from their initial values, while the standard deviations (denoted by $\hat{\sigma}_r$) and the mixture weights (denoted by $\hat{\omega}_r$) changed dramatically, which accounted for the expected overlap between the mixture components. On the other hand, it is noticeable that $\hat{\mu}_1 < \hat{\mu}_2$ and $\hat{\omega}_1 \gg \hat{\omega}_2$. This again implied that much more passengers might have taken the faster route, which was similarly labelled by $r=1$ and referred to as Route1. Correspondingly, the slower route, labelled by $r = 2$, was referred to as Route2. The LNM model even suggested a relatively more lopsided situation that Route1 might have taken more than 80% of the passenger traffic between this O-D.

In line with the process of model testing as demonstrated in ***Case-1***, the distributions of the estimated posterior choice probabilities of passengers are illustrated in **Figure 4.8** below.

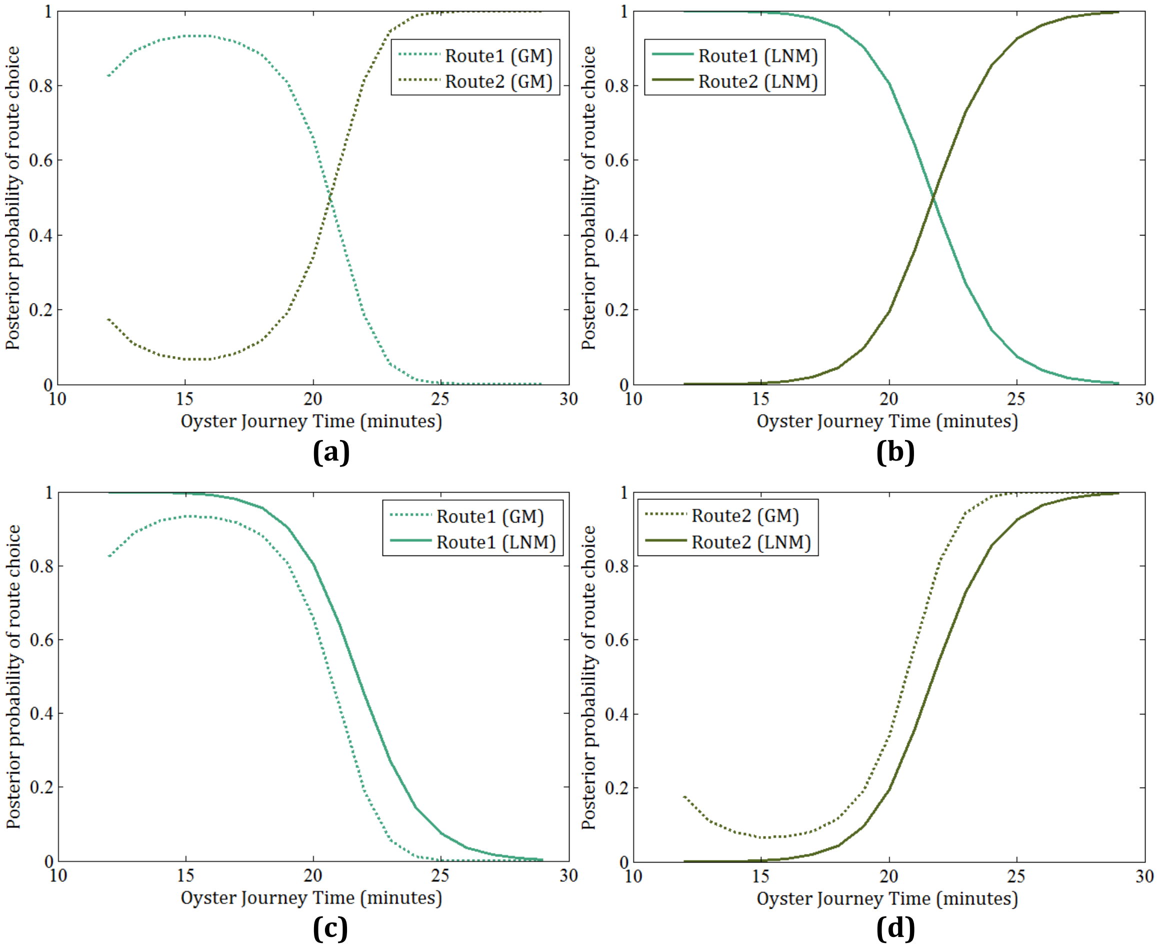


Figure 4.8 Posterior probabilities of route choices given OJT^{OBS} for **Euston – St. James's Park** ($n = 22{,}379$) :

(a) for both routes, based on GM; **(b)** for both routes, based on LNM;
(c) for Route1, based on GM and LNM; and **(d)** for Route2, based on GM and LNM.

On the basis of that, the inference of passenger-flow proportions between the alternative routes on this O-D were calculated, and the results are presented in **Table 4.9** (*see next page*). Furthermore, the computation of the route-specific average journey times are shown in **Table 4.10** (*see next page*).

Table 4.9 Inferences of proportion of passenger traffic on each alternative route connecting **Euston** to **St. James's Park** ($n = 22,379$)

	GM		LNM	
Component-label	$r = 1$	$r = 2$	$r = 1$	$r = 2$
$\hat{\omega}_r$ (%)	72.4	27.6	82.8	17.2
$n_r^{\mathrm{INF}_0}$	17,766	4,613	19,322	3,057
$\omega_r^{\mathrm{INF}_0}$ (%)	79.4	20.6	86.3	13.7
n_r^{INF}	16,152	6,227	18,512	3,867
ω_r^{INF} (%)	72.2	27.8	82.7	17.3

Table 4.10 Expected journey times of simulated samples for each alternative route connecting **Euston** to **St. James's Park**

	Calculated average travel time (minutes)	
l' – l'' s	Victoria – Circle/District Victoria	Northern – Circle/District Embankment
Journey segment		
$t_{l',o}^{\mathrm{ACC}}$	4.0	2.4
$t_{l',o,1}^{\mathrm{WFD}}$ / $t_{l',o,2}^{\mathrm{WFD}}$	0.6 / 2.6	1.8 / 5.1
$t_{l',[o,s]}^{\mathrm{OBT}}$	7.0	8.0
$t_{[l',l''],s}^{\mathrm{TIC}}$	2.0	2.2
$t_{l'',s,1}^{\mathrm{WIC}}$ / $t_{l'',s,2}^{\mathrm{WIC}}$	1.6 / 3.8	1.5 / 3.6
$t_{l'',[s,d]}^{\mathrm{OBT}}$	1.0	3.0
$t_{l'',d}^{\mathrm{EGR}}$	0.5	0.5
Route-label	$h = 1$	$h = 2$
Total average		
$t_h(1,1)$	16.7	19.4
$t_h(2,1)$	18.7	22.7
$t_h(1,2)$	18.9	21.5
$t_h(2,2)$	20.9	24.8
t_h^{REF}	18.8	22.1

Table 4.11 (*see next page*) demonstrates the comparisons between the models' estimates and the survey results in order to interpret the route-labels and to validate those estimates.

Table 4.11 Matching the estimated mixture components with the real-world routes for **Euston – St. James's Park**

		r **matches** h	
Component-label r		$r=1$	$r=2$
Journey time (minutes)			
$\hat{\mu}_r$	GM	17.6	21.2
	LNM	17.8	22.3
t_h^{REF} $(\hat{\sigma}_h^{\text{SEM}})$		18.8 (0.9)	22.1 (1.1)
CI for h	95% CL	[16.1, 21.5]	[18.5, 25.7]
Traffic distribution (%)			
$\hat{\omega}_r$	GM	72.4	27.6
	LNM	82.8	17.3
ω_h^{ROD} (n_h^{ROD})	AM Peak	42.8 (187)	57.2 (250)
	A weekday	46.4 (225)	53.6 (260)
Route-label h		$h=1$	$h=2$
		Victoria – Circle/District Victoria	Northern – Circle/District Embankment

By comparing $\hat{\mu}_r$ with t_h^{REF} for each alternative route in this case, the situation was also very similar to that in ***Case-1***. In view of the fact that $\hat{\mu}_1$ and $\hat{\mu}_2$ fell within the 95% CI of t_1^{REF} and t_2^{REF}, respectively, we could preliminarily match Route1 (*i.e.* $r=1$) to the route that goes through **Victoria** station (*i.e.* $h=1$: **Victoria** – **Circle**/**District**); and also regard Route2 (*i.e.* $r=2$) as the alternative route via **Embankment** station (*i.e.* $h=2$: **Northern** – **Circle**/**District**).

However, in this case, there was an issue on validating the mixture models with the RODS results. Take the estimates from GM model for example. According to the RODS, ω_1^{ROD} (= 42.8%) was slightly smaller than ω_2^{ROD} (= 57.2%), which suggested that the quicker route shared less of the total passenger traffic than the slower transit service. Given the existing information, we could not find out the reason to this point; but we might doubt that the ω_h^{ROD} in this case was not quite credible. On the other hand, $\hat{\omega}_r$ derived from either GM or LNM model in this case seemed to make more sense, as a much larger proportion of passenger traffic was assigned to Route1 for a quicker service. The following three possibilities might account for puzzled situation: (*see next page*)

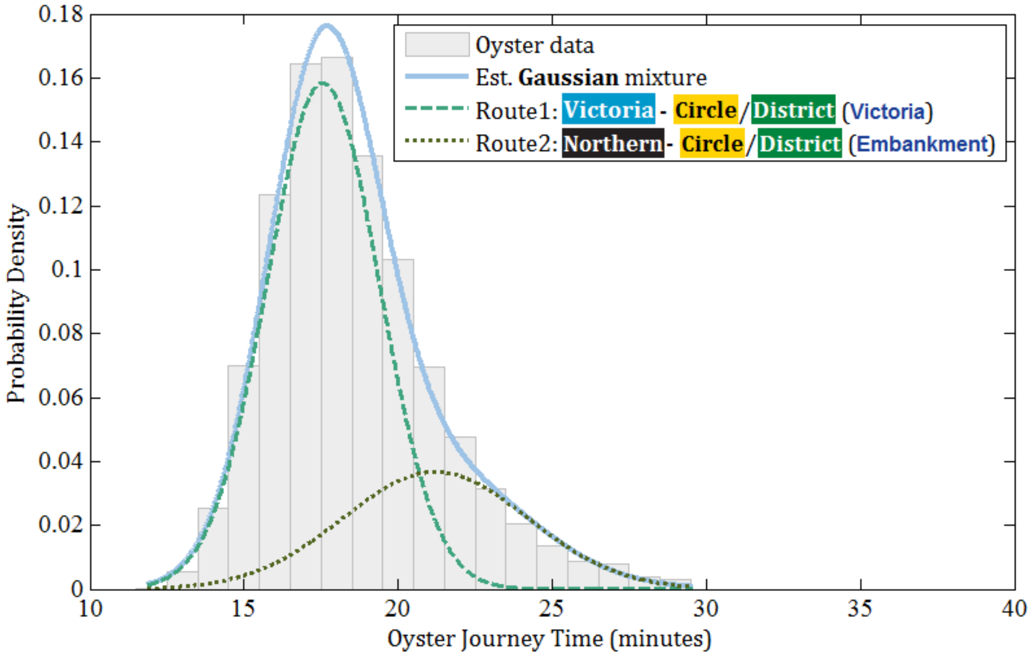
i. the RODS results for this case might not be accurate mainly because they were aggregated on a rolling basis, notwithstanding the presence of a large sample;

ii. some attributes of the slower route were possibly more preferable to passengers[5], *e.g.* shorter walking distance and wait time; and/or

iii. neither the GM nor LNM model were suitable for this case, but other models should be further tested.

Despite all this, we should accept both GM and LNM models in this case, given their sensible estimates.

Figure 4.9 below (*and also next page*) shows the estimated PDFs of the GM and LNM distributions, which showed that both models could fit the journey time data very well. Yet, the difference between the two was not as immediately noticeable as that in ***Case-1***.

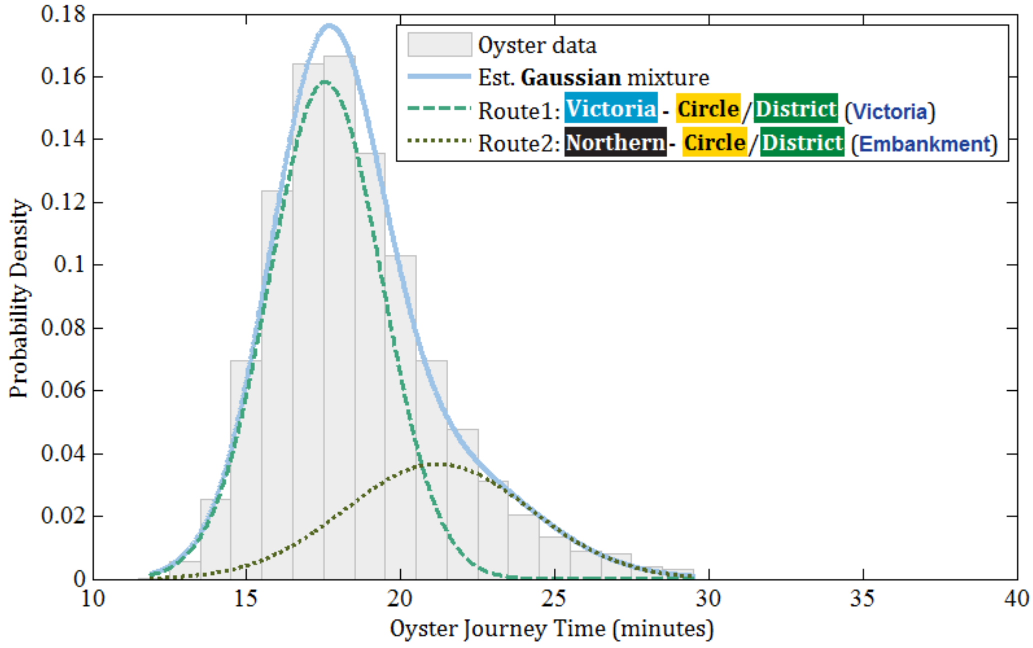


(a)

Figure 4.9 Estimated mixture distributions, and weighted components thereof, of OJT for **Euston – St. James's Park** $(n = 22,379)$:

(a) estimated GM model; and
(b) estimated LNM model (*see next page*).

[5] This will be further examined in **Chapter 6.**



(b)

Figure 4.9 (*Continued.*)

Furthermore, as shown in **Table 4.12** below, the result of the goodness-of-fit test suggested that the LNM model should be more suitable in this case, as it gave a lower average *gof* as well as has a much greater rate of gaining a better fit.

Table 4.12 Goodness-of-fit test result for **Euston – St. James's Park**

The calculation of *gof* was repeated 1,000 times for each model.

	GM	**LNM**
Rate of obtaining lower *gof* (%)	19.1	80.9
Average *gof*	0.115	0.113

4.3.1.3 *Case-3*: Victoria – Liverpool Street

The last of the three cases involving two alternative routes being studied was the O-D pair: **Victoria – Liverpool Street**. Its network is illustrated in **Figure 4.10** (*see next page*), where both the O-D stations are highlighted with green-shaded circles.

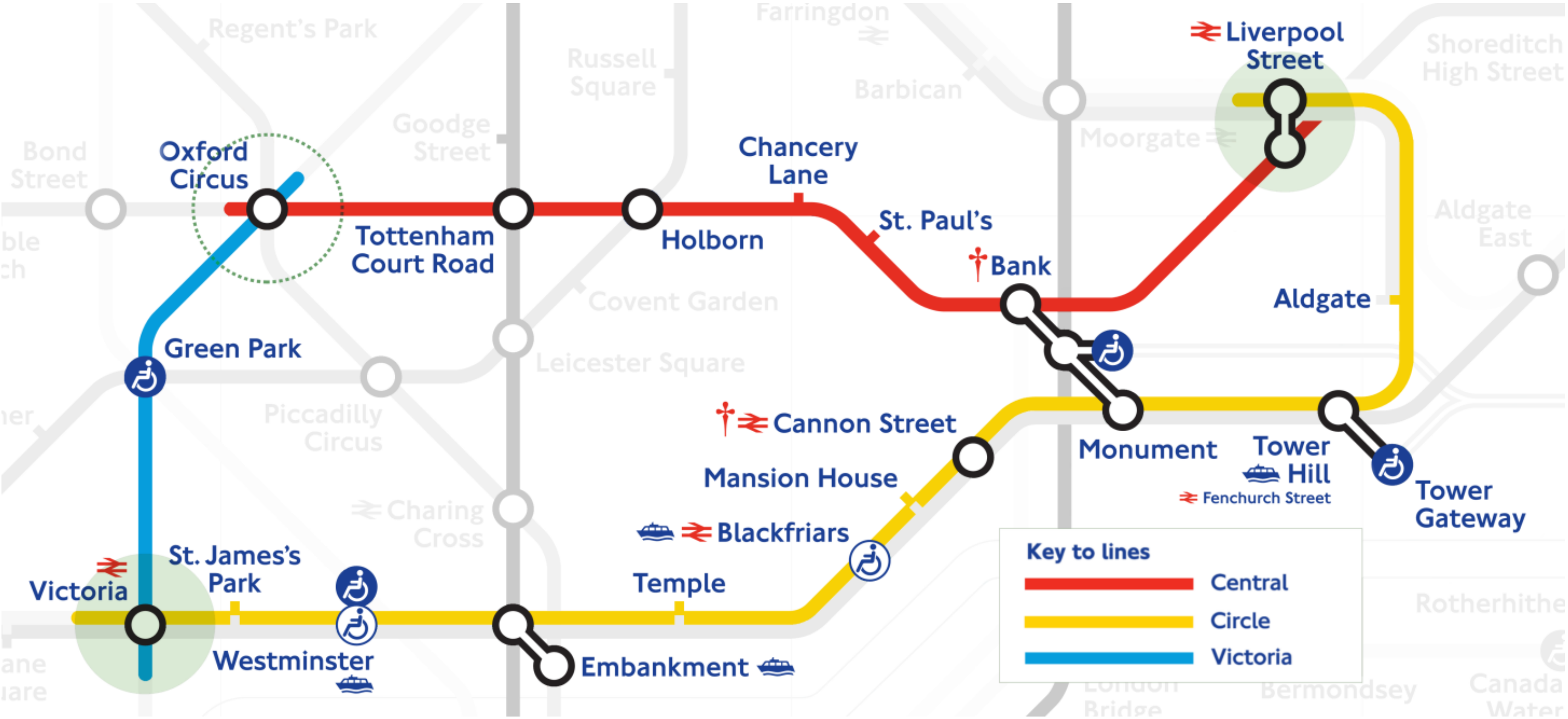
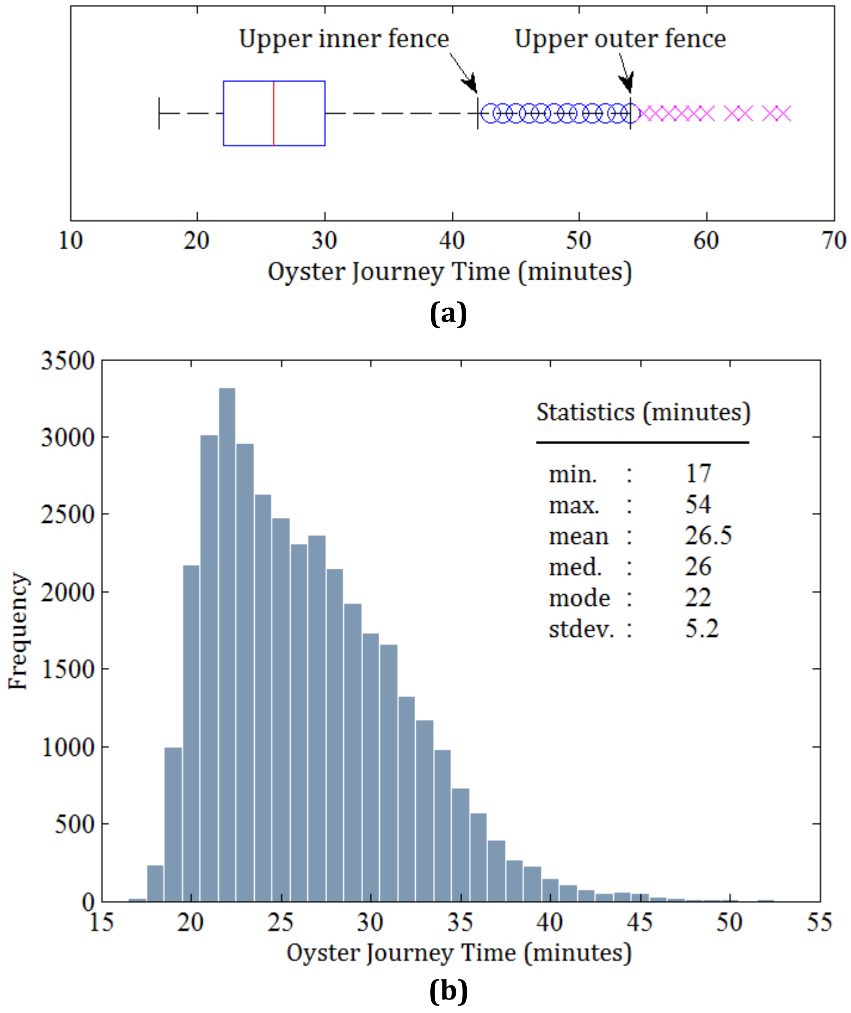


Figure 4.10 The LU network connecting the O-D pair: Victoria – Liverpool Street.

As mentioned earlier, this O-D provides travellers with both direct and indirect services. More specifically, all the passengers at the origin station, Victoria, could use either the Circle line (eastbound) serving as a direct route, labelled $h = 1$; or choose to take the Victoria line (northbound) first but would then transfer to the Central line (eastbound) at Oxford Circus, labelled by $h = 2$.

It is noted that anyone choosing the **Circle** line may also jump to the **Central** line at the station of **Monument/Bank** complex (or simply **Bank**). However, that route was not considered in this case, because of its overlong connection paths for interchange; and we also assumed that any passenger who had already chosen a direct service would not usually change to an indirect service during his/her journey. Moreover, as reported by RODS, this route was rarely used in practice.

Figure 4.11 below summarises the OJT^{OBS} data to be modelled in this case. Unlike ***Case-1*** and ***Case-2***, the frequency distribution of OJT^{OBS} for this O-D, as shown in **Figure 4.11(b)**, appeared to be a bimodal profile, with the major and minor modes being 22 and 27 minutes, respectively, though the minor one was less obvious.



(a)

(b)

Figure 4.11 Summary of OJT^{OBS} data for **Victoria – Liverpool Street**:

(a) a box-and-whisker plot of the raw data $(n_0 = 36,668)$; and
(b) a histogram of the valid data $(n = 36,262)$.

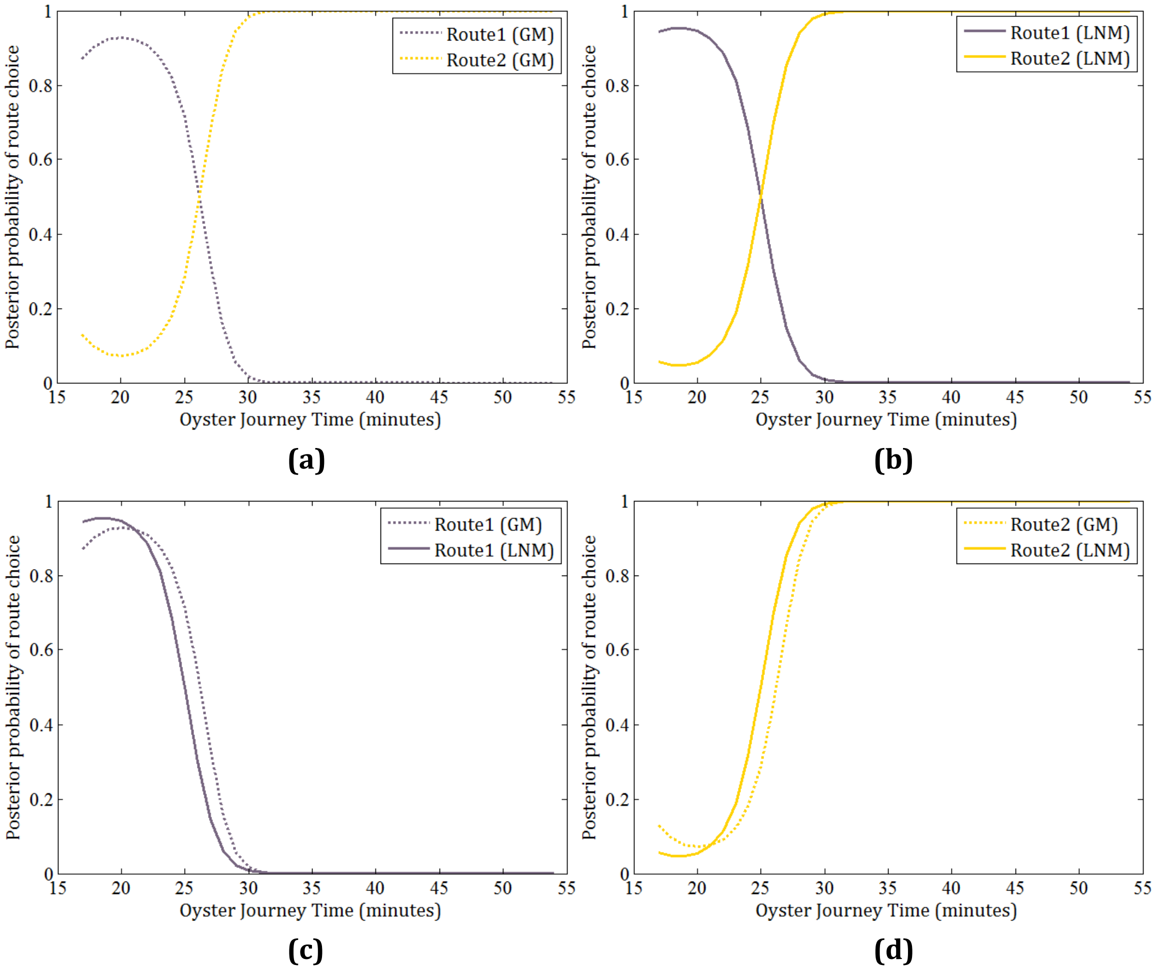
Based on the sample of 36,262 individuals' journey times, the initial values as well as the estimates of mixture model parameters were obtained, which are presented in **Table 4.13** below.

Table 4.13 Parameter estimates of GM and LNM models based on OJT^{OBS} data for **Victoria – Liverpool Street**

The initial values and the model parameters were estimated using the *K*-means clustering and the EM algorithm, respectively. $n = 36,262$.

	GM		**LNM**	
Component-label	$r=1$	$r=2$	$r=1$	$r=2$
Initial values				
η_r^{KMS} (minute)	23.0	30.0	22.1	30.2
σ_r^{KMS} (minute)	2.1	3.9	1.8	3.8
ω_r^{KMS} (%)	55.6	44.5	49.2	50.8
Parameter estimates				
$\hat{\mu}_r$ (minute)	22.8	30.3	22.3	29.7
$\hat{\sigma}_r$ (minute)	2.3	4.6	2.1	4.5
$\hat{\omega}_r$ (%)	50.6	49.4	43.6	56.4

We discerned that both the major and minor modes shown in the frequency distribution had been roughly captured and retrieved by the estimation of the two mixture models, where $\hat{\mu}_1$ and $\hat{\mu}_2$ were around 22.5 and 30.0 minutes, respectively. By comparison with the previous two O-D cases, a significant difference in the estimation results for this case was reflected in the estimates of mixture weights, $\hat{\omega}_r$. So far, all the testing mixture models (in the previous two cases) suggested that the traffic volume on faster routes would be higher than the slower alternative. However, notwithstanding a large gap between $\hat{\mu}_1$ and $\hat{\mu}_2$ in this case, $\hat{\omega}_1$ was almost the same as $\hat{\omega}_2$ for the GM model; and the situation was even the opposite given the LNM model, namely, $\hat{\omega}_1 < \hat{\omega}_2$ while $\hat{\mu}_1 < \hat{\mu}_2$. The LNM estimates then suggested that a larger proportion of passengers tended to pay nearly eight minutes more for the slower service. The most likely reason might be that more travellers would be inclined to avoid the interchange when making a journey.[1] **Figure 4.12** below depicts the estimates of posterior probabilities of the passengers' route choices.



(a) **(b)** **(c)** **(d)**

Figure 4.12 Posterior probabilities of route choices given OJT^{OBS} for **Victoria – Liverpool Street** $(n = 36,262)$:

(a) for both routes, based on GM; **(b)** for both routes, based on LNM;
(c) for Route1, based on GM and LNM; and **(d)** for Route2, based on GM and LNM.

Generally, what was happening on this point was very much similar to ***Case-2***. If passengers' journey times were shorter than the sample mean of the given data set for this O-D, their choice probabilities for Route1 (*i.e.* the faster route) were believed to be higher than for Route2; whereas for those who spent more than about 26 or 27 minutes, the probability of choosing the slower route would

[1] This will be further examined in **Chapter 6.**

become higher. This could be reasonable given the fact that the minor mode was about 27 minutes and the estimated mean for Route2 was greater than that.

The inference of passenger-traffic distribution on this O-D is presented in **Table 4.14** (*see next page*). For each of the two mixture models, the traffic share $\omega_r^{\mathrm{INF}_0}$ (that was based on the *naive* inference) showed the same trend as ω_r^{INF} (that was based on the *effective* inference), notwithstanding the difference of implications between the two models.

Similar to the previous two cases, the average travel times for each alternative route as well as for each of their journey segments are presented in **Table 4.15** (*see next page*); and the comparison of the mixture models is set out in **Table 4.16** (*see* p.105).

Now we also take the GM model as an example to demonstrate the way of matching a component-label to a real route. From the information in **Table 4.15**, we could see that $t_1^{\mathrm{REF}} > \hat{\mu}_1$, and also that $t_2^{\mathrm{REF}} > \hat{\mu}_2$. This was mainly because the calculation of t_h^{REF} $\forall h = 1, 2$ considered equally the four distinct circumstances specified in **Section 3.5.1** (*see* p.62) However, it could also be noticed that $\hat{\mu}_1$ was close to $t_1(1,1)$ of the indirect route, while $\hat{\mu}_2$ approximated $t_2(1,1)$ of the direct route. Additionally, as shown in **Table 4.16**, $\hat{\mu}_1$ and $\hat{\mu}_2$ fell within the 95% CI of t_1^{REF} and t_2^{REF}, respectively. In view of these evidence, Route1 and Route2 could be deemed as the indirect service and the direct service, respectively. This should then suggest that most passengers travelling on this O-D could successfully board the firstly arriving train at both the origin and interchange stations (*cf.* circumstance II-i, *see also* p.62).

Table 4.14 Inferences of proportion of passenger traffic on each alternative route connecting Victoria to Liverpool Street ($n = 36,262$)

	GM		LNM	
Component-label	$r=1$	$r=2$	$r=1$	$r=2$
$\hat{\omega}_r$ (%)	50.6	49.4	43.6	56.4
$n_r^{\mathrm{INF}_0}$	20,145	16,117	17,835	18,427
$\omega_r^{\mathrm{INF}_0}$ (%)	55.6	44.4	49.2	50.8
n_r^{INF}	18,077	18,185	15,617	20,645
ω_r^{INF} (%)	49.9	50.1	43.1	56.9

Table 4.15 Expected journey times of simulated samples for each alternative route connecting Victoria to Liverpool Street

	Calculated average travel time (minutes)	
$l' - l''$ s	Victoria – Central Oxford Circus	Circle –
Journey segment		
$t^{\mathrm{ACC}}_{l',o}$	2.7	1.7
$t^{\mathrm{WFD}}_{l',o,1}$ / $t^{\mathrm{WFD}}_{l',o,2}$	0.8 / 2.8	4.8 / 14.8
$t^{\mathrm{OBT}}_{l',[o,s]}$	3.0	20.0
$t^{\mathrm{ICT}}_{[l',l''],s}$	3.3	-
$t^{\mathrm{ICW}}_{l'',s,1}$ / $t^{\mathrm{ICW}}_{l'',s,2}$	1.3 / 3.6	-
$t^{\mathrm{OBT}}_{l'',[s,d]}$	10.0	-
$t^{\mathrm{EGR}}_{l'',d}$	1.7	1.8
Route-label	$h=1$	$h=2$
Total average		
$t_h(1,1)$	22.8	28.4
$t_h(2,1)$	24.8	38.3
$t_h(1,2)$	25.2	-
$t_h(2,2)$	27.2	-
t^{REF}_h	25.0	33.3

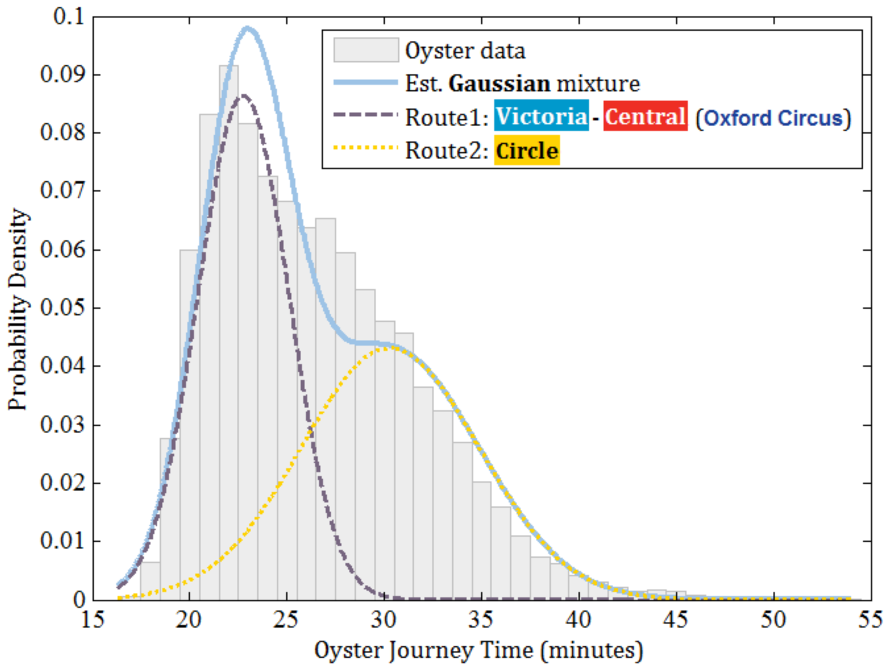
In regard to the proportions of passenger traffic, neither the GM estimates nor the corresponding *naive* inference were consistent with the RODS result; the GM model might lead to a contradictory conclusion on the traffic shared between Route1 and Route2. By reference to $\omega^{\mathrm{ROD}}_h \quad \forall h$ for a typical whole day on this O-D, which was based on a much larger sample (*see also* **Table 4.16**), it showed that a majority of passengers would rather spend a relatively longer journey time than make an interchange for a quicker transit service.

From a combined view of the information in both **Table 4.15** and **Table 4.16**, we shall conclude that Route1 (*i.e.* $r=1$) was most likely the indirect route (*i.e.* $h=1$); and Route2 (*i.e.* $r=2$) must be the direct service (*i.e.* $h=2$). And we shall also consider both models to be eligible.

Figure 4.13 (*see next page*) shows the estimated PDFs of the GM and LNM distributions.

Table 4.16 Matching the estimated mixture components with the real-world routes for **Victoria – Liverpool Street**

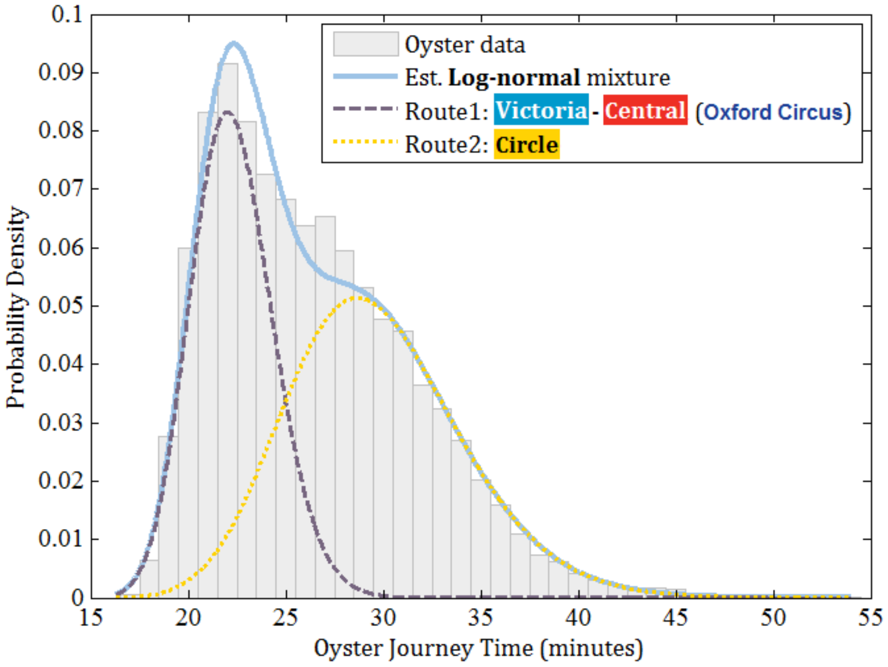
		r **matches** h	
Component-label r		$r=1$	$r=2$
Journey time (minutes)			
$\hat{\mu}_r$	GM	22.8	30.3
	LNM	22.3	29.7
t_h^{REF} $(\hat{\sigma}_h^{\text{SEM}})$		25.0 (0.9)	33.3 (2.9)
CI for h	95% CL	[22.2, 27.8]	[24.2, 42.5]
Traffic distribution (%)			
$\hat{\omega}_r$	GM	50.6	49.4
	LNM	43.8	56.4
ω_h^{ROD} (n_h^{ROD})	AM Peak	48.1 (268)	51.9 (289)
	A weekday	38.9 (1,042)	61.1 (1,634)
Route-label h		$h=1$	$h=2$
		Victoria – **Central** **Oxford Circus**	**Circle** –



(a)

Figure 4.13 Estimated mixture distributions, and weighted components thereof, of *OJT* for **Victoria – Liverpool Street** $(n = 36,262)$:

(a) estimated GM model; and
(b) estimated LNM model (*see next page*).



(b)

Figure 4.13 (*Continued.*)

Table 4.17 Goodness-of-fit test result for **Victoria – Liverpool Street**

The calculation of *gof* was repeated 1,000 times for each model.

	GM	**LNM**
Rate of obtaining lower *gof* (%)	0	100
Average *gof*	0.11	0.07

From **Table 4.17** above, the goodness-of-fit results in this case showed that the LNM could always provide a relatively lower *gof* , of which the average was very close to 0; and compared with the GM model, the LNM model had an absolute better-fit to the sample data.

4.3.2 More than two alternative routes (*Case-4* and *Case-6*)

In this section, we further challenge the applicability of GM and LNM model in the context that more than two alternative routes are available for a given O-D. For this purpose, we selected four typical O-D pairs, where two were for cases of three routes, with each associating with a three-component mixture distribution, and the other two for the cases of four alternative routes, with each associating with a four-component mixture distribution, accordingly. However, here we only describe one case study in each of the two circumstances. This is because the modelling process for three or more components are essentially the same as for the two component case studies in **Section 4.3.1**. The main difference between different case studies lies mostly in the conditions for matching a route-label to a real route.

To start this section, a case study given the availability of three alternative routes is presented in **Section 4.3.2.1**, which is code-named ***Case-4***. Then **Section 4.3.2.2** examines an O-D pair connected by four alternative routes, which is code-named ***Case-6***. For the other remaining two case studies, we present in **Appendix C** only the relevant estimation results, as the basic principles have been demonstrated in the previous section. We code-name the case with three alternative routes ***Case-5*** and that with four alternative routes ***Case-7***.

In the same way as we dealt with the two-route examples, both GM and LNM models were applied to fit the OJT^{OBS} data available for all these four cases, so as to test whether the two standard mixture models could also be suitable. The identification of the route-choice set for each of the four O-D's had also been verified with the RODS results, and are described based on the edited Tube maps presented in the corresponding subsections.

4.3.2.1 A case of three routes (*Case-4*): Angel – Waterloo

This section describes a case study on an O-D pair connected by three alternative routes, where the origin and destination are the stations of **Angel** and **Waterloo**, respectively. The network linking this O-D pair is illustrated in **Figure 4.14** (*see next page*), with both the O-D stations being marked with shaded circles, and the relevant interchange stations being circled with dots.

In this case, all passengers starting their journeys from **Angel** station (shown in the upper right corner of the map) may choose either a northbound or a southbound train of the **Northern** line for the first journey leg.

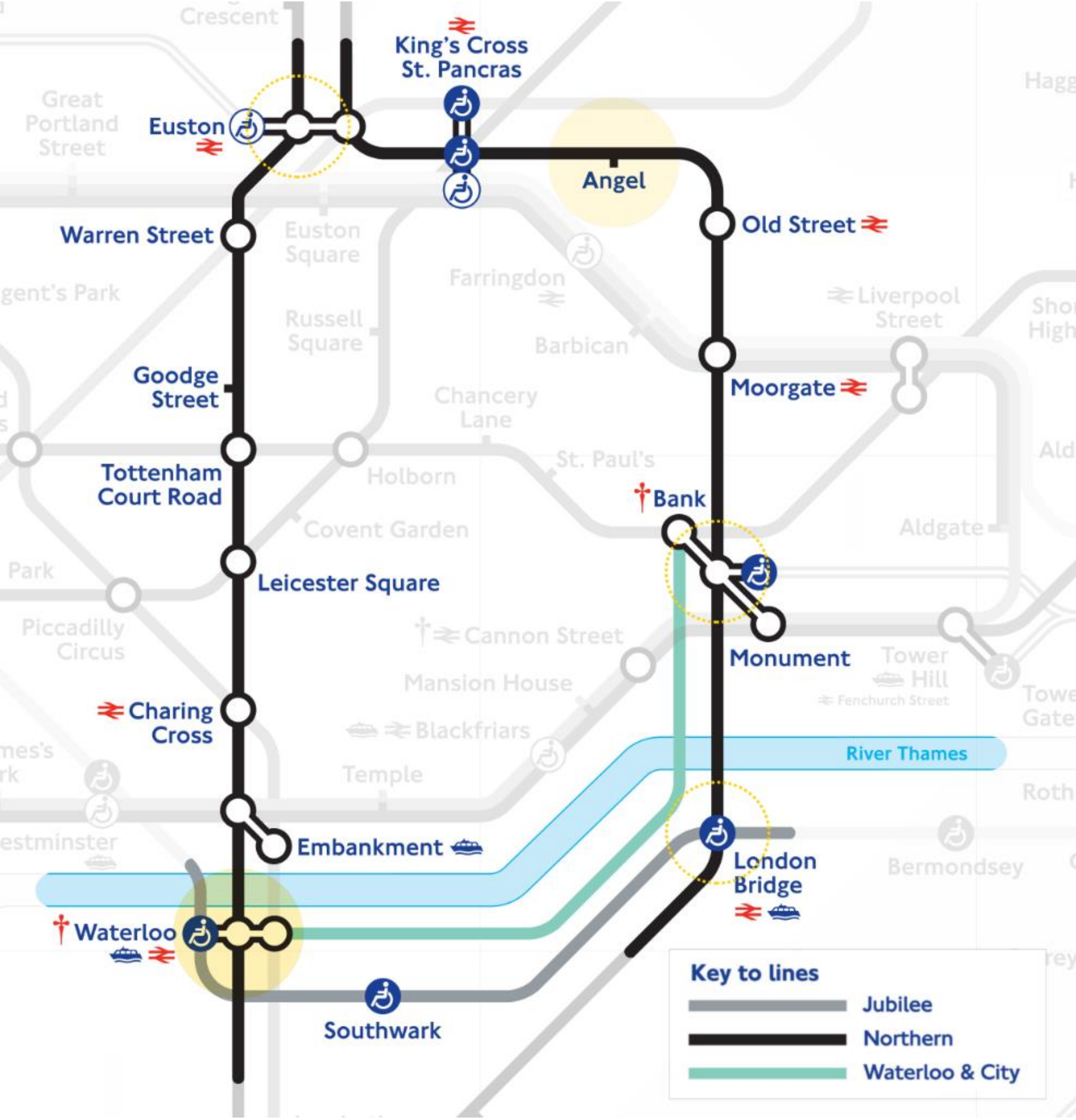
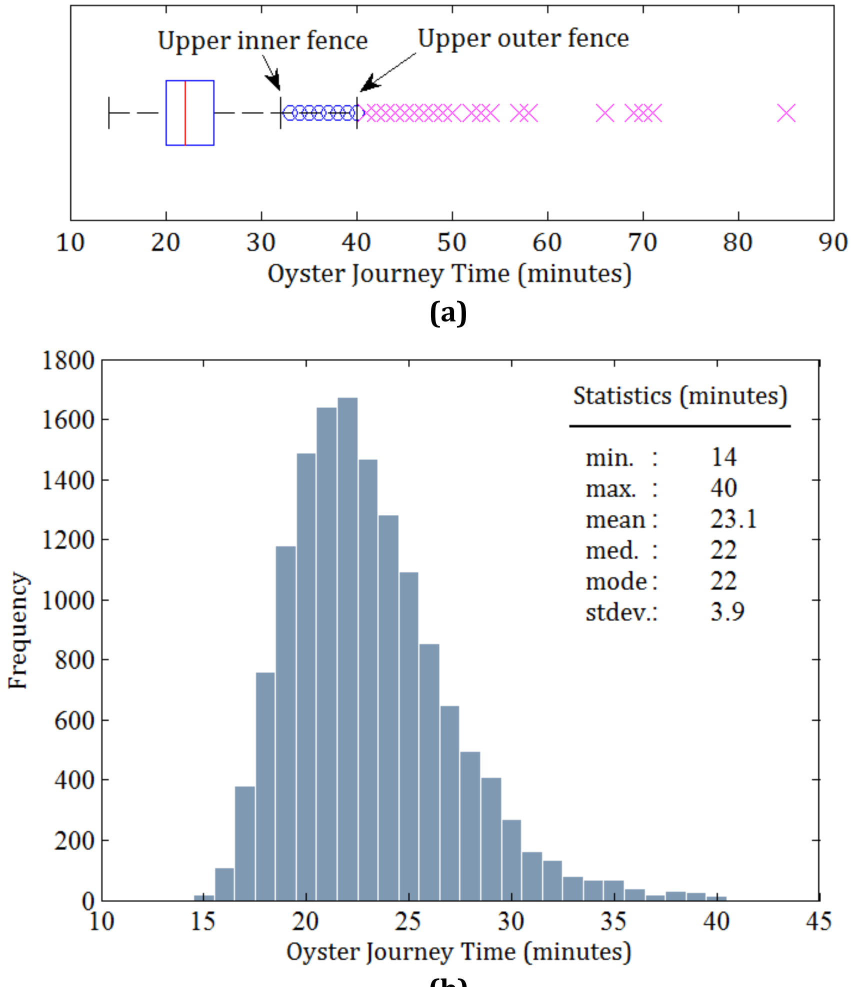


Figure 4.14 The LU network connecting the O-D pair: Angel – Waterloo.

For the former option (*i.e.* a northbound train), the passengers would need to transfer at Euston station, to a southbound Northern Line train. For the latter (*i.e.* a southbound train), two alternative interchange stations are available. That is to say, the passengers could choose to alight at Bank station and transfer to a connecting service on the Waterloo & City line (southbound); or they may remain on the southbound Northern line train (via Bank) and travel a bit further to the station of London Bridge, where they could transfer to a southbound train of the Jubilee line so as to reach Waterloo station. According to the map-view, to make an interchange at Bank would seem to be more attractive than the others, as that route involve only three intermediate stops in total. By contrast, it might possibly cost a much longer journey time to transfer at Euston.

Within the specified AM Peak (from 07:00 *a.m.* to 10:00 *a.m.*), 25,122 journeys were recorded during the observation period of 193 days, with a sample size of 24,760 OJT^{OBS} being considered valid. **Figure 4.15** below gives the statistical summary of the sample data set for this case. As shown in **Figure 4.15(b)**, the frequency distribution still presented a unimodal profile, with the single mode being 22 minutes. This might also imply that the locations (or rather, the location parameters) of the journey time distributions for the three alternative routes were possibly close to each other, which stacked around the mode of the mixture. Otherwise, in light of experience gained from the previous case studies, the journey time distribution of the relatively slower route among the three alternatives might have a higher degree of dispersion.



(a)

(b)

Figure 4.15 Summary of OJT^{OBS} data for **Angel – Waterloo**:

(a) a box-and-whisker plot of the raw data $(n_0 = 14,673)$; and
(b) a histogram of the valid data $(n = 14,419)$.

Using this data sample, we obtained the estimates of the initial values from the *K*-means clustering and that of the mixture model parameters by the EM algorithm. The estimated results are presented in **Table 4.18** below.

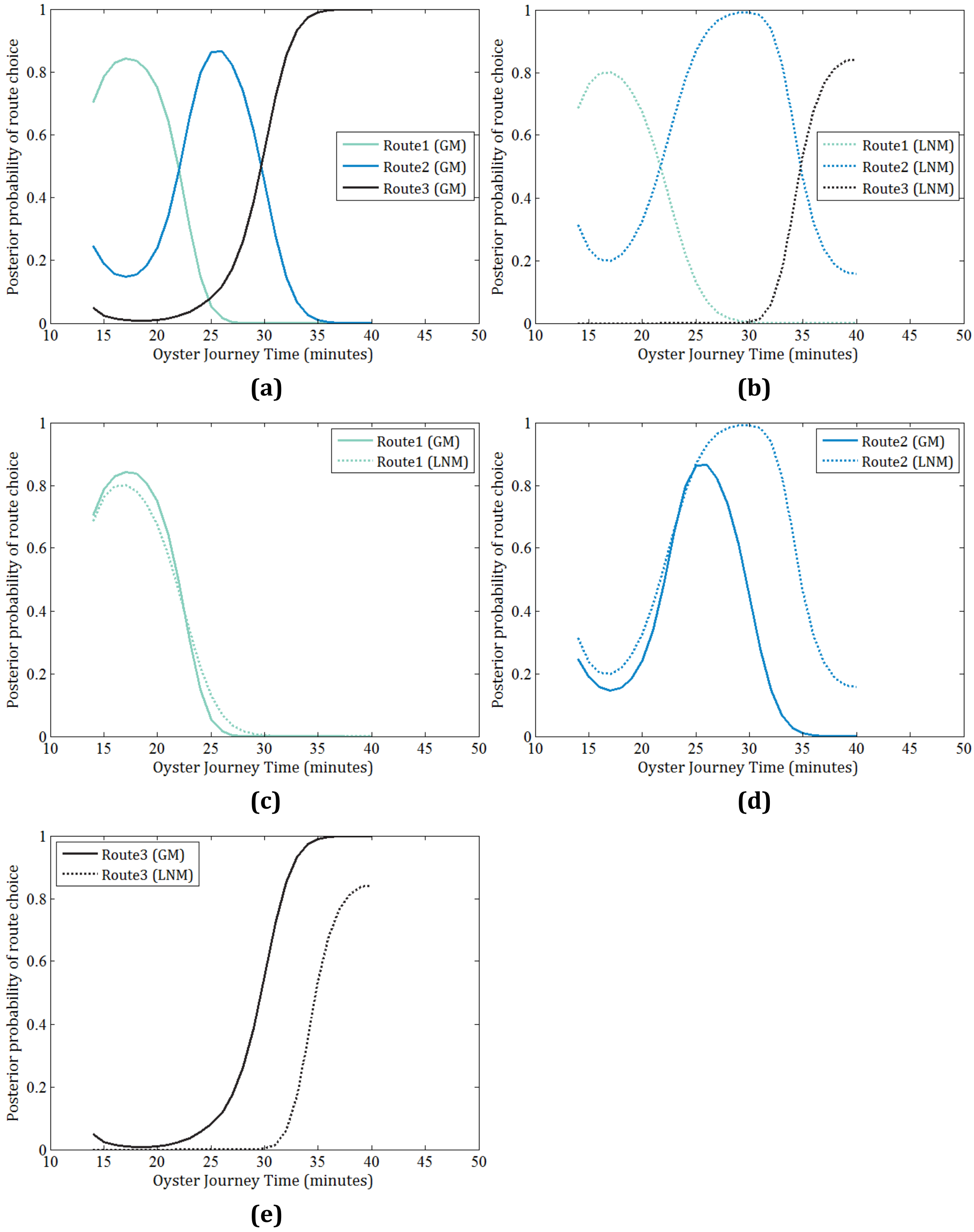
Table 4.18 Parameter estimates of GM and LNM models based on OJT^{OBS} data for **Angel – Waterloo**

The initial values and the model parameters were estimated using the *K*-means clustering and the EM algorithm, respectively. $n = 14,419$.

	GM			LNM		
Component-label	$r=1$	$r=2$	$r=3$	$r=1$	$r=2$	$r=3$
Initial values						
η_r^{KMS} (minute)	20.0	23.0	28.0	19.0	22.0	27.1
σ_r^{KMS} (minute)	1.4	1.1	2.9	1.2	1.1	2.7
ω_r^{KMS} (%)	38.7	38.3	23.0	27.3	42.1	30.6
Parameter estimates						
$\hat{\mu}_r$ (minute)	20.3	24.0	29.2	20.5	24.4	36.0
$\hat{\sigma}_r$ (minute)	1.9	2.9	4.5	2.2	3.5	2.3
$\hat{\omega}_r$ (%)	38.8	49.6	10.6	39.0	59.7	1.3

For Route1 (*i.e.* $r=1$) and Route2 (*i.e.* $r=2$), we could see that η_1^{KMS} and η_2^{KMS} were around the mixture mode for both GM and LNM, while Route3 (*i.e.* $r=3$) tended to be representing a slower route as σ_3^{KMS} appeared to be much larger. For the GM model, ω_1^{KMS} and ω_2^{KMS} were nearly equal to each other. Nonetheless, for the LNM model, ω_1^{KMS} was the smallest among the three routes/components. This might potentially lead to a similar situation in the estimation of $\hat{\omega}_r$ for the mixture models.

As also exhibited in **Table 4.18** above, the estimates of both the mixture models indicated that the journey time distribution of Route1 (that provides the fastest service among the three routes) was shaped by a relatively smaller proportion of the sample OJT^{OBS}. In comparison, Route2 (*i.e.* a slightly slower route) shared the largest portion of the whole passenger traffic. For Route3, the slowest service, $\hat{\mu}_3$ differed significantly between the GM and LNM. This might serve as a crucial point to judge whether the model was acceptable or not.



(a)

(b)

(c)

(d)

(e)

Figure 4.16 Posterior probabilities of route choices given OJT^{OBS} for **Angel – Waterloo** $(n = 14,419)$:

(a) for all alternatives, based on GM; **(b)** for all alternatives, based on LNM;
(c) for Route1, based on GM and LNM; **(d)** for Route2, based on GM and LNM; and
(e) for Route3, based on GM and LNM.

A batch of graphs presented in **Figure 4.16** above illustrates the posterior probabilities of passengers' route choices estimated from the GM as well as LNM models. Take the estimates of the GM for example. As shown in **Figure 4.16(a)** (*see previous page*), the solid curve suggested that if passengers' journey times were less than about 22 minutes, the probability of choosing Route1 was higher than the other two alternative routes. Route2 was believed to be more likely to be chosen by those passengers whose journey times were within the range between about 22 and 30 minutes. If the journey times were longer than about 35 minutes, it was believed by GM model that the passengers had definitely chosen Route3, because in that case both the posterior probabilities of choosing Route1 and Route2 were estimated as approximating zero. Comparing the estimates of each alternative route between the two mixture models, the LNM also suggested a similar trend. In the case of Route1, as shown in **Figure 4.16(c)**, GM and LNM gave similar results; whereas for each of Route2 (*see* **Figure 4.16(d)**) and Route3 (*see* **Figure 4.16(e)**), there existed a substantial gap between the GM and LNM estimates of the choice probabilities.

The distribution of passenger traffic, inferred from the estimated posterior probabilities of individuals' choices, among the alternative routes on this O-D is presented in **Table 4.19** below.

Table 4.19 Inferences of proportion of passenger traffic on each alternative route connecting **Angel** to **Waterloo** ($n = 14,419$)

	GM			LNM		
Component-label	$r = 1$	$r = 2$	$r = 3$	$r = 1$	$r = 2$	$r = 3$
$\hat{\omega}_r$ (%)	39.8	49.6	10.6	39.0	59.7	1.3
$n_r^{\mathrm{INF}_0}$	7,254	6,250	915	5,580	8,645	194
$\omega_r^{\mathrm{INF}_0}$ (%)	50.3	43.3	6.3	38.7	60.0	1.3
n_r^{INF}	5,688	7,189	1,542	5,604	8,626	189
ω_r^{INF} (%)	39.4	49.4	10.7	38.9	59.8	1.3

The consistency between $\hat{\omega}_r$ and ω_r^{INF} in both GM and LNM models again assures the practical significance of the method for *effective* inference. Note that in the case of the GM model, $\omega_r^{\mathrm{INF}_0}$ from the *naive* inference suggested that a larger portion of the passengers might take Route1 (*i.e.* the fastest route), which also seemed to be reasonable. Notwithstanding this, the judgement had to be made after further review of, in our case, the RODS data.

Similar to all the previous cases, the computation of route-specific average journey times is demonstrated in **Table 4.20** below, which was used to support the examination of the estimation results.

Table 4.20 Expected journey times of simulated samples for each alternative route connecting Angel to Waterloo

	Calculated average travel time (minutes)		
l' – l'' s	Northern – Waterloo & City Bank	Northern – Jubilee London Bridge	Northern – Northern Euston
Journey segment			
$t^{\mathrm{ACC}}_{l',o}$	3.8	3.8	3.8
$t^{\mathrm{WFD}}_{l',o,1}$ / $t^{\mathrm{WFD}}_{l',o,2}$	1.5 / 4.8	1.5 / 4.8	1.3 / 4.4
$t^{\mathrm{OBT}}_{l',[o,s]}$	6.0	8.0	4.0
$t^{\mathrm{ICT}}_{[l',l''],s}$	5.2	3.4	3.3
$t^{\mathrm{ICW}}_{l'',s,1}$ / $t^{\mathrm{ICW}}_{l'',s,2}$	1.7 / 4.6	1.4 / 3.7	1.7 / 5.1
$t^{\mathrm{OBT}}_{l'',[s,d]}$	3.0	3.0	10.0
$t^{\mathrm{EGR}}_{l'',d}$	1.0	3.1	2.5
Route-label	$h=1$	$h=2$	$h=3$
Total average			
$t_h(1,1)$	22.1	24.1	26.5
$t_h(2,1)$	25.4	27.4	29.7
$t_h(1,2)$	25.0	26.4	29.9
$t_h(2,2)$	28.3	29.7	33.0
t^{REF}_h	25.2	26.9	29.8

Given $t^{\mathrm{REF}}_h \ \forall h$, the route labelled $h=3$, *i.e.* "Northern – Northern, via Euston station", was believed to be the longest among all the three alternatives; and another route labelled $h=1$, *i.e.* "Northern – Waterloo & City, via Bank station", appeared to be the fastest. This was consistent with our conjecture based on the Tube map for this O-D (*see* **Figure 4.14**). On this basis, in the first instance, we could simply perceive that the estimated mixture component with the largest $\hat{\mu}_r$ (*i.e.* $r=3$, referred to as Route3) should be possibly the slowest route that goes through Euston, and that the component with the smallest $\hat{\mu}_r$ (*i.e.* $r=1$, referred to as Route1) should be likely to be the fastest route via Bank. Thus, Route2 (*i.e.*

$r=2$) was believed to be referring to the route, "Northern – Jubilee, via London Bridge", which was labelled $h=2$. All these matching pairs are displayed in **Table 4.21** below.

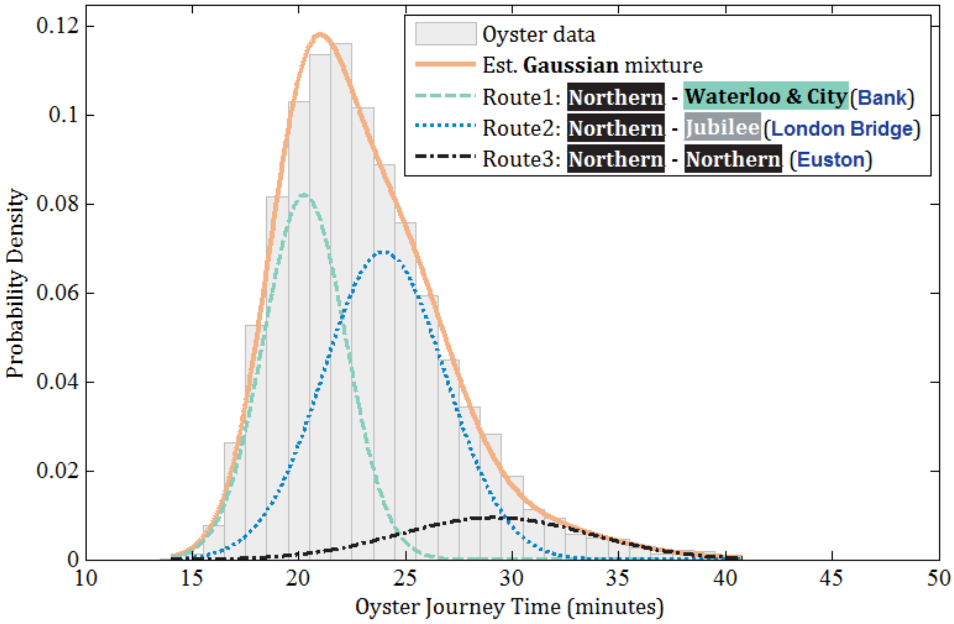
Table 4.21 Matching the estimated mixture components with the real-world routes for Angel – Waterloo

		r **matches** h		
Component-label r		$r=1$	$r=2$	$r=3$
Journey time (minutes)				
$\hat{\mu}_r$	GM	20.3	24.0	29.2
	LNM	20.5	24.4	36.0
t_h^{REF} ($\hat{\sigma}_h^{\text{SEM}}$)		25.2 (1.3)	26.9 (1.2)	29.8 (1.3)
CI for h	95% CL	[21.2, 29.2]	[25.6, 34.0]	[23.2, 30.6]
Traffic distribution (%)				
$\hat{\omega}_r$	GM	39.8	49.6	10.6
	LNM	39.0	59.7	1.3
ω_h^{ROD} (n_h^{ROD})	AM Peak	42.9 (33)	44.1 (34)	13.0 (10)
	A weekday	54.9 (508)	24.2 (224)	20.9 (193)
Route-label h		$h=1$	$h=2$	$h=3$
		Northern – Waterloo & City Bank	Northern – Jubilee London Bridge	Northern – Northern Euston

Let us take for example the estimates of GM model. We could see from **Table 4.21** that $\hat{\mu}_2$ and $\hat{\mu}_3$ were both within the 95% CI of their corresponding t_h^{REF}. Despite $\hat{\mu}_1$ being slightly smaller than the lower CI boundary, it was still perceived acceptable in view of $\hat{\mu}_1$ being closely approximated that boundary; whereas in the case of LNM model, $\hat{\mu}_3$ ($=36.0$ minutes) was far beyond the upper boundary of the corresponding 95% CI. Thus, that estimate was deemed not appropriate, and hence the LNM model would not be considered to be suitable in this case.

For further examination, we compared $\hat{\omega}_r$ to ω_h^{ROD}. Two issues here should be noted. On the one hand, the sample size of RODS data for the AM Peak was small. There might be a higher risk of lack of credibility. On the other hand, in each of the mixture models, a higher mixture weight was assigned to Route2 (via **London Bridge**), than to Route1 (via **Bank**). In other words, $\hat{\omega}_1 < \hat{\omega}_2$. This was also reflected by the RODS result, notwithstanding the presence of the small sample size. However, for the estimated mean journey times, we had $\hat{\mu}_1 < \hat{\mu}_2$. Considering both the Tube map as well as the results of t_h^{REF}, Route1 (via **Bank**) would be expected to be more attractive. This supposition could be supported by evidence from the RODS data in the context of a much larger sample size. As also shown in **Table 4.21**, nearly 55% of passengers chose the quickest route on a typical weekday. On this account, the estimates from GM were still acceptable, though both testing models, especially the LNM, were potentially over-fitting the data.

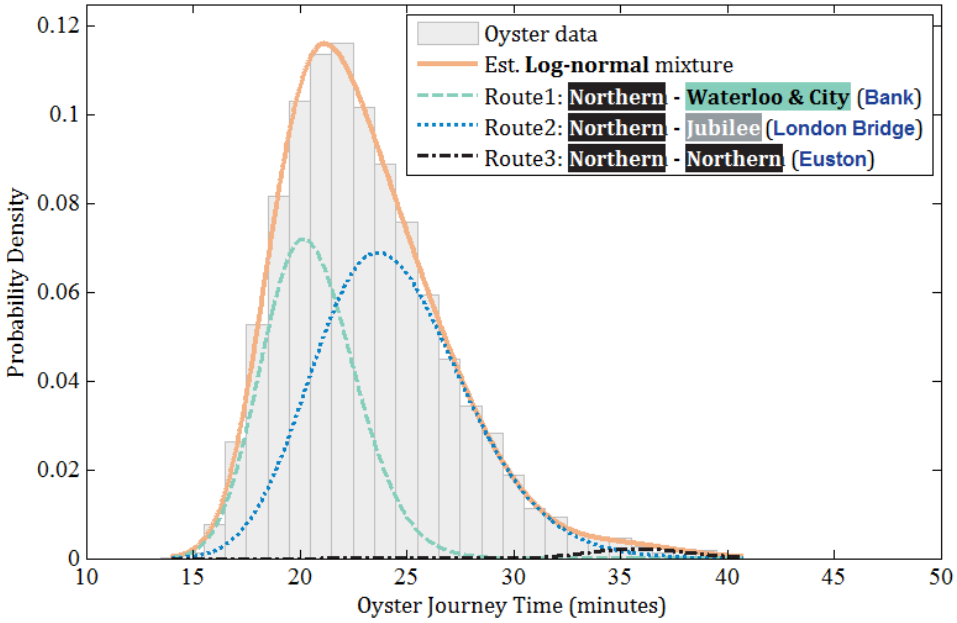
The estimated mixture distributions are illustrated in **Figure 4.17** below (*and also next page*). From the appearances of the two graphs, both the GM and LNM models could fit the sample OJT^{OBS} data very well. Nevertheless, given the estimated parameters for the LNM, it did not seem possible to put a plausible interpretation on which route each of the mixture components might refer to. The LNM would therefore be ignored, compared with the GM.



(a)

Figure 4.17 Estimated mixture distributions, and weighted components thereof, of OJT for **Angel – Waterloo** $(n = 14,419)$:

(a) estimated GM model; and
(b) estimated LNM model (*see next page*).



(b)

Figure 4.17 (*Continued.*)

From the analyses above, the test of goodness of fit in this case was actually not necessary, since we have already made a judgement that the GM model would be relatively more suitable than the LNM. For demonstration purpose, we still present, in **Table 4.22** below, the goodness-of-fit test result. The LNM model, though could have a much better fit than the GM model, might over-fit the sample data in this case.

Table 4.22 Goodness-of-fit test result for **Angel – Waterloo**

The calculation of *gof* was repeated 1,000 times for each model.

	GM	LNM
Rate of obtaining lower *gof* (%)	17.7	82.3
Average *gof*	0.081	0.078

4.3.2.2 A case of four routes (*Case-6*): Euston – South Kensington

In this section, we turn our attention to test the applicability of the GM and LNM models on an O-D pair being served by four alternative routes. The origin and destination are **Euston** and **South Kensington**, respectively. The network of the O-D is illustrated in **Figure 4.18** (*see next page*).

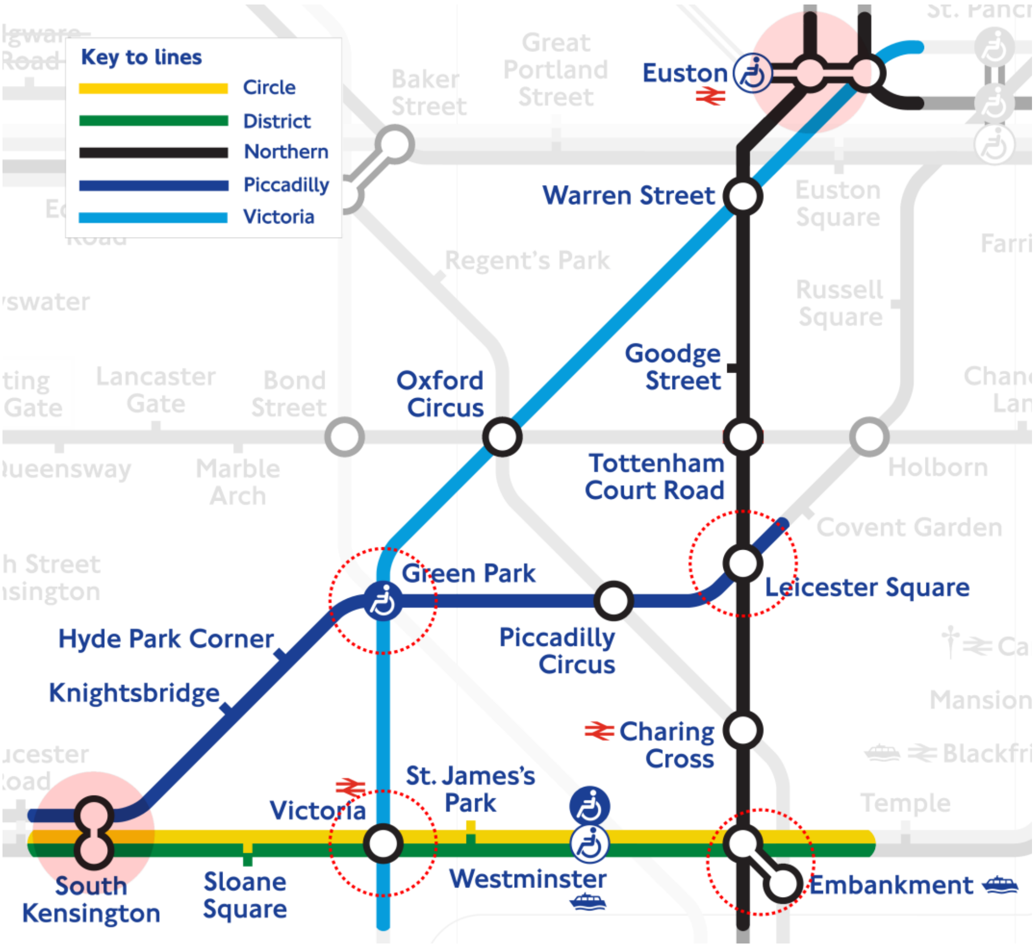
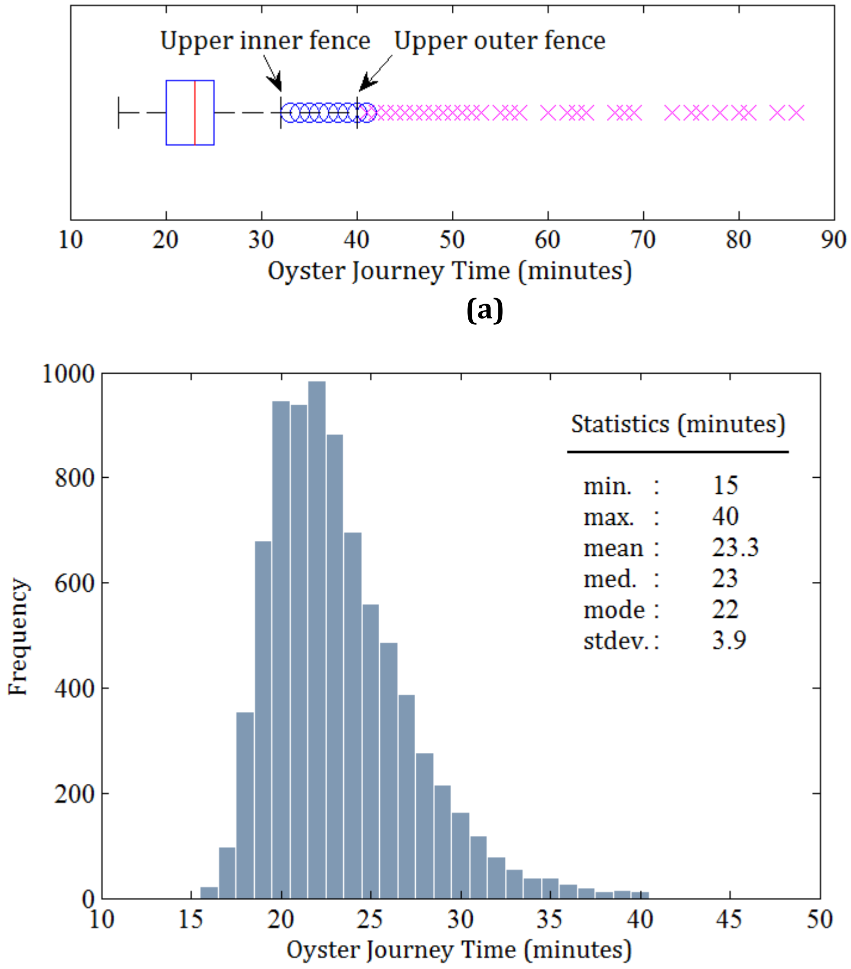


Figure 4.18 The LU network connecting the O-D pair: **Euston – South Kensington**.

In this case, all passengers departing from **Euston** station are supposed to choose between the **Victoria** line and the **Northern** line (southbound). For those who take the former, they may then choose to transfer to a westbound train of the **Piccadilly** line at **Green Park** station, or alight at **Victoria** station but transfer to another westbound train on one of the **Circle**/**District** lines. For those who go for the latter option (*i.e.* taking the **Northern** line for the first journey leg), they may make an interchange at either the stations of **Leicester Square** or **Embankment**. Likewise, it would also lead to a line choice between the **Piccadilly** line and the common lines.

A summary of the OJT^{OBS} data available for this O-D is illustrated in **Figure 4.19** (*see next page*). A total of 8,277 journey records were recorded within the AM Peaks during the 192-day observation period, where a sample size of 8,116 OJT^{OBS} were valid and thus used for estimation of four-component GM and LNM in this case.



(a)

(b)

Figure 4.19 Summary of OJT^{OBS} data for **Euston – South Kensington**:

(a) a box-and-whisker plot of the raw data $(n_0 = 8{,}277)$; and
(b) a histogram of the valid data $(n = 8{,}116)$.

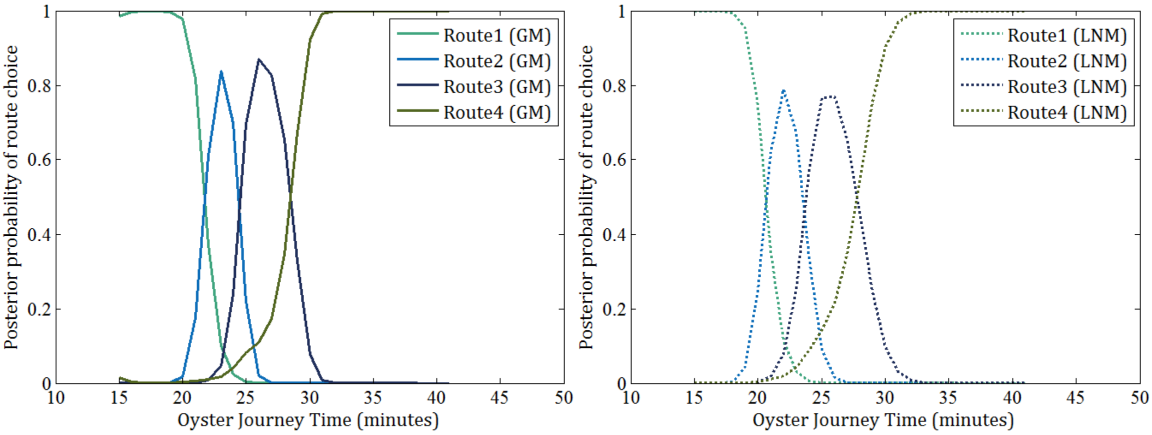
As shown in **Figure 4.19(b)** above, the bimodality is presented in the frequency distribution of the valid data, with the major and minor modes being about 22 and 20 minutes, respectively. This was also reflected in the estimates of η_r^{KMS} as shown in **Table 4.23** (*see next page*); and similar $\hat{\mu}_r$ were obtained for two of all the component distributions, which should characterise the two fastest routes among all the four alternatives. In addition, it appeared that $\hat{\omega}_r$ $\forall r$ were fairly reasonable, which generally suggested that most passengers might prefer faster routes.

Table 4.23 Parameter estimates of GM and LNM models based on OJT^{OBS} data for **Euston – South Kensington**

The initial values and the model parameters were estimated using the *K*-means clustering and the EM algorithm, respectively. $n = 8{,}116$.

	GM				**LNM**			
Component-label	$r=1$	$r=2$	$r=3$	$r=4$	$r=1$	$r=2$	$r=3$	$r=4$
Initial values								
η_r^{KMS} (minute)	20.0	23.0	26.0	31.0	19.0	22.0	25.0	30.1
σ_r^{KMS} (minute)	1.2	0.8	1.1	2.9	1.0	0.8	1.1	2.8
ω_r^{KMS} (%)	37.5	31.5	21.1	9.9	25.9	34.5	26.2	13.3
Parameter estimates								
$\hat{\mu}_r$ (minute)	20.0	22.9	26.0	30.3	19.5	22.0	25.2	29.3
$\hat{\sigma}_r$ (minute)	1.4	1.0	1.4	3.7	1.3	1.2	1.7	3.8
$\hat{\omega}_r$ (%)	40.9	26.6	19.8	12.7	28.4	30.9	24.3	16.4

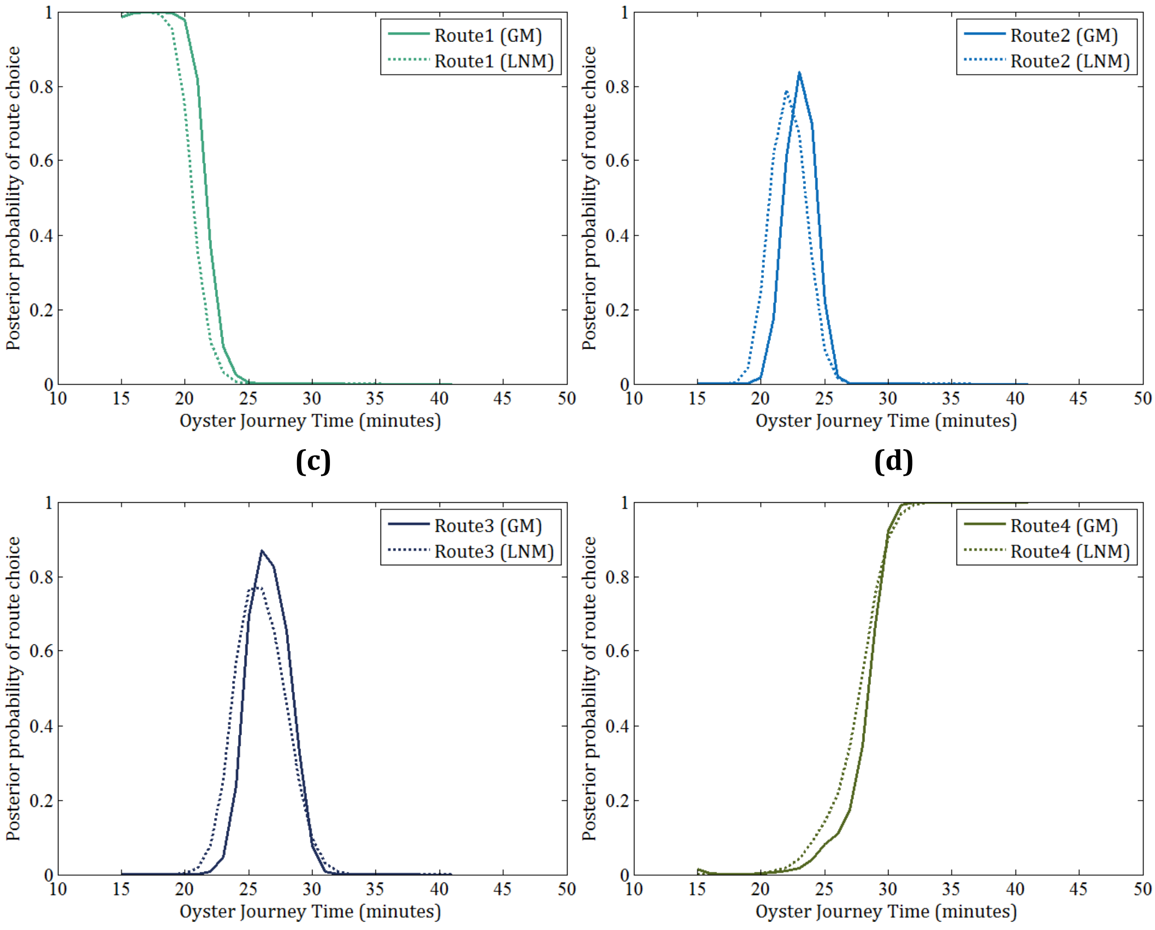
Moreover, the estimated posterior probabilities of passengers' route choices are illustrated in **Figure 4.20** below (*and also next page*). **Figure 4.20(a)** and **(b)** below present the estimation results from the GM and LNM models, respectively.



(a) **(b)**

Figure 4.20 Posterior probabilities of route choices given OJT^{OBS} for **Euston – South Kensington** $(n = 8{,}116)$:

(a) for all alternatives, based on GM; **(b)** for all alternatives, based on LNM;
(c) for Route1, based on GM and LNM (*see next page*);
(d) for Route2, based on GM and LNM (*see next page*);
(e) for Route3, based on GM and LNM (*see next page*); and
(f) for Route4, based on GM and LNM (*see next page*).



(c) **(d)**

(e) **(f)**

Figure 4.20 (*Continued.*)

Figure 4.20(c)–(f) above show comparisons of the choice probabilities for each alternative route between the two models. It could be seen from the four graphs that those both models suggested a similar trend of the route-choice probability condition on journey time.

The inferences of the passenger-traffic distributions among the four alternative routes were presented in **Table 4.24** (*see next page*). For both models, $\omega_r^{\mathrm{INF}_0}$ was close to ω_r^{INF} for each route. This would largely reduce the indeterminacy of the judgement on route-matching and model validation.

To proceed to find out each route-label in this case, the computation of expected average journey times is presented in **Table 4.25** (*see next page*), where the four routes were labelled by $h=1,\ 2,\ 3,$ and 4, respectively. It is noticeable that t_h^{REF} $\forall h$ were clearly distinct from one another. This would greatly facilitate the route-matching process.

Table 4.24 Inferences of proportion of passenger traffic on each alternative route connecting Euston to South Kensington ($n = 8{,}116$)

	GM				LNM			
Component-label	$r=1$	$r=2$	$r=3$	$r=4$	$r=1$	$r=2$	$r=3$	$r=4$
$\hat{\omega}_r$ (%)	40.9	26.6	19.8	12.7	28.4	30.9	24.3	16.4
$n_r^{\mathrm{INF}_0}$	3,047	2,563	1,712	804	2,107	2,807	2,132	1,080
$\omega_r^{\mathrm{INF}_0}$ (%)	37.5	31.5	21.1	9.9	25.9	34.5	26.2	13.3
n_r^{INF}	3,314	2,147	1,654	1,011	2,315	2,514	1,934	1,363
ω_r^{INF} (%)	40.8	26.4	20.4	12.4	28.5	30.9	23.8	16.8

Table 4.25 Expected journey times of simulated samples for each alternative route connecting Euston to South Kensington

	Calculated average travel time (minutes)			
l' – l'' s	Victoria – Circle/District Victoria	Victoria – Piccadilly Green Park	Northern – Piccadilly Leicester Sq.	Northern – Circle/District Embankment
Journey segment				
$t_{l',o}^{\mathrm{ACC}}$	4.0	4.0	2.4	2.4
$t_{l',o,1}^{\mathrm{WFD}}$ / $t_{l',o,2}^{\mathrm{WFD}}$	0.6 / 2.6	0.6 / 2.6	1.8 / 5.1	1.8 / 5.1
$t_{l',[o,s]}^{\mathrm{OBT}}$	7.0	5.0	5.0	8.0
$t_{[l',l''],s}^{\mathrm{TIC}}$	2.1	3.4	2.6	2.2
$t_{l'',s,1}^{\mathrm{WIC}}$ / $t_{l'',s,2}^{\mathrm{WIC}}$	1.6 / 3.8	1.5 / 3.9	1.2 / 3.6	1.5 / 3.6
$t_{l'',[s,d]}^{\mathrm{OBT}}$	4.0	6.0	9.0	10.0
$t_{l'',d}^{\mathrm{EGR}}$	1.1	3.6	3.6	1.1
Route-label	$h=1$	$h=2$	$h=3$	$h=4$
Total average				
$t_h(1,1)$	20.3	24.0	25.6	27.0
$t_h(2,1)$	22.3	26.0	28.9	30.3
$t_h(1,2)$	22.5	26.4	27.9	29.1
$t_h(2,2)$	24.5	28.4	31.3	32.5
t_h^{REF}	22.4	26.2	28.4	29.7

According to the criteria specified in **Section 3.5.2** and experience gained from the previous case studies, we could always preliminarily match a route-label to a real route given the similarity between $\hat{\mu}_r$ and t_h^{REF}. Since no relevant information was available, the validation of such conjecture must be further supported by evidence that $\hat{\mu}_r$ should fall within the CI of t_h^{REF} (at a given CL) for each alternative route. Meanwhile, $\hat{\omega}_r$ also need to be checked with some prior information (*e.g.* ω_h^{ROD} in our case, though, which was not completely reliable).

We take for example the estimates from GM model to illustrate the component-route matching in this case. As demonstrated in **Table 4.26** below, we had two important facts here: (a) $\hat{\mu}_1 < \hat{\mu}_2 < \hat{\mu}_3 < \hat{\mu}_4$ and (b) $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$ and $\hat{\mu}_4$ were within the 95% CI of t_1^{REF}, t_2^{REF}, t_3^{REF} and t_4^{REF}, respectively. This information could then shed light on the preliminary route-matching. That is, Route1 (*i.e.* $r=1$), Route2 (*i.e.* $r=2$), Route3 (*i.e.* $r=3$) and Route4 (*i.e.* $r=4$) correspond, respectively, to the alternative routes via **Victoria**, **Green Park**, **Leicester Square** and **Embankment**.

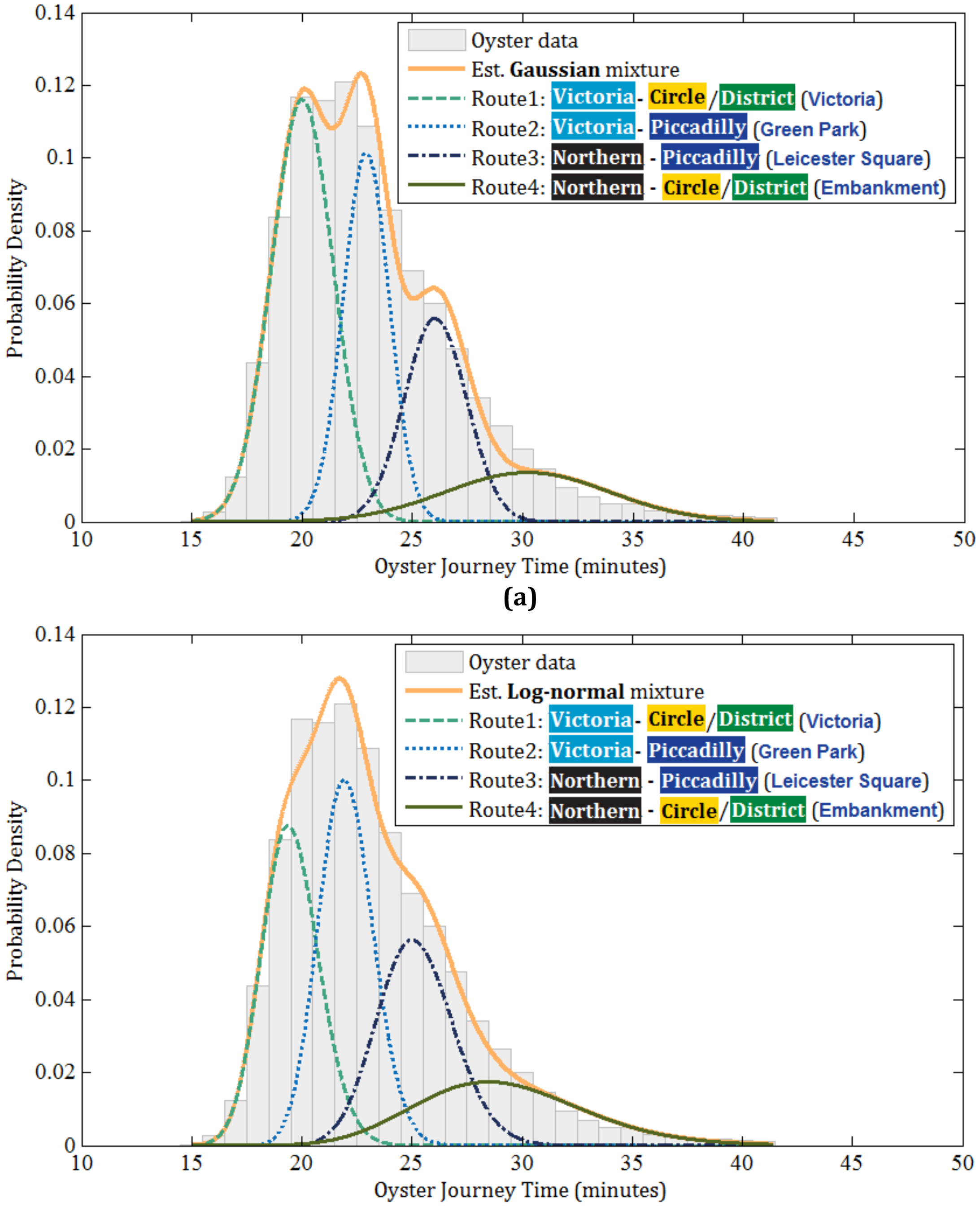
Table 4.26 Matching the estimated mixture components with the real-world routes for **Euston – South Kensington**

		r **matches** h			
Component-label r		$r=1$	$r=2$	$r=3$	$r=4$
Journey time (minutes)					
$\hat{\mu}_r$	GM	20.0	22.9	26.0	30.3
	LNM	19.5	22.0	25.2	29.3
t_h^{REF} ($\hat{\sigma}_h^{\mathrm{SEM}}$)		22.4 (0.8)	26.2 (0.9)	28.4 (1.2)	29.7 (1.1)
CI for h	95% CL	[19.7, 25.1]	[22.8, 29.5]	[24.7, 32.2]	[26.1, 33.4]
Traffic distribution (%)					
$\hat{\omega}_r$	GM	40.9	26.6	19.8	12.7
	LNM	28.4	30.9	24.3	16.4
ω_h^{ROD} (n_h^{ROD})	AM Peak	57.4 (120)	21.1 (44)	21.1 (44)	0.5 (1)
	A whole day	44.0 (176)	31.8 (127)	23.3 (93)	1.0 (4)
Route-label h		$h=1$	$h=2$	$h=3$	$h=4$
		Victoria – Circle/District Victoria	Victoria – Piccadilly Green Park	Northern – Piccadilly Leicester Sq.	Northern – Circle/District Embankment

In the comparison between $\hat{\omega}_r$ and ω_h^{ROD} for each of the mixture components, the general trend of $\hat{\omega}_r$ $\forall r$ also appeared to be consistent with the RODS results.

Therefore, we could then confirm the preliminary judgement and that could be deemed as the conclusion in this case.

Given all the parameter estimates, **Figure 4.21** below illustrates the estimated mixture distributions of both the GM and LNM models.



(a)

(b)

Figure 4.21 Estimated mixture distributions, and weighted components thereof, of OJT for **Euston – South Kensington** $(n = 8{,}116)$:

(a) estimated GM model; and
(b) estimated LNM model.

Notwithstanding the goodness-of-fit test results presented in **Table 4.27** below, the GM model was considered more suitable for this O-D case due to its more reasonable parameter estimates.

Table 4.27 Goodness-of-fit test result for **Euston – South Kensington**

The calculation of gof was repeated 1,000 times for each model.

	GM	**LNM**
Rate of obtaining lower gof (%)	17.4	82.6
Average gof	0.091	0.087

4.4 Summary and conclusions

Following the idea proposed and the method discussed in the previous chapter, this chapter has implemented the mixture models of passengers' journey times on a real underground network. The applications and features of both GM and LNM models have been demonstrated separately, and compared, on seven different O-D pairs based on the LU network, whereof five cases have been described in detail in this chapter. The other two cases have been exhibited in Appendix C that shows only the estimation results.

Among all the seven cases, there was barely bi- or multi-modality exhibited in the mixture journey time distributions per se. This intrinsic feature, however, does not matter much for the application of the mixture models. In most cases, K-means could effectively capture the modes of the mixture distribution, which would greatly facilitate the delivery of sensible estimates by EM algorithm. It has been noted that the mixture model estimates, especially the estimated values of means, did not differ greatly from the initial values given by K-means. This might be partly because K-means is a special case of the EM algorithm; and partly because in some cases, K-means clustering might afford satisfied estimates, to a certain extent. For future research, more experiments could be done to test the influence of different initial values may have on the estimation results, using different methods other than K-means.

In addition, when the number of alternative routes is small, say only two, GM and LNM models could afford similar results, where LNM may often provide a relatively better goodness of fit. As the route-choice set grows larger, LNM may be more likely to produce 'extreme' component estimates, whereas GM would suit better the cases with an increasing number of alternative routes.

In all case studies, the calculated average journey times t_h^{REF} $\forall h$ were always greater than the estimated means, $\hat{\mu}_r$ $\forall h$. The most likely reason for this was that the computation of t_h^{REF} $\forall h$ always considered $t_h(\phi,\psi)$ equally for all the four circumstances specified in **Section 3.5.1** (*i.e.* $\forall \phi = 1,2$ and $\forall \psi = 1,2$); and $t_h(2,2)$ accounted for 25% of t_h^{REF}, which might be too high. It will surely be better to have a weighted average of t_h hence a better reference value of t_h^{REF}. Still, it has been shown that $\hat{\mu}_r$ in most cases could fall within the 95% CI of t_h^{REF}, which largely supports the identification of each route-label. Nevertheless, as has been briefly summarised in **Section 4.3.2.2**, the identification process in this thesis was rather subjective (*cf.* **Section 3.5.2**). On this account, an algorithm for automatic identification of the route-labels should be further studied in future research.

In another regard, the level of traffic congestion would vary even within the specified three-hour period of study (*i.e.* the AM-Peak, defined as between 7:00 *a.m.* and 10:00 *a.m.*), so that passengers' perceptions to route choices may change as well. The case studies carried out in this thesis investigated only the AM-Peak as a whole. Further studies should be devoted to a shorter term with a larger sample given a relatively stable congestion level. Also, it is possible to obtain different mixture/component distributions given data from different time-bands of a day. Comparisons between the distributions by different time-band of a day (*e.g.* between the Peaks and Off-peak) may thus assist us to draw some more general conclusions about passengers' travel behaviour, such as whether they would tend to avoid busy stations at rush hour.

In general, the outcomes of those case studies have shown that the finite mixture models could be a qualified inference framework for passengers' probabilistic route choices at the aggregate level. It also enhances the potential of making use of the smart-card data to estimate passengers' probabilistic route choices on any other similar public transport networks.

Additionally, in some special cases on the LU network, the Oyster travellers are advised to swipe their Oyster cards on a 'pink' reader at some interchange stations, except for the ticket validation required at both the O-D gatelines. In that way, the cardholders' fares would be calculated properly according to the specific routes they have chosen. Otherwise, the maximum fare will be charged for travelling between the corresponding O-D. Accordingly, in such cases, the information about where and when passengers made interchanges is readily available. It will thus be worth examining these cases for future work, where we may firstly put aside the interchange data but estimate a mixture model; and then compare the model estimates with the real information of interchanges. This will greatly assist in testing the applicability of the mixture model in estimating passengers' probabilistic route choices, and also improve the odds of obtaining a more appropriate model.

It needs to be stated again that the mixture model allows for the observed journey time (*i.e.* OJT^{OBS} in our case) to be an only condition for estimating the posterior probabilities. Therefore, different passengers, who were observed to have spent the same amount of η_r^{KMS} travelling on this (and any other) O-D, are supposed to share an identical posterior probability of choosing each alternative route. This issue will be further investigated in the following chapter.

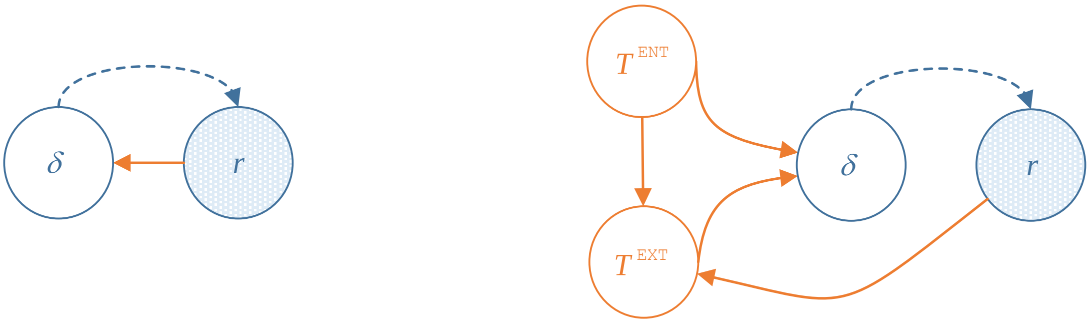
Chapter 5
Updating route choice probabilities

5.1 Introduction

In this chapter, we revaluate and update every passenger's choice probabilities with additional consideration of the determinants in each of their journey times. To this end, we recall two elementary events, $choice_{qr}$ as well as $\boldsymbol{\delta}_q$, which have been discussed in **Chapter 3**; and trace them back to the simple network of o-d illustrated in **Figure 3.1**. As has been defined in **Section 3.2**, for each individual passenger, $choice_{qr}$ represents a statistical event that passenger q chose route r when he/she travelled from o to d; and the symbol $\boldsymbol{\delta}_q$ represents another event that the observed journey time of q is δ_q^{OBS}. For Bayesian inference, $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q)$, as a conditional probability function of the two events recalled, represents a posterior probability of q choosing r. It was conceived to be a straightforward representation of the probabilistic route choices made by q, given the common set, R, of route choices (*cf.* **Section 3.4**). Additionally, the likelihood of $choice_{qr}$ occurring was predicated on the understanding that the journey time of q has been known.

Up to this point, it must be noted that the journey time has been serving as the only explanatory variable for the measurement of the route-choice probabilities. According to the mixture models (that has been implemented and demonstrated in the **Chapter 4**), passengers who were observed to have the same journey times were assumed to have the same choice probabilities for all the alternative routes. In other words, for any two individuals sampled from the passenger population, who are labelled i and j (where $i, j \in Q$, $\forall i \neq j$), respectively, if $\delta_i^{\text{OBS}} = \delta_j^{\text{OBS}}$, then it should be taken for granted by the mixture model that the posterior probabilities that they might have chosen the same route, say route r ($\forall r \in R$), are equivalent. In that case, we should obtain the following equation: $\Pr(choice_{ir} \mid \boldsymbol{\delta}_i) = \Pr(choice_{jr} \mid \boldsymbol{\delta}_j)$. The two passengers may be thus regarded as having the same preference for every alternative route, even if they had actually used different routes. Or conversely, although the two passengers i and j made their journeys by the same route, the fact that different journey time observations (*i.e.* $\delta_i^{\text{OBS}} \neq \delta_j^{\text{OBS}}$) might result in them having different posterior route-choice probabilities, thereby leading to biased estimates at the individual level. This is actually an inherent drawback of the posterior probabilities directly derived from the mixture models. Although the mixture model appeared to perform well for the estimation of route choice probabilities at aggregate level, we cannot infer that $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q)$ presents individuals' probabilistic route choices with a high degree of confidence.

Basically, the mixture model allows for an oversimplified assumption on the probabilistic relationship between the two variables: passengers' journey time, δ, and their possible route choice, r. Such correlation could be simplistically represented by a graphic structure, as shown in **Figure 5.1(a)** below (*see also* **Appendix A**). The solid, **orange** coloured arc that joins the r-node to the δ-node represents a real-world causality. It indicates that any passenger's journey time can be observed only after his/her journey has been completed, for which the passenger must have made a route choice. That is to say, a journey through route r brings about (the observations of) δ, with a probability distribution $p(\delta \mid r)$.



(a) **(b)**

Figure 5.1 Bayesian-network structures for investigating passengers' probabilistic route choices:

(a) a simplistic graphic structure showing probabilistic relationship between journey time and route choice; and

(b) an extended structure showing causal conditions between entry time, exit time, journey time and route choice.

But since r is unobservable, we may only be able to learn about it in view of the journey time observation, with a posterior probability distribution $p(r^{(q)} \mid \delta)$, where $r^{(q)}$ represents a possible route choice that q might have made. This dependency is indicated by the dashed arc in **Figure 5.1(a)**, which has also been our primary concern in the previous two chapters.

In essence, the journey time δ in the mixture model is treated as an independent variable. This implies that evidence about different entry times T^{ENT} (or exit times T^{EXT}) among the individual passengers would not have effect on learning about the distribution of δ. Accordingly, two equations hold: $p(\delta) = p(\delta \mid T^{\text{ENT}})$ and $p(\delta) = p(\delta \mid T^{\text{ENT}})$. That is, T^{ENT} and δ were assumed to be independent and so were δ and T^{EXT}, but only the difference, $\delta = T^{\text{EXT}} - T^{\text{ENT}}$, matters. As has been mentioned at the outset of this chapter, it is only the variation in the journey times that causes the individual passengers to be assigned different choice probabilities by the mixture model. In this sense, the estimates from the mixture model purely suggest the average route-choice probability (or average level of preference to each alternative route) of passengers who spent about the same journey time.

However, passengers' entry time T^{ENT} actually acts as an important influencing factor in the journey time variations. In addition, their exit time T^{EXT} would be largely dependent on the specific routes they choose after they touch in, and hence the corresponding journey times differ. That is to say, the passengers' journey times are caused jointly by T^{ENT}, T^{EXT} as well as their route choices. The dependencies among these three variables are illustrated in **Figure 5.1(b)** (*see previous page*). Now in this renewed framework, by comparison to the structure in **Figure 5.1(a)**, T^{ENT} becomes an independent variable, so that passengers' journey times are considered to have an indirect dependency on their route choices given both T^{ENT} and T^{EXT}.

Nevertheless, suppose that we have known which route the passengers have chosen. As discussed in **Chapter 2**, their journey times would also hinge upon the linkage between the passengers' within-station movements at each journey segment and the transit services (*e.g.* the timetable of different transit lines). It involves a bunch of factors, such as layouts of passages within the stations, individuals' walking speeds, how many attempts made to successfully board trains, and the trains' timetable as well as service reliability. Considered from this perspective, **Figure 5.1(b)** simplistically skips over a sequence of serially dependent time variables that cause observations of T^{EXT}.

A main objective of this chapter is to explore a way of involving the linkage in learning the route choice probabilities for each individual passenger. Efforts would be focused on combining the passengers' entry and the trains' running schedule, which provides extra evidence for the implications of the passengers' route choices. We will then attempt to factor such additional information in the individual choice probabilities estimated from the mixture model, in order to acquire a set of more credible posterior probabilities of each individual possible route choice.

The rest of this chapter is arranged as follows. **Section 5.2** expands on the points that have raised in the current section, where the problem being concerned are reduced to a single variable. It is demonstrated in **Section 5.3** that how the single variable, as an additional condition, could be involved into the previously estimated posterior probabilities from the mixture model. This is followed by **Section 5.4**, where we draw detailed Oyster data samples[1] from the same LU O-D pairs studied in **Chapter 4**, so as to present an illustrative example showing a comparison of the before-and-after individual route-choice probabilities. **Section 5.5** summarises and concludes this chapter.

A set of symbols that will be used in the following sections is listed below.

Notation:

T_q^{ENT}	**ent**ry $(\mathtt{ENT})$ time of passenger q
T_{qr}^{EXT}	**exit** $(\mathtt{EXT})$ time of q, given that he/she chooses r
δ_{qr}	journey time of passenger q making a single journey by r
$\delta_{qr}^{\mathrm{EXP}}$	**exp**ected $(\mathtt{EXP})$ journey time of q using r, given T_q^{ENT}, average walking time and trains' timetables
$\boldsymbol{\delta}_{qr}$	elementary event that the expected journey time of passenger q is $\delta_{qr}^{\mathrm{EXP}}$, given that he/she chooses route r and his/her entry time is T_q^{ENT}
n	sample size of a given data set
$f_{qr}(\delta_{qr})$	PDF of distribution of δ_{qr}

(*Continued*)

[1] The information of each individual passenger's entry and exit times is available.

Notation: (*Continued.*)

ϑ_{qr}	vector of a parameter (or parameters) for $f_{qr}(\delta_{qr})$
$\hat{\mu}_r$	estimate of mean of journey time of r
$\hat{\sigma}_r$	estimate of standard deviation of journey time of r
$\pi_{qr}^{\texttt{UMM}}$	**u**pdated posterior probability of passenger q choosing route r, based on the estimate, $\pi_{qr}^{\texttt{MIX}}$, from a **mi**xture **m**odel ($\texttt{UMM}$)
$\mathbf{\Pi}_{\Delta_{5\%}}^{\texttt{UMM}}$	$n \times N_R$ matrix that enumerates all $\pi_{qr}^{\texttt{UMM}}$
$\hat{\gamma}_{qr}$	estimate of the location parameter for PDF of $f_{qr}(\delta_{qr})$
$\hat{\varsigma}_r$	estimate of the scale parameter for PDF of $f_{qr}(\delta_{qr})$
$\omega_r^{\texttt{UPD}}$	proportion of passenger using route r, based on *effective* inference from **upd**ated ($\texttt{UPD}$) route-choice probabilities

5.2 Correlation between passengers' entry and trains' timetable

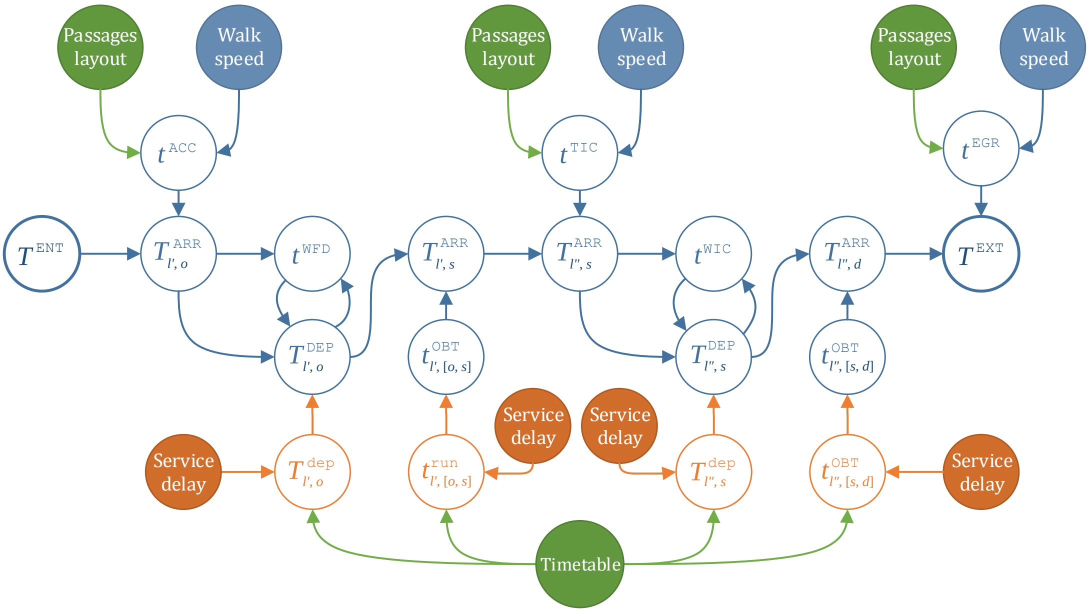

Passages layout
Walk speed
Passages layout
Walk speed
Passages layout
Walk speed
$t^{\texttt{ACC}}$
$t^{\texttt{TIC}}$
$t^{\texttt{EGR}}$
$T^{\texttt{ENT}}$
$T_{l',o}^{\texttt{ARR}}$
$t^{\texttt{WFD}}$
$T_{l',s}^{\texttt{ARR}}$
$T_{l'',s}^{\texttt{ARR}}$
$t^{\texttt{WIC}}$
$T_{l'',d}^{\texttt{ARR}}$
$T^{\texttt{EXT}}$
$T_{l',o}^{\texttt{DEP}}$
$t_{l',[o,s]}^{\texttt{OBT}}$
$T_{l'',s}^{\texttt{DEP}}$
$t_{l'',[s,d]}^{\texttt{OBT}}$
Service delay
Service delay
Service delay
Service delay
$T_{l',o}^{\texttt{dep}}$
$t_{l',[o,s]}^{\texttt{run}}$
$T_{l'',s}^{\texttt{dep}}$
$t_{l'',[s,d]}^{\texttt{OBT}}$
Timetable

Figure 5.2 A Bayesian-network structure showing the causality between passengers' entry time and exit time.

From the considerations above, it is necessary for us to expand the arcs which, in **Figure 5.1(b)**, indicate the causal relationships between T^{ENT} and T^{EXT}, as well as between r and T^{EXT}. To this end, a much more complex structure is built accordingly, as illustrated in **Figure 5.2** (*see previous page*). This framework incorporates a sequence of time variables relating to different journey segments of a specific route, which delineates the way that the determinants of any individual's journey time and the transit service interrelate.

Particularly, we use T_q^{ENT} to represent the entry time of an individual passenger q at his/her origin station; and suppose that q has chosen route r. His/her exit time, which we represent by T_{qr}^{EXT}, would therefore be affected by all of the variables relating to r shown in **Figure 5.2**. In addition, we let δ_{qr} denote the journey time that the passenger q has spent in travelling on the route r, and so we have $\delta_{qr} = T_{qr}^{\text{EXT}} - T_q^{\text{ENT}}$. Apparently, given a certain entry time T_q^{ENT}, then T_{qr}^{EXT} (and hence δ_{qr}) may vary for q due to various circumstances, such as delays in transit services (or inconsistency of train punctuality) as well as passenger-traffic congestion (or even overcrowding) leading to passengers' failures to board the trains. As a matter of fact, this framework could be viewed as a Markov chain (Kleinrock, 1975, pp.21-22). However, we turn to a general way of looking at this problem by reducing its inherent complexity.

It is commonly assumed that arrivals of passengers, and hence the arrival times, at their origin stations would be uniformly distributed during a certain period. In our case, we consider their entry times at the gateline, which similarly follow a uniform distribution and T_q^{ENT} $\forall q$ are independent of each other. Recall the calculated average travel time of each alternative route, which is represented by $t_h(\phi,\psi)$. It is defined as a sum of all the travel time variables of journey segments (*cf.* **Section 3.5.1**) and has been used for the interpretation and validation of the mixture model. Since each of the component-labels, r, has been paired up with a route-label, h, they are exchangeable, and hence $t_h(\phi,\psi) = t_r(\phi,\psi)$, given that r matches h.

As T_q^{ENT} is independent, δ_{qr} is equivalent to $t_r(\phi,\psi)$ on condition that T_q^{ENT} is known:

$$\delta_{qr} = t_r(\phi,\psi) \mid T_q^{\text{ENT}}.$$

Let δ_{qr}^{EXP} denote the (conditional) mean of δ_{qr}. It is then regarded as a conditional expectation of the route average journey time, and represented by

$$\delta_{qr}^{\text{EXP}} = E\left[t_r(\phi,\psi) \mid T_q^{\text{ENT}}\right]. \tag{5-1}$$

where $E\left[t_r(\phi,\psi)\right]$ was calculated as the sum of the averages for all the journey segments. Note that δ_{qr}^{EXP} does not necessarily equal $E[t_r(\phi,\psi)]$; nevertheless, as a conditional variable, it is supposed to differ among passengers having different entry times. This is mainly because of the variations among passengers' wait times for boarding a train, at the origin and/or interchange station. To a large extent, δ_{qr}^{EXP} could account for the facts that (a) passengers might experience the same journey time but actually travelled by different routes, and (b) passengers might experience different journey times though travelled by the same route. It could therefore be of great value for refreshing the passengers' route-choice probabilities estimated from the mixture models.

To obtain the values of δ_{qr}^{EXP} $\forall q \in Q$ and $\forall r \in R$, several assumptions are made as follows. We assume that the station facilities, especially the layouts of all the passages, are fixed. In this way, the average walking times to access, egress and interchange are calculated given the average speeds within the passenger population. In addition, we assume that there is consistent punctuality of transit services, whereby individuals' wait times could be calculated.

The proximate cause of journey time variation is reduced to only the entry time and trains' timetables, but reflected by the wait time. **Figure 5.3** below depicts a simplified structure, compared to that in **Figure 5.2** (*see* p.131), where those shaded nodes with dashed outlines represent the averages of the corresponding variables based on the assumptions above. The plain nodes with dashed outlines would then be fixed given the observation of T_q^{ENT}, and thereby δ_{qr}^{EXP} is derived.

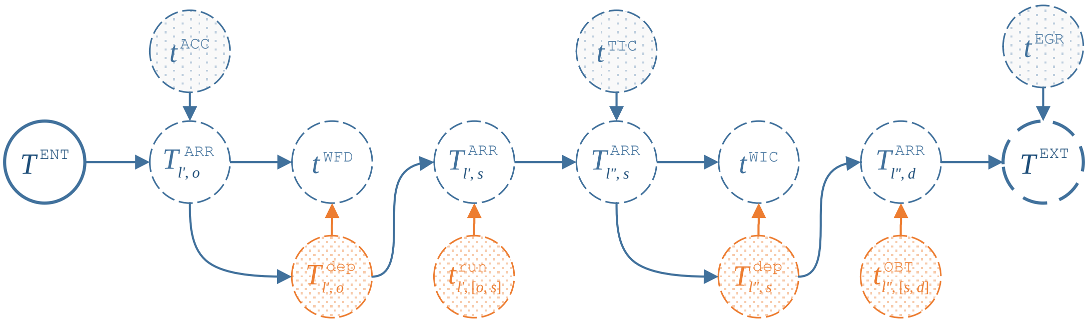


Figure 5.3 A simplified structure of passengers' journey.

Let $\boldsymbol{\delta}_{qr}$ represent an elementary event that the expected journey time of passenger q travelling by route r is $\delta_{qr}^{\mathrm{EXP}}$, given his/her entry time T_q^{ENT} is observed. This additional information would be considered for updating each individual route-choice probability $\Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q)$ estimated from the mixture model in **Chapter 4**. That is, we are now trying to calculate the choice probability of passenger q by taking account of two conditions including both $\boldsymbol{\delta}_q$ and $\boldsymbol{\delta}_{qr}$. This posterior probability is denoted by $\Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$ accordingly.

For all alternative routes, all the corresponding probabilities must also sum to one:

$$\sum_{r \in R} \Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) = 1 . \tag{5-2}$$

This constraint is again to specify that passenger q only chooses one of all the alternative routes. Of central interest to us now is that how we could deal with the additional information of $\boldsymbol{\delta}_{qr}$ and work out $\Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$ $\forall r \in R$.

5.3 Updating the posterior route-choice probabilities

5.3.1 Factoring additional condition

By definition of conditional probability, we have

$$\Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) = \frac{\Pr(choice_{qr}, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})}{\Pr(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})} , \tag{5-3}$$

provided that $\Pr(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$ exist and that $\Pr(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) > 0$. In conformity with the product rule (*cf.* Russell and Norvig, 2010, pp.485-486), $\Pr(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$ can be further expressed as follows:

$$\Pr(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) = \Pr(\boldsymbol{\delta}_q \,|\, \boldsymbol{\delta}_{qr}) \Pr(\boldsymbol{\delta}_{qr}) = \Pr(\boldsymbol{\delta}_{qr} \,|\, \boldsymbol{\delta}_q) \Pr(\boldsymbol{\delta}_q) . \tag{5-4}$$

It should be noted that, $\boldsymbol{\delta}_{qr}$ would occur for sure given the observation of q's entry time T_q^{ENT}, which does not affect the probability of $\boldsymbol{\delta}_q$. The two events $\boldsymbol{\delta}_q$ and $\boldsymbol{\delta}_{qr}$ are conditionally independent given the entry time of q is observed.

For the numerator, $\Pr(choice_q, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$, it is a joint probability that all the three events would occur simultaneously. By applying the chain rule (*cf.* Russell and Norvig, 2010, pp.514-515), it could be factored in several ways as the order of events in the joint probability does not matter. Besides the equivalence presented by equation (5-3) itself, we also have

$$\begin{aligned}\Pr(choice_{qr}, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) &= \Pr(\boldsymbol{\delta}_q \,|\, \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(\boldsymbol{\delta}_{qr} \,|\, choice_{qr}) \Pr(choice_{qr}) \\ &= \Pr(\boldsymbol{\delta}_q \,|\, \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(choice_{qr} \,|\, \boldsymbol{\delta}_{qr}) \Pr(\boldsymbol{\delta}_{qr}) \\ &= \Pr(\boldsymbol{\delta}_{qr} \,|\, \boldsymbol{\delta}_q, choice_{qr}) \Pr(\boldsymbol{\delta}_q \,|\, choice_{qr}) \Pr(choice_{qr}) \\ &= \Pr(\boldsymbol{\delta}_{qr} \,|\, \boldsymbol{\delta}_q, choice_{qr}) \Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q) \Pr(\boldsymbol{\delta}_q).\end{aligned}$$

Moreover, it should also be noted that the expected journey time $\delta_{qr}^{\mathrm{EXP}}$ is derived under the premise that q has actually chosen route r. The event, $choice_{qr}$, as a condition, would provide no more information about the occurrence of $\boldsymbol{\delta}_{qr}$, and vice versa. On this account, the two events, $choice_{qr}$ and $\boldsymbol{\delta}_{qr}$, are deemed to be independent. That is,

$$\Pr(\boldsymbol{\delta}_{qr} \,|\, choice_{qr}) = \Pr(\boldsymbol{\delta}_{qr}), \tag{5-5}$$

and

$$\Pr(choice_{qr} \,|\, \boldsymbol{\delta}_{qr}) = \Pr(choice_{qr}). \tag{5-6}$$

Therefore, the number of alternatives for equation (5-3) could be reduced to three, and so

$$\begin{aligned}\Pr(choice_{qr}, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) &= \Pr(\boldsymbol{\delta}_q \,|\, \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(\boldsymbol{\delta}_{qr}) \Pr(choice_{qr}) \\ &= \Pr(\boldsymbol{\delta}_{qr} \,|\, \boldsymbol{\delta}_q, choice_{qr}) \Pr(\boldsymbol{\delta}_q \,|\, choice_{qr}) \Pr(choice_{qr}) \\ &= \Pr(\boldsymbol{\delta}_{qr} \,|\, \boldsymbol{\delta}_q, choice_{qr}) \Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q) \Pr(\boldsymbol{\delta}_q).\end{aligned}$$

Still, there are six combinations for the fraction on the right-hand side of equation (5-3), which are enumerated as follows:

(i) $$\frac{\Pr(\boldsymbol{\delta}_q \,|\, \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(choice_{qr})}{\Pr(\boldsymbol{\delta}_q \,|\, \boldsymbol{\delta}_{qr})}$$

(ii) $$\frac{\Pr(\boldsymbol{\delta}_q \,|\, \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(\boldsymbol{\delta}_{qr}) \Pr(choice_{qr})}{\Pr(\boldsymbol{\delta}_{qr} \,|\, \boldsymbol{\delta}_q) \Pr(\boldsymbol{\delta}_q)}$$

(iii) $$\frac{\Pr(\boldsymbol{\delta}_{qr} \,|\, \boldsymbol{\delta}_q, choice_{qr}) \Pr(\boldsymbol{\delta}_q \,|\, choice_{qr}) \Pr(choice_{qr})}{\Pr(\boldsymbol{\delta}_q \,|\, \boldsymbol{\delta}_{qr}) \Pr(\boldsymbol{\delta}_{qr})}$$

(*see next page*)

(iv) $$\frac{\Pr(\boldsymbol{\delta}_{qr} \mid \boldsymbol{\delta}_q, choice_{qr}) \Pr(\boldsymbol{\delta}_q \mid choice_{qr}) \Pr(choice_{qr})}{\Pr(\boldsymbol{\delta}_{qr} \mid \boldsymbol{\delta}_q) \Pr(\boldsymbol{\delta}_q)}$$

(v) $$\frac{\Pr(\boldsymbol{\delta}_{qr} \mid \boldsymbol{\delta}_q, choice_{qr}) \Pr(choice_{qr} \mid \boldsymbol{\delta}_q) \Pr(\boldsymbol{\delta}_q)}{\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}) \Pr(\boldsymbol{\delta}_{qr})}$$

(vi) $$\frac{\Pr(\boldsymbol{\delta}_{qr} \mid \boldsymbol{\delta}_q, choice_{qr}) \Pr(choice_{qr} \mid \boldsymbol{\delta}_q)}{\Pr(\boldsymbol{\delta}_{qr} \mid \boldsymbol{\delta}_q)}$$

To find a solution to $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$, the focal issue to be addressed now is to select the most suitable form. Certainly, the following selection criteria must be fulfilled: firstly, the knowledge derived from the mixture model must be considered to be furthering the learning process on this issue; and secondly, $\boldsymbol{\delta}_{qr}$ must act as a condition. By looking through all the six formulas, only the term (i) can meet both the criteria. Therefore, we consider

$$\Pr(choice_{qr} \mid \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) = \frac{\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(choice_{qr})}{\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr})}, \tag{5-7}$$

where $\Pr(choice_{qr})$ is the prior probability and has been estimated from the mixture model. Regarding the other term of the numerator, it is reasonable that $\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}, choice_{qr})$ could be interpreted as the likelihood of observing δ_q^{OBS} given the fact that q has chosen r and the expected journey time was δ_{qr}^{EXP} according to his/her entry. In this sense, this term actually corresponds to the journey time distribution of the individual q conditional on T_q^{ENT}, which is in essence the probability distribution of the variable δ_{qr}.

Let $f_{qr}(\delta_{qr} \mid \vartheta_{qr})$ represent the PDF of the distribution of δ_{qr}, where ϑ_{qr} denotes a vector of parameter(s). Thus, we could have

$$\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}, choice_{qr}) \approx f_{qr}(\delta_{qr} = \delta_q^{\text{OBS}} \mid \vartheta_{qr}). \tag{5-8}$$

Now our focus is shifted to learn the conditional PDF, $f_{qr}(\delta_{qr} \mid \vartheta_{qr})$.

5.3.2 Conditional journey time distribution

For each individual q, his/her journey time δ_{qr} may be following a certain distribution. Suppose that we have obtained a huge data sample of passengers' journey times between the $o\text{-}d$, which is collected in a given period. One possible way to learn about $f_{qr}(\delta_{qr} \mid \boldsymbol{\vartheta}_{qr})$ is to sort out each individual's journey time observations from the whole sample. If we could obtain a sufficiently large subsample for q, it would be most likely that q, as a frequent traveller, might always choose the same route. But obviously this is not suitable for every individual within the data sample or the passenger population.

An alternative way is to assume, for any passenger q ($q = 1,\ldots,n$, n is the sample size), that δ_{qr} is distributed according to the probability distribution of δ_r, but with its own measures of central tendency. By dint of the mixture model, we have already gained some knowledge about each of the component distributions, $c_r(\delta_r \mid \boldsymbol{\theta}_r)$, where $\boldsymbol{\theta}_r$ represents a vector of parameter(s) being estimated. In this regard, we are actually assuming that the variables $\delta_{1r},\ldots,\delta_{nr}$ $\forall r$ are independent and share the same statistical parameters $\boldsymbol{\theta}_r$, except for the location parameters.

In order to better illuminate this point, let us suppose, for example, that δ_r is normally distributed, *i.e.* $\delta_r \sim \mathcal{N}(\mu_r, \sigma_r^2)$, where μ_r and σ_r denote its mean and standard deviation, respectively. Based on the hypotheses stated in **Chapter 3**, both μ_r and σ_r, hence the distribution $c_r(\delta_r \mid \mu_r, \sigma_r)$, could be obtained from estimating the corresponding GM model relying on a data sample. Given the estimates of the mean and standard deviation (still denoted by $\hat{\mu}_r$ and $\hat{\sigma}_r$, respectively), we shall therefore believe that $\delta_r \sim \mathcal{N}\left(\hat{\mu}_r, \hat{\sigma}_r\right)$. Meanwhile, based on the current assumption (stated in the previous paragraph), the probability distribution of δ_{qr} is also considered to be Gaussian. That is, for all passengers within the sample data, the journey-time variables $\delta_{1r},\ldots,\delta_{nr}$ are homoscedastic, and would be assumed independently, normally distributed. Note that δ_{qr} and δ_r are not necessarily identically distributed. In this case, still, the standard deviation of δ_{qr} remains unknown, which would then be assumed to be the estimated value according to the GM model, that is, we would have $\hat{\boldsymbol{\vartheta}}_{qr} = (\delta_{qr}^{\mathrm{EXP}}, \hat{\sigma}_r)$, hence $\delta_{qr} \sim \mathcal{N}(\delta_{qr}^{\mathrm{EXP}}, \hat{\sigma}_r)$. In this way, the likelihood that the journey time of q would be δ_q^{OBS} can be roughly approximated to the probability density $f_{qr}(\delta_{qr} = \delta_q^{\mathrm{OBS}} \mid \delta_{qr}^{\mathrm{EXP}}, \hat{\sigma}_r)$, given the information of his/her entry time and trains' timetable.

5.3.3 Deriving updated posterior probabilities

Now the only term that remains unknown in the fraction of formula (5-7) is $\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr})$. Since $\boldsymbol{\delta}_{qr}$ and $choice_{qr}$ are independent, as stated by formula (5-6), we could also apply the law of total probability to $\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr})$:

$$\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}) = \sum_{r\in R} \Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(choice_{qr} \mid \boldsymbol{\delta}_{qr}),$$

namely,

$$\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}) = \sum_{r\in R} \Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(choice_{qr}). \tag{5-9}$$

As such, $\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr})$ is actually equivalent to the sum of the numerator of the right part of formula (5-7), which also guarantees the condition (5-2). Therefore, we have

$$\Pr(choice_{qr} \mid \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) = \frac{\Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(choice_{qr})}{\sum_{r\in R} \Pr(\boldsymbol{\delta}_q \mid \boldsymbol{\delta}_{qr}, choice_{qr}) \Pr(choice_{qr})}. \tag{5-10}$$

Given equation (3-18), *i.e.* $\Pr(choice_{qr}) = \omega_r$, and equation (5-8), all the terms in formula (5-10) can be computed depending on the knowledge that has been held. Therefore, for each individual passenger, the updated posterior probability of each alternative route being chosen is derived.

$$\Pr(choice_{qr} \mid \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr}) = \frac{\omega_r f_{qr}(\delta_{qr} = \delta_q^{\text{OBS}} \mid \boldsymbol{\vartheta}_{qr})}{\sum_{r\in R} \omega_r f_{qr}(\delta_{qr} = \delta_q^{\text{OBS}} \mid \boldsymbol{\vartheta}_{qr})}, \tag{5-11}$$

We use π_{qr}^{UMM} (in contrast to π_{qr}^{MIX}, *cf.* formula (3-22)) to represent the estimate of $\Pr(choice_{qr} \mid \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$ $\forall q = 1, \dots, n$, $\forall r \in R$, with the superscript 'UMM' indicating that is an **u**pdated estimate based on the result of a **m**ixture **m**odel.

Similar to $\boldsymbol{\Pi}_\Delta^{\text{MIX}}$, the updated set of posterior estimates are also enumerated in a $n \times N_R$ matrix, which we represent by

$$\boldsymbol{\Pi}_\Delta^{\text{UMM}} = \begin{pmatrix} \pi_{11}^{\text{UMM}} & \cdots & \pi_{1N_R}^{\text{UMM}} \\ \pi_{21}^{\text{UMM}} & \cdots & \pi_{2N_R}^{\text{UMM}} \\ \vdots & \ddots & \vdots \\ \pi_{n1}^{\text{UMM}} & \cdots & \pi_{nN_R}^{\text{UMM}} \end{pmatrix}, \tag{5-12}$$

with (*see next page*)

$$\pi_{qr}^{\text{UMM}} = \frac{\hat{\omega}_r f_{qr}(\delta_{qr} = \delta_q^{\text{OBS}} \mid \hat{\boldsymbol{\vartheta}}_{qr})}{\sum_{r \in R} \hat{\omega}_r f_{qr}(\delta_{qr} = \delta_q^{\text{OBS}} \mid \hat{\boldsymbol{\vartheta}}_{qr})}; \tag{5-13}$$

and evidently,

$$\sum_{r \in R} \pi_{qr}^{\text{UMM}} = 1. \tag{5-14}$$

The following section will compare the estimates between $\mathbf{\Pi}_{\Delta}^{\text{MIX}}$ and $\mathbf{\Pi}_{\Delta}^{\text{UMM}}$.

5.4 Implementation of the updating approach

It is the quality of the update – the extent to which $\boldsymbol{\delta}_{qr}$, as extra information, would modify the estimates of each individual passenger's choice probabilities – that is vital for the inference as well as understanding of the passengers' actual route choices. For illustrating what effect of such alteration of conditions would be, we follow up the seven cases of O-D pairs, which have been previously examined, and implement the proposed updating method by exploiting the estimates derived from **Chapter 4**. So far, for each of the single O-D networks, we have employed both GM and LNM distributions to fit their respective data sample of OJT^{OBS}. We will only take advantage of the model that performed relatively better than the competitor in each case (in light of the test results of gof). From the elected mixture model, we have already obtained $\mathbf{\Pi}_{\Delta}^{\text{MIX}}$, along with the model parameters. On that basis, we would then be able to update individually those estimated posterior probabilities by bringing in the further consideration about $\boldsymbol{\delta}_{qr}$.

With respect to the application of the mixture model, it might be confronted with completely different situations given different O-D cases and different number of alternative routes, especially for matching up route-labels with their real-world counterparts. Unlike that situation, in this section, the application of the update would be only to change the posterior probabilities of passengers' route choices for each alternative routes, wherein the demonstration per se would be analogous for all the O-D cases. For this reason, in this section, we present only one case, ***Case-1***: **Victoria – Holborn**, as an illustrative example, with the results of the other six cases (***Case-2*** – ***Case-7***) exhibited in **Appendix D**.

5.4.1 Data issue

Information of passengers' entry times is one of the essentials for the updating approach. However, the data samples used for the mixture model estimation, as has been declared in **Chapter 4**, were each retrieved from a processed data set where the entry time data for each individual was unavailable, but merely OJT^{OBS}. In that situation, the updating approach was not be able to be applied on the same samples.

Instead, we had to draw support from another sample of Oyster data, which details the timestamps of everyone's entry as well as exit. This data was gathered in a period of 28 consecutive days, from 6th February (Sunday) to 5th March (Saturday) in 2011; and it is confined to (a sample of) 5% of the Oyster journey records across the whole LU network (during the 28-day period). Given the 5%-sample data, which we represent by $\Delta_{5\%}$ (in contrast to Δ as the larger sample), for each of the O-D cases being considered, its valid OJT^{OBS} is also delimited by an upper outer fence. The value of this fence was set to be the same as that of the sample used for estimating the mixture models, rather than using the upper outer fence of the 5%-sample itself. This is because the sample size of the latter is much larger and that data was collected during a much longer period, which is believed to deliver a more representative statistical boundary. Moreover, although the sampled passengers may have several journey records presented in the sample data, different journey records made by the same individual were each regarded as an independent journey of the others. The sample size and the mixture model used for each of the O-D cases is briefly summarised in **Table 5.1** below.

Table 5.1 Summary of sample sizes and elected mixture models for seven case studies

Case-	N_R	The relatively better mixture model	Sample size of $\Delta_{5\%}$
1	2	LNM	105
2	2	LNM	89
3	2	LNM	140
4	3	GM	85
5	3	GM	92
6	4	GM	48
7	4	GM	42

5.4.2 An example based on *Case-1*

With the individual Oyster journey records of the valid 5%-sample of ***Case-1***, in total, 105 valid records were obtained in respect of this O-D within the period of AM Peak (07:00 *a.m.* – 10:00 *a.m.*), which is still denoted by $\Delta_{5\%}$. This section compares $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$ and $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ at both the individual and aggregate level, where $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$ represents the posterior estimates of $\mathbf{\Pi}_{\Delta}^{\mathtt{MIX}}$ for the sample data set $\Delta_{5\%}$, while $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ represents the corresponding updated choice probabilities. Note that for most of the sampled passengers, $\Delta_{5\%}$ did also contain multiple observations for each of them. However, we still assume that all the journey time observations are independent of each other. As such, the context is equivalent to that each passenger has completed only one journey; and more specifically, $\Delta_{5\%}$ were supposed to be a sample of 105 passengers.

Further to *Case-1*: Victoria – Holborn

Recall ***Case-1*** from **Chapter 4**. For this case study, we investigated the pair of O-D stations: **Victoria – Holborn**, which is connected by two indirect routes. Every passenger (still denoted by q) travelling between this O-D would have to transfer at either **Oxford Circus** (referred to as Route1, and labelled $r=1$) or **Green Park** (referred to as Route2, and labelled $r=2$). Details about this network has been described in **Section 4.3.1**.

Given $\Delta_{5\%}$, the OJT^{OBS} of each of the 105 passengers (each being denoted by OJT_q^{OBS} and represented by an orange cross, ×) are depicted against their entry times in **Figure 5.4** (*see next page*). It is noted that there was a 'gap' in the data between about 07:15 and 07:30 *a.m.* The main reason for this is that the data was sourced from the 5% of all the Oyster data sampled on the basis of a certain group of travellers on the entire LU network. It was possible that none of the sampled travellers for this O-D made journeys during that 15-minute interval. Similar situations also occurred to all the other O-D cases in this thesis, except ***Case-3*** and ***Case-4*** (*see* **Appendix D**).

Additionally, given the entry time of each individual sampled, the expected journey times $\delta_{qr}^{\mathrm{EXP}}$ of each passenger in the sample is also illustrated, with the purple triangles, ▲, and blue circles, ●, representing $\delta_{q1}^{\mathrm{EXP}}$ and $\delta_{q2}^{\mathrm{EXP}}$, respectively. The computation of $\delta_{q1}^{\mathrm{EXP}}$ and $\delta_{q2}^{\mathrm{EXP}}$ followed the steps of deriving the expected route-specific journey time (referring to formula (5-1); *see also* **Section 3.5.1**).

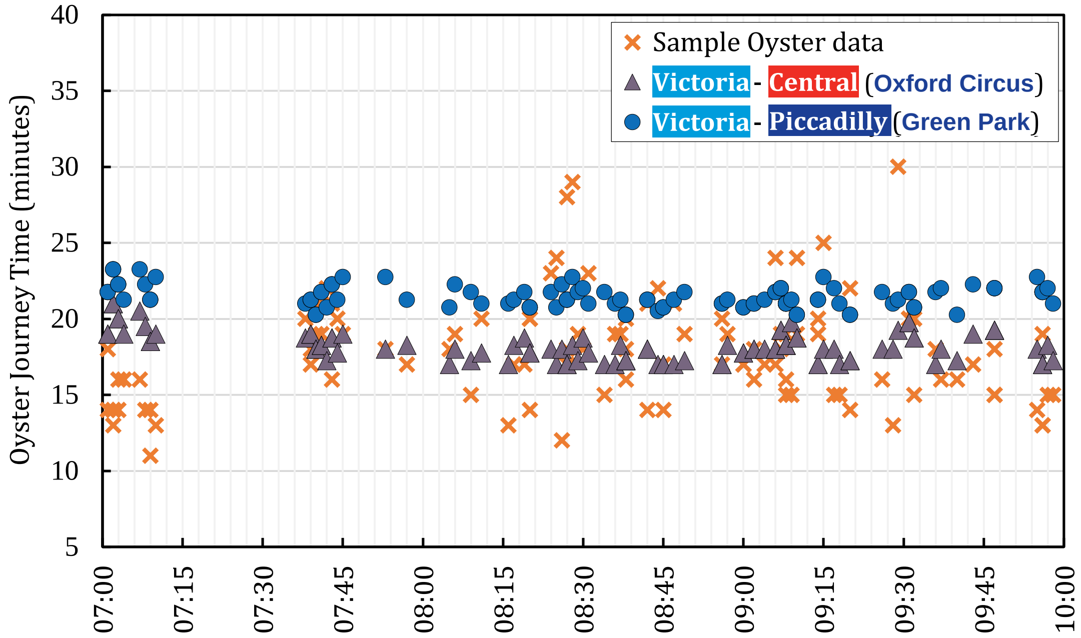


Figure 5.4 Comparisons between OJT_q^{OBS} and $\delta_{qr}^{\mathrm{EXP}}$ $\forall q, r$, given $\Delta_{5\%}$ for Victoria – Holborn.

As can be seen from **Figure 5.4**, given passengers' entry times (within AM Peak), $\delta_{q1}^{\mathrm{EXP}}$ and $\delta_{q2}^{\mathrm{EXP}}$ turned out to be bouncing between roughly 17 and 20 minutes, and between 20 and 23 minutes, respectively. These ranges also approximate the 95% CIs for ***Case-1*** (as shown in **Table 4.6**, p.88). Moreover, given different entry times, passengers' OJT^{OBS} fluctuated significantly; and their journey times might differ sharply given the same entry time. These facts have effectively verified our previous statements (*see* **Section 5.1**).

According to the test results of gof for this case (*see* **Table 4.7**, p.90), the LNM model was believed to have outperformed the GM, given the data set Δ containing a sample size of 24,760 individuals' OJT^{OBS}. On that basis, the estimates of the mixture weights (still denoted by $\hat{\omega}_r$; *see* **Table 4.6**) of the LNM was entered into the revaluation/update of each of the sampled individual's posterior route-choice probabilities.

In addition, in this case, the journey time of any passenger (still denoted by q) travelling by any of the two alternative routes (denoted by r) was treated as a random variable, denoted by δ_{qr} $\forall r = 1, 2$ and $\forall q$. And δ_{qr} was assumed to be log-normally distributed accordingly, with scale parameter being the same as the estimated scale parameter (denoted by $\hat{\sigma}_r$) of the r-th LNM component distribution. That is, there shall be two hypothetical log-normal distributions for each q. Note that the estimators of parameters for a log-normal distribution is different from a Gaussian case exemplified in **Section 5.3.2**. In this case, $\delta_{qr}^{\mathrm{EXP}}$ and $\hat{\sigma}_r$ are not parameters of the conjectural log-normal distribution. In this case, we use the symbols γ_{qr} and ς_r to represent, respectively, the location and scale parameters of the r-th hypothetical log-normal distribution for passenger q, In turn, we could represent the hypothetical distributions by $\delta_{q1} \sim \log \mathcal{N}(\gamma_{q1}, \varsigma_1)$ and $\delta_{q2} \sim \log \mathcal{N}(\gamma_{q2}, \varsigma_2)$.

Denote by $\hat{\gamma}_{qr}$ and $\hat{\varsigma}_r$ the parameter estimates. They should then be calculated, respectively, as follows (*cf.* Walck, 1996, p.86):

$$\hat{\gamma}_{qr} = \log\left((\delta_{qr}^{\mathrm{EXP}})^2 \Big/ \sqrt{\hat{\sigma}_r^{\ 2} + (\delta_{qr}^{\mathrm{EXP}})^2} \right) \tag{5-15}$$

and

$$\hat{\varsigma}_r = \sqrt{\log\left(1 + (\hat{\sigma}_r / \delta_{qr}^{\mathrm{EXP}})^2\right)}. \tag{5-16}$$

Given $\hat{\boldsymbol{\vartheta}}_{qr} = (\hat{\gamma}_{qr}, \hat{\varsigma}_r)$, according to formula (5-13), we would have

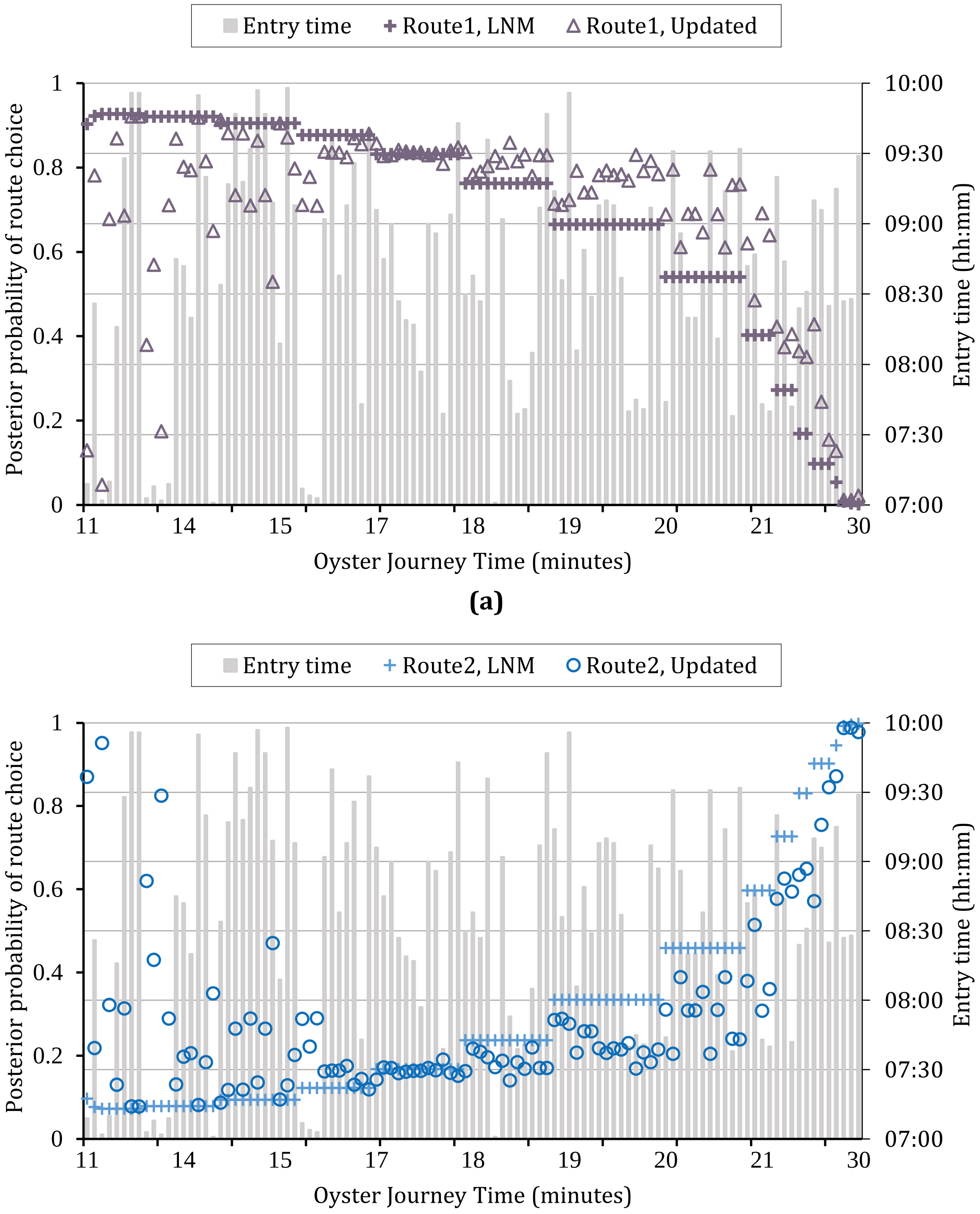
$$\pi_{q1}^{\mathrm{UMM}} = \frac{\hat{\omega}_1 f_{q1}(\delta_q^{\mathrm{OBS}} \mid \hat{\gamma}_{q1}, \hat{\varsigma}_1)}{\sum_{r=1}^{2} \hat{\omega}_r f_{qr}(\delta_q^{\mathrm{OBS}} \mid \hat{\gamma}_{qr}, \hat{\varsigma}_r)} \tag{5-17}$$

and

$$\pi_{q2}^{\mathrm{UMM}} = \frac{\hat{\omega}_2 f_{q2}(\delta_q^{\mathrm{OBS}} \mid \hat{\gamma}_{q2}, \hat{\varsigma}_2)}{\sum_{r=1}^{2} \hat{\omega}_r f_{qr}(\delta_q^{\mathrm{OBS}} \mid \hat{\gamma}_{qr}, \hat{\varsigma}_r)}, \tag{5-18}$$

whereby $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ could be gained according to equation (5-12). Note that here δ_q^{OBS} is equivalent to OJT^{OBS} of individual q.

Both $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{MIX}}$ (based on LNM) and $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ for this O-D pair (***Case-1***) are depicted in **Figure 5.5** (*see next page*), showing the differences between the two sets of posterior choice probabilities. The plus signs, coloured in **purple** in **Figure 5.5(a)** and **blue** in **Figure 5.5(b)**, illustrate, respectively, $\pi_{q,r=1}^{\mathrm{MIX}}$ and $\pi_{q,r=2}^{\mathrm{MIX}}$ on the basis of the sample data set $\Delta_{5\%}$.



(a)

(b)

Figure 5.5 Comparisons between π_{qr}^{MIX} (based on LNM) and π_{qr}^{UMM} for **Victoria – Holborn**:

(a) Route1: **Victoria** – **Central** (via **Oxford Circus**); and
(b) Route2: **Victoria** – **Piccadilly** (via **Green Park**).

The interval between the tick-marks on the horizontal axis spans 10 bars each relating to an individual/journey record in the Oyster data.

For comparison, the purple empty-triangles △ as well as the blue empty-circles, ○, illustrate, respectively, $\pi_{q,r=1}^{\mathrm{UMM}}$ and $\pi_{q,r=2}^{\mathrm{UMM}}$. Each of the symbols represents one observation of the data set $\Delta_{5\%}$. In addition, a grey bar indicates the related entry time of the passenger. Take, for instance, Route1. Given the same OJT^{OBS}, $\pi^{\text{MIX}}_{q,r=1}$ was a constant (for all passengers q). In contrast, $\pi^{\text{UMM}}_{q,r=1}$ could vary significantly as passengers' entry times differed.

Let us look further at one of the sampled individuals, labelled i, who entered **Victoria** station at $T^{\text{ENT}}_i = 07{:}39$ *a.m.* and exited from **Holborn** at 07:57 *a.m.* Then his/her journey time OJT^{OBS}_i was 18 minutes. According to the estimation result from the LNM model for this O-D, given OJT^{OBS}_i, the posterior probability of i choosing Route1 was $\pi^{\text{MIX}}_{i,r=1} = 76.3\%$, while that for Route2 was $\pi^{\text{MIX}}_{i,r=2} = 23.7\%$. On the other hand, given T^{ENT}_i, along with the information of timetable as well as average walking times for AEI between this O-D, the expected journey times that each alternative route for this passenger could be calculated as per formula (3-45). That is, an expected journey time for i travelling by Route1 were calculated to be $\delta^{\text{EXP}}_{i,r=1} = 19$ minutes; and $\delta^{\text{EXP}}_{i,r=2} = 21$ minutes by Route2. From this, intuitively, we could say that passenger i might be more likely to have chosen Route1, since OJT^{OBS}_i was less than and closer to $\delta^{\text{EXP}}_{i,r=1}$. This conjecture was also supported by the evidence that the mixture model estimate $\pi^{\text{MIX}}_{i,r=1}$ was much higher than $\pi^{\text{MIX}}_{i,r=2}$.

In order to justify the conjecture, we updated π^{MIX}_{ir} $\forall r = 1,2$ by taking into account the information about the differences between the OJT^{OBS}_i and δ^{EXP}_{ir} $\forall r = 1,2$. To this end, it was assumed that $\delta_{i,r=1}$ and $\delta_{i,r=2}$ were each following a log-normal distribution. The distribution parameters were calculated using formulas (5-15) and (5-16), given $\hat{\sigma}_r$ (estimated from the LNM model, where $\hat{\sigma}_1 = 2.4$ and $\hat{\sigma}_2 = 4.4$; *see* **Table 4.3**) as well as δ^{EXP}_{ir} $\forall r = 1,2$. In this case, $\delta_{i,r=1} \sim \log\mathcal{N}(2.9, 0.1)$ and $\delta_{i,r=2} \sim \log\mathcal{N}(3.0, 0.2)$. Note again that the journey time variable δ_{ir} might possibly follow any other probability distributions in reality; however, for simplicity, we considered it being log-normally distributed only in the scope of his thesis (*cf.* **Section 5.3.2**).

Then the updated choice probabilities could be derived from calculations based on formulas (5-17) and (5-18), respectively, where the estimated LNM weights were also involved (*i.e.* $\hat{\omega}_1 = 69.1\%$ and $\hat{\omega}_2 = 30.9\%$; *see also* **Table 4.3**, p.84). For passenger i, $\pi^{\text{UMM}}_{i,r=1} = 81.5\%$ and $\pi^{\text{UMM}}_{i,r=2} = 18.5\%$. $\pi^{\text{UMM}}_{i,r=1}$ was greater than $\pi^{\text{MIX}}_{i,r=1}$ while $\pi^{\text{UMM}}_{i,r=2}$ was less than $\pi^{\text{MIX}}_{i,r=2}$. Evidently, this result further justified the conjecture that i might be more likely to have chosen Route1.

It is noticeable that the application of the proposed updating method caused some sharp reversals of the choice probabilities for those faster travellers. That is, π_{qr}^{UMM} was diametrically opposed to π_{qr}^{MIX} that were estimated from the LNM model. And among those quickest journeys (with journey times being less than, say, 15 minutes), the updating had dramatically altered, or rather, decreased the probabilities π_{q1}^{MIX}. A possible reason was that when OJT^{OBS} was small, both the hypothetical distributions of δ_{q1} and δ_{q2} might suggest that there was a small likelihood of choosing either route given the OJT^{OBS}. As a consequence, both of $f_{q1}(OJT^{\mathrm{OBS}} \,|\, \hat{\gamma}_{q1}, \hat{\varsigma}_1)$ and $f_{q2}(OJT^{\mathrm{OBS}} \,|\, \hat{\gamma}_{q2}, \hat{\varsigma}_2)$ were rather small. In that case, if the former were slightly less than the latter, then that would result in a huge difference between π_{q1}^{UMM} and π_{q2}^{UMM}, since their sum should be equal to one (*see also* formulas (5-13) as well as (5-14)). It must be recognised that this is actually a drawback of the proposed updating approach, which would potentially bias the *naive* inference of passenger traffic distribution between this O-D (*cf.* **Section 3.4.1**). Notwithstanding, But for future research, a possible way to improve it could be to test different probability distributions for each passenger for each alternative route.

Besides, as OJT^{OBS} became longer, the updated choice probabilities π_{q1}^{UMM} would be much higher than π_{q1}^{MIX}, given the corresponding entry times. This could be reasonable, because the estimated 95% CI upper boundary of the mean journey time of Route1 was nearly 22 minutes, which may imply that the sampled passengers might experience longer journey time on Route1 (as well as on Route2) in the context of rush hour. In that sense, the update, to some extent, might also reflect the impact of passenger-traffic congestion or service delay, which possibly lead to passengers' boarding failures.

Moreover, aggregate measures are presented in **Table 5.2** (*see next page*), where ω_r^{ROD}, ω_r^{INF} and ω_r^{UPD}, respectively, represent the proportion of respondents who chose route *r* according to the RODS result (up to 2010), the proportion of passenger-traffic on route *r* given *effective* inference from the mixture model and that according to updated choice probabilities.

Table 5.2 Proportion of passenger traffic for each alternative route on Victoria – Holborn

In this case, ω_r^{INF} is calculated on the basis of LNM model estimates.

	Sample size	**Proportion of passenger-traffic** (%)	
		Victoria – Central Oxford Circus $r=1$	Victoria – Piccadilly Green Park $r=2$
ω_r^{ROD}	526	71.3	28.7
$\hat{\omega}_r$	24,760	69.1	30.9
ω_r^{INF}	24,760	69.2	30.8
ω_r^{INF}	105	69.8	30.2
ω_r^{UPD}	105	69.7	30.3

As can be seen from **Table 5.2**, ω_r^{UPD} and ω_r^{INF} are almost the same before and after the update. The estimates derived from the larger sample of journey times modelled by mixture distribution were retrieved from the much smaller sample. Thus the updating method is believed not to affect the inference of passengers' average choice probabilities of different alternative routes.

5.5 Summary and conclusions

On the basis of the mixture model of passengers' journey times, this chapter proposes an approach to update the previously estimated choice probabilities for each individual passenger. The update is achieved by taking into account additional information about the occurrence of $\boldsymbol{\delta}_{qr}$, which refers to a conditional expected average journey time of each route for each passenger. In that way, the posterior probability $\Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q)$ estimated from the mixture model in **Chapter 4** has now been updated to a newly formed posterior route-choice probability, *i.e.* $\Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$. It should be particularly noted that it is the prior probability, $\Pr(choice_{qr})$, rather than the posterior probability per se, that directly enters the calculation of $\Pr(choice_{qr} \,|\, \boldsymbol{\delta}_q, \boldsymbol{\delta}_{qr})$, where $\Pr(choice_{qr})$ is an estimate of the mixture weight for a mixture model. However, the estimation of $\Pr(choice_{qr})$ is reliant on the posterior estimates. Thus, the mixture model in effect provides prior knowledge for updating the posterior choice probabilities.

This extra condition $\boldsymbol{\delta}_{qr}$ was derived given the following assumptions. First, every passenger's journey time is examined by a set of hypothetical probability distributions, each of which is based on a situation that the passenger always chooses one of the alternative routes. The number of such distributions for a passenger was in line with the size of route-choice set for the passenger. In that way, the likelihood that an observed data δ_q^{OBS} was from a certain route was learnt for each individual passenger, which is distinct from the likelihood function $\Pr(\boldsymbol{\delta}_q \,|\, choice_{qr})$ considered in the mixture models. Additionally, it was assumed that consistent punctuality for all trains was assured, in which case the timetable data is practical for use. Third, all station facilities (*e.g.* passages, ticket gates) were assumed to remain unchanged. These assumptions thus made allow the calculation of the expected average journey time of each alternative route for every passenger, given the observation of the individual's entry time.

With the use of detailed Oyster card data gathered from the seven O-D pairs studied in **Chapter 4**, for each case, a comparison is made between the choice probabilities before and after the update. It must be pointed out that a major issue here is the inconsistency of the observation period of the two data sets: the data used for demonstrating the updating of choice probabilities and the OJT^{OBS} samples used in estimating the mixture model. This is due to the shortage of the detailed individual Oyster data. Notwithstanding, at the aggregate level, the average shares of passenger traffic distribution among alternative routes presents little difference between the two scenarios. In view of the limited sample size of the detailed data, the mixture model shows high adaptability for estimation of aggregate measures. At the individual level, passengers' choice probabilities will fluctuate significantly as their entry times vary. To some extent, such differences demonstrate that the influences of the additional condition, as well as reflects a more realistic range of individual taste variance in different alternative routes. Still, there is not convincing evidence that demonstrate whether the update exerts positive or negative influence on the learning of passengers' probabilistic route choices. This will be further tested in the next chapter.

Chapter 6
A latent route choice model

6.1 Introduction

From the previous chapters, we have already contrived to obtain two different sets of posterior probabilities of passengers' route choices between any given pair of multi-route O-D. Based on the GM and LNM models, the probabilities that an individual might have chosen each of the alternative routes have been derived for all passengers in a data sample of their actual journey times, which we represented by Δ. That set of estimates was represented by $\mathbf{\Pi}_{\Delta}^{\mathtt{MIX}}$ (derived from **Chapter 3** and **Chapter 4**), and later by $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$ relating to the detailed individual data sample $\Delta_{5\%}$ (in **Chapter 5**). On that basis, such posterior probabilities for each individual have been updated in light of supplementary knowledge on the expectation of their journey times that were deterministically calculated. The updated set of posterior estimates was then represented by $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$. Nonetheless, the extent to which either of the two sets can reflect the passengers' true choice probabilities has not been evaluated, though, theoretically, $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ should be more sensible than $\mathbf{\Pi}_{\Delta}^{\mathtt{MIX}}$. In other words, the credibility of those estimated posterior probabilities may not be fully guaranteed.

As mentioned in **Chapter 3**, the true choice probability that a passenger q would place on a route r would essentially be due to his/her own personal propensity (*cf.* **Section 3.2**). From the theory of random utility models (McFadden, 2000), q may be more willing to choose r if he/she perceives it to have a relatively higher utility than other alternatives, and will thus choose the one that offers the highest utility. It is noted, however, that the estimates of those posterior probabilities of q having likely chosen r were actually compliant with Bayes' theorem. Moreover, it in effect quantifies a subjective degree of our belief about the occurrence of the route-choice event, irrespective of how/why the route choices were actually made by q. Notwithstanding this irrelevance, ideally, we would still expect that the posterior estimates of the passengers' choice probabilities for each route would be as close to the true values as possible.

Nevertheless, to gain insight into the true choice probabilities would then entail the modelling of the passengers' decision-making process per se. We are thus again led to think of the means of discrete choice models, as it fundamentally takes into consideration a range of factors that relate to passengers' travel behaviour. Depending on the model specification, the parameter estimates of those factors may shed light on how/why passengers would choose a specific route. An understanding of such route choice behaviour of the passengers is of great interest to us; and it would also be a valuable asset for effective planning of local public transport (*cf*. **Section 1.1**). Yet, again, the development of such route choice models, or more specifically, the estimation of the models' parameters, would necessarily be reliant on observations of each individual's actual route choice, which, however, are not available in our case. In this regard, the conventional process for estimating the discrete choice models would be suspended for the lack of the route-choice data (*cf*. **Section 2.5**).

In such a context, this chapter pursues a route choice model, which will cope with the passengers' route choices within the probabilistic setting, rather than the actual route-choice observations. It is hereby referred to as a latent route choice model, wherein the term 'latent route choice' is interpreted to mean that the passengers' actual route choices are not observed (or not observable), but could be known up to a choice probability; and such probabilities of all alternatives correspond to the posterior probabilities being estimated otherwise.

This chapter is intended for two objectives of equal importance. On the one hand, we seek to assess the previously estimated posterior probabilities of route choices and to validate the updating approach proposed in **Chapter 5**. On the other hand, we also aim to gain an understanding of why passengers would choose a specific route between any given O-D pair. Accordingly, this chapter is to develop a latent route choice model, with the posterior probabilities of passengers' route choices being used as input into the representation of choice probability as well as the estimation of the choice model. Therefore, whether the posterior probabilities are trustworthy would largely depend on whether the latent choice model could yield meaningful estimates of relative sensitivities to explanatory variables that are specified.

To these ends, the rest of this chapter is arranged as follows. **Section 6.2** presents a brief review on the choice modelling techniques. The latent choice model is then introduced in **Section 6.3** that illuminates the idea of how the previously estimated posterior probabilities play a part in the model estimation. **Section 6.4** presents two empirical examples of applying the proposed approach by estimating a simple multinomial logit model, with use of the detailed Oyster data on the LU. At the end, a summary and conclusions are presented in **Section 6.5**.

6.2 Choice modelling

6.2.1 Standard logit choice probability

Any perceptible changes in the transit services (concerning *e.g.* frequency of lines/trains, transfer cost as well as accessibility of passageways) between a given O-D might easily have influence on passengers' route choice behaviour. Discrete choice models have certainly been the predominant approach to understand such behavioural process as to how the route-choice decisions are made.

In contrast to the route-choice modules embedded in most deterministic transit assignment models, which typically minimise the passengers' generalised travel cost function (*cf.* **Section 2.2.2**), discrete choice models, however, look at their perceived 'utilities' of each alternative route, with the specification of utility functions. As such, in general, a utility function measures the 'attractiveness' of a particular route – relative to its alternatives – to each individual passenger. Based on the premise that a passenger would always seeks the most attractive route to him/her, only the route that can offer the highest, or the maximum utility will be chosen by the passenger.

Let us keep looking at the *o-d* network with N_R alternative routes (illustrated in **Figure 3.1**, in **Section 3.2**). In this background, we could let U_{qr} denote the utility that passenger q perceives he/she may gain from choosing route r; and it can be specified in the simplistic form as follows:

$$U_{qr} = V_{qr} + \varepsilon_{qr},$$

where V_{qr} expresses the deterministic utility of r, and ε_{qr} acts as an error term. Intrinsically, V_{qr} is linked to a number of factors that potentially affect q's decision on whether to choose r, which could be related to various attributes of the route r (*e.g.* their transit services) and of the passenger q per se. We shall treat these factors as quantitative random variables that can be numerically measured for each individual passenger. As such, V_{qr} could be further expressed as a function of the variables, which is parameterised by a vector of coefficients. In reality, there must be some other factors that also exert impact on the true utility perceived by q but might not (be able to) be represented by V_{qr}. Concerning those unknown/unobservable factors, they are then ascribed to ε_{qr} as a completely random variable.

As the most popular discrete choice model, the logit model is based on the premise that ε_{qr} $\forall q \in Q$, $\forall r \in R$ are independent and each following the type I extreme value distribution (*cf.* Train, 2009, p.34). This assumption then serves as the necessary and sufficient condition for the derivation of the standard structure of logit probability formula, which, in the context of this chapter, could be expressed as follows:

$$P_{qr} = \frac{\exp\left(V_{qr}\right)}{\sum_{r} \exp\left(V_{qr}\right)}.$$

This is the probability of passenger q choosing route r.

6.2.2 Route choice models

A variety of applications of discrete choice models for route choices have been developed by looking into many factors (*e.g.* travel time variability, fare) that may have effect on the travellers' route-choice behaviour, which allow for varying degrees of responses of the individual passengers. Prato (2009) conducted a comprehensive review of the choice modelling approaches. A range of route choice models with diverse modifications on the structure of the standard logit formula were surveyed in the context of the route choice on road traffic network.

Given a multi-route O-D pair, alternative routes may potentially correlate with each other due to their overlaps, in which cases a station or some route sections might be shared by more than one transit lines (*e.g.* the overlap between the **Circle** and **District** lines on the LU). Without consideration on such correlated issues, the estimates of the sensitivities to attributes affecting route choice may be biased when developing the relative choice models. For modelling this common lines problem and approximating the correlation among these routes, based on the simple logit structure, correlation terms have been introduced into the utility functions, where amendments are made to the deterministic part of the function (Cascetta *et al.*, 1996; Ben-Akiva and Bierlaire, 1999; and Bovy *et al.*, 2008). Cascetta *et al.* (1996) firstly brought forward a **c**ommonality **f**actor (CF), which is used to capture the similarity between each route and its alternatives within the choice set. Each CF is associated with each route choice. From this, the degree of the similarity could be measured. On the basis of the standard logit structure, the choice probability function of C-Logit is specified as follows:

$$P_{qr} = \frac{\exp\left(V_{qr} + \beta^{\mathrm{CF}} \cdot \mathrm{CF}_{qr}\right)}{\sum_{r} \exp\left(V_{qr} + \beta^{\mathrm{CF}} \cdot \mathrm{CF}_{qr}\right)},$$

where β^{CF} represents the coefficient of the CF as an additional variable. It is supposed to be negative, so as to indicate the utility of a route is in inverse proportion to that of the other alternative routes.

Ben-Akiva and Bierlaire (1999) took into account a **p**ath **s**ize (PS) attribute for each alternative route, given those alternatives overlap, or rather, share some route sections. The PS, as a correction factor, enters the deterministic part of the utility associated with each route, which result in the original logit choice probability turning into the following expression:

$$P_{qr} = \frac{\exp\left(V_{qr} + \beta^{\mathrm{PS}} \cdot \log(\mathrm{PS}_{qr})\right)}{\sum_{r} \exp\left(V_{qr} + \beta^{\mathrm{PS}} \cdot \log(\mathrm{PS}_{qr})\right)}.$$

where β^{PS} is to be estimated as the parameter of the PS. Such a choice model is termed thus **p**ath **s**ize **l**ogit (PSL). Further, Bovy *et al.* (2008) updated the PS with a path size correction term, in which case the model is known as **p**ath **s**ize **c**orrection **l**ogit (PSCL) and may yet yield similar estimation results to PSL. It is noted that the correlations between alternative routes could only be partially explained by PSL or PSCL.

Likewise, more intuitive corrections to the utility function hence the choice probability have also been made to improve the interpretation through more advanced generalised extreme value models. Typical examples include paired combinatorial logit model and the more general cross nested logit model. Both were adapted for route choice modelling by Prashker and Bekhor (1998), and the generalised nested logit model by Bekhor and Prashker (2001). Still, both of the follow issues remain: restricted taste variation and disability of handling with temporally correlation in error terms. Besides, models within mixed logit structure, *e.g.* multinomial probit model (Daganzo and Sheffi, 1977) and logit kernel approaches (Bekhor *et al.*, 2002), are computationally expensive due to their choice probabilities taking a non-closed form.

The subnetwork model, which is an error components logit model developed by Frejinger and Bierlaire (2007), considers that the correlation between different routes is primarily caused by overlapping route sections of key routes. It is noted that such correlations involve not only physical overlapping but also perceptual relevance. The context, though, was specific to road networks.

As noted above, the evolution in the discrete choice models for route choices rely on researchers to customise the modification of and to improve the structure of the logit probability term. But all in all, to estimate the parameters for the variables specified in those models would essentially still depend on the observations of travellers' actual choices. In other words, the estimation of the models requires availability of data of each individual's route choice that either is stated or has actually been made. In this regard, a shortage and/or an absence of the route-choice data may often be an obstacle for the model development.

6.2.3 Data for choice modelling

In a conventional way, as aforementioned, the estimation of a discrete route choice model is essentially reliant on us obtaining the data of each individual's actual route choice. On this account, collection of the data is often supposed to be a vital issue for the analysts to deal with. Either unavailability or shortage of the data would cause the model estimation to fail, which may be an obstacle to the model development.

As mentioned in **Section 2.4.2**, the route-choice data could be collected through two approaches. One is by conducting surveys verbally or in a written form, which could obtain the respondents' text descriptions; and the other turns to employ intelligent devices of passive monitoring, *e.g.* GPS tracking units, which automatically gathers digitised information. In practice, however, either approach could be very costly for gathering sufficiently large data samples. Occasionally, the data that has been acquired might be inexplicable due to a lack of accuracy or loss of crucial information. For instance, as stated by Bierlaire and Frejinger (2008), a GPS unit may track a traveller in terms of formatted geographical coordinates recording his or her routes, and as a consequence, the observed data would not yet be immediately interpretable. That is, effective data of route choices would have to be retrieved through certain conversion prior to its being put into estimating the models. Meanwhile, such manipulation itself might also induce error information unexpectedly, and hence biased model estimates.

With regard to road traffic networks, much progress has been made to tackle the aforementioned issues. For the purpose of narrowing down the differences between the observed data and the real choices on road traffic networks, initially, Ben-Akiva *et al.* (1984) assigned descriptive labels to choices of *e.g.* fastest or shortest route. Later, in the context of route choice of long-distance travels by car, Bierlaire *et al.* (2006) looked at 'aggregate observations' instead of exact data of route choices, which allows for several routes to correspond to one 'observation' given a shrunken choice set. In this case, survey respondents only need to list sequentially approximate locations that they passed through during the course of a journey, rather than the specific positions. A possible approach to forming a whole route that is the most likely actual choice is to assume the route sections as the shortest routes between each of these sequential location points. This concept was later formulated, by Bierlaire and Frejinger (2008), as 'Domain of Data Relevance (DDR)' that relates an area to a list of network elements including notional nodes/links, *etc*. It was then further illustrated by Chen (2013) and Bierlaire *et al.* (2013). On this understanding, more valid data becomes accessible since the precise information would not be indispensable, although explicit rules of delimitating a DDR is uncertain and would largely depend on specific situations in practice.

While much progress has been made to tackle issue with the indeterminate data of individual route choice in context of road traffic networks, no applications in particular to that on the public transport have been made, mainly due to its complexity and data accessibility.

6.3 Latent choice probability

This section describes a modelling approach to overcome the challenge of modelling route choices on public transport without observing the route choices. Consider the estimation process of the standard logit model. As for the traditional procedure, we are used to employ a binary indicator being 0 or 1 as an exponent of a passenger's choice probability, P_{qr}. In that case, for all the alternative routes that are surely not chosen by the passenger, the exponents must be equal to 0, while only the probability term associated with the chosen route is raised to the power of 1 (*cf.* Train, 2009, pp.60-63). Namely, the choice probabilities for each individual passenger's actual chosen route are finally entered into the likelihood function for estimating the model coefficients (denoted by a vector, $\boldsymbol{\beta}$). And more specifically, in the context of this thesis, the log-likelihood function of $\boldsymbol{\beta}$, given a data set, say Δ of sample size n, should be:

$$\log \ell(\boldsymbol{\beta};\Delta) = \sum_{q=1}^{n}\sum_{r=1}^{N_R} \alpha_{qr} \log P_{qr} \; .$$

where α_{qr} represents the binary indicator. From that, if an individual was observed to have chosen a certain route, denoted by i, it is anticipated from the model that the probability of the route i being chosen by passenger q would be as close as possible – though not exactly – to 1, given the estimates of $\boldsymbol{\beta}$.

In view of the fact that the route choices that passengers have actually made are unknown in our case, we have postulated in **Chapter 3** that each alternative route has its own probability of being chosen from Bayesian perspective. And such has been further estimated as being the posterior probability of a passenger choosing a given route that he/she might have actually chosen, which are expected to reflect the true individual preference on different alternatives. We now replace the 0-1 indicators in the contribution by passenger q or the likelihood function through a weighted average of the probabilities of all possible choices for that passenger, where the weights are given by the posterior probabilities. In the case where the route choices would be observed with certainty, a single one of these would be equal to 1, with all others being 0, bringing us back to the original log-likelihood function. Therefore, we expect that the choice model could reproduce as close to the true choice probabilities as possible. On this account, we shall weigh each of the exponentials of the observed utilities, $\exp\left(V_{qr}\right)$, in the logit choice probability by the corresponding estimates of the posterior probabilities (which we represent here by the π_{qr} as a general form).[1] Then, in the estimation of a latent choice model, P_{qr} would thus become a weighted average:

$$P_q = \frac{\pi_{q1}\exp\left(V_{q1}\right) + \cdots + \pi_{qN_R}\exp\left(V_{qN_R}\right)}{\exp\left(V_{q1}\right) + \cdots + \exp\left(V_{qN_R}\right)},$$

which is further generalised to

$$P_q = \frac{\sum_{r\in R}\pi_{qr}\exp\left(V_{qr}\right)}{\sum_{r\in R}\exp\left(V_{qr}\right)}, \tag{6-1}$$

In formula (6-1), $\pi_{qr}\ \ \forall q,r$ are the posterior probabilities of all passengers' chose routes.

This probability term can be interpreted as the likelihood of observing the actual route choice that is unknown to us. In other words, when we are predicting a given passenger's route choice that is being unobserved, P_q is supposed to be the probability with which the actual choice could be predicted. More specifically, we are predicting the choice with a probability of P_q; that is to say, we are having a probability of P_q to find out the real choice. Finally, P_q given by formula (6-1) will be entered into the likelihood function.

Thus, with the given data sample, Δ, we could estimate a number of model parameters (still denoted by $\boldsymbol{\beta}$) associated with the attributes of the alternative route (*e.g.* travel time, fare and interchange) based on the maximum likelihood estimation. The traditional procedures of deriving the likelihood function of $\boldsymbol{\beta}$, hence its log-likelihood function, would be adapted accordingly, as the likelihood function turns out to be

$$\ell(\boldsymbol{\beta};\Delta) = \prod_{q=1}^{n} P_q\,;$$

and the log-likelihood

$$\log \ell(\boldsymbol{\beta};\Delta) = \sum_{q=1}^{n} \log P_q\,.$$

[1] It could be either π_{qr}^{MIX} or π_{qr}^{UMM}.

6.4 Empirical examples on the London Underground network

To test the proposed approach, this section presents an empirical study, making use of the previously estimated posterior probabilities of route choices as inputs into choice model estimation. Relying on some notation used in **Section 3.5.1**, we employ the following notation to elaborate on how the latent route choice model works.

Notation:

$\pi_{qr}^{\texttt{MIX}}$	posterior probability that q chose route r (given $\delta_{q}^{\texttt{OBS}}$), estimated from a mixture model on a data set
$\mathbf{\Pi}_{\Delta_{5\%}}^{\texttt{MIX}}$	matrix that contains $\pi_{qr}^{\texttt{MIX}}$ for all observations in $\Delta_{5\%}$ (*cf.* $\mathbf{\Pi}_{\Delta}^{\texttt{MIX}}$ defined in **Section 3.3.2**)
$\pi_{qr}^{\texttt{UMM}}$	updated posterior probability that q chose route r (given $\delta_{q}^{\texttt{OBS}}$ and $\delta_{qr}^{\texttt{EXP}}$) based on $\pi_{qr}^{\texttt{MIX}}$
$\mathbf{\Pi}_{\Delta_{5\%}}^{\texttt{UMM}}$	matrix that contains $\pi_{qr}^{\texttt{UMM}}$ for all observations in $\Delta_{5\%}$
$t_{qr}^{\texttt{WLK}}$	total **walk**ing ($\texttt{WLK}$) time of passenger q's access at an origin station and egress at a destination station by using route r
$t_{qr}^{\texttt{WFD}}$	q's **w**aiting time to board a train **f**or **d**eparture ($\texttt{WFD}$) from an origin station by using route r
$t_{qr}^{\texttt{OBT}}$	q's total **o**n-**b**oard **t**ravel ($\texttt{OBT}$) time by using r
$t_{qr}^{\texttt{TIC}}$	q's walking time to **t**ransfer between platforms at an **i**nter**c**hange ($\texttt{TIC}$) station on r
$t_{qr}^{\texttt{WIC}}$	q's **w**aiting time to board a train for departure from an **i**nter**c**hange ($\texttt{WIC}$) station on r
$\mathbf{t}_{qr}$	vector that contains all travel time variables for q choosing r
$\beta^{\texttt{WLK}}$	coefficient of $t_{qr}^{\texttt{WLK}}$
$\beta^{\texttt{WFD}}$	coefficient of $t_{qr}^{\texttt{WFD}}$
$\beta^{\texttt{OBT}}$	coefficient of $t_{qr}^{\texttt{OBT}}$
$\beta^{\texttt{TIC}}$	coefficient of $t_{qr}^{\texttt{TIC}}$
$\beta^{\texttt{WIC}}$	coefficient of $t_{qr}^{\texttt{WIC}}$
$\boldsymbol{\beta}$	vector that contains all coefficients, each being associated with a travel time variable
V_{qr}	deterministic (or observable) portion of utility U_{qr}

(*Continued*)

Notation: (*Continued.*)

U_{qr} utility that q perceives he/she may gain from choosing r to make a journey

ε_{qr} error term in utility U_{qr}

$\Delta_{N_R=2}$ set of posterior probabilities of passengers' route choice for selected O-D pairs, each of which involves two alternative routes ($N_R = 2$)

$\Delta_{N_R\leq 3}$ set of posterior probabilities of passengers' route choice for selected O-D pairs, any one of which involves no more than three alternative routes ($N_R \leq 3$)

$\Delta_{N_R\leq 4}$ set of posterior probabilities of passengers' route choices for selected O-D pairs, any one of which involves no more than four alternative routes ($N_R \leq 4$)

$t_{qr}^{\mathtt{AEI}}$ total walking time of q's journey by using r, including **a**ccess and **e**gress, as well as his/her walking time for **i**nterchange ($\mathtt{AEI}$)

$t_{qr}^{\mathtt{WTT}}$ total **wa**i**t**ing **t**ime of q's journey by using r, including his/her waiting times at both the origin and interchange stations

v_r dummy variable that indicates whether r involves an interchange

$\beta^{\mathtt{AEI}}$ coefficient of $t_{qr}^{\mathtt{AEI}}$

$\beta^{\mathtt{WTT}}$ coefficient of $t_{qr}^{\mathtt{WTT}}$

$\beta^{\mathtt{I/C}}$ coefficient of v_r for **inter**c**hange**/non-interchange ($\mathtt{I/C}$)

6.4.1 Data description

The data is comprised of (a) the posterior route-choice probabilities (*i.e.* $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$ and $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{UMM}}$) for every passenger being sampled and (b) 'observed values' of the travel time variables for those individuals.

Firstly, $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$ and $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{UMM}}$, have been already obtained from **Chapter 4** and **Chapter 5**, respectively (*see also* **Appendix C** and **Appendix D**). With regard to the data for the explanatory variables, in practice, each individual's travel times along different journey segments were not observed for each of the seven O-D pairs under study, nor were they available from the smart-card, especially for the journeys involving interchanges from one line to another. On this account, we could utilise the calculated average travel times for each journey segment for each O-D pair (*see* **Section 4.3**).

However, it is obviously the case that for any given pair of O-D (still denoted by o-d), the walking times (including $t^{\mathtt{ACC}}_{q,l',o}$, $t^{\mathtt{TIC}}_{q,[l',l''],s}$ and $t^{\mathtt{EGR}}_{q,l'',d}$) as well as the on-board travel times (including $t^{\mathtt{OBT}}_{q,l',[o,s]}$ and $t^{\mathtt{OBT}}_{q,l'',[s,d]}$) were actually all assumed to be constants and equal for each passenger travelling between that O-D, but only the individuals' wait times (including $t^{\mathtt{WFD}}_{l',o,1}$ and $t^{\mathtt{WIC}}_{l'',s,1}$) differ (*cf.* **Section 5.1**). For this reason, we may put together all the smart-card samples from the seven O-D pairs, which provided us with a combined data set with 601 journey records in total. This thus allowed for all the observed values of the explanatory variables to be varied among all the sampled individual passengers. Since the study area covers a large portion of Zone-1 of the LU network (shown in **Figure 4.1**, p.72), the choice model to be estimated may reflect, to some extent, the passengers' route choice behaviour within that area.

6.4.2 Utility function specification

Given the combined data set, the travel time along different journey segments are taken into account as explanatory variables. It might be arguable that the factors that influence passengers' route-choice decisions are quite subjective and may not be quantitatively measured. Notwithstanding, given the data available in this thesis, it would be necessary to assume that passengers would base their choices on the travel times of different journey segments. On that basis, we would like to see the passengers' sensitivities to the travel time for different journey segments. Particularly, we would like to understand how much different the passengers' sensitivities towards travel times would be at the interchanges from that when they are travelling at the origin/destination stations. Therefore, the walking time and waiting time at the interchange station will also be considered separately.

Let $\mathbf{t}_{qr}$ denote a vector of observed variables with respect to route r . Given $r :\Leftrightarrow (o,l',s,l'',d)$, we can represent the travel time variable of each journey segment of route r by $\mathbf{t}_{qr} = (t^{\mathtt{WLK}}_{qr}, t^{\mathtt{WFD}}_{qr}, t^{\mathtt{OBT}}_{qr}, t^{\mathtt{TIC}}_{qr}, t^{\mathtt{WIC}}_{qr})$, where

$$t^{\mathtt{WLK}}_{qr} = t^{\mathtt{ACC}}_{qr} + t^{\mathtt{EGR}}_{qr} ;$$

$$t^{\mathtt{WFD}}_{qr} = \frac{t^{\mathtt{WFD}}_{l',o,1} + t^{\mathtt{WFD}}_{l',o,2}}{2} ; \text{ (\textit{see next page})}$$

$$t_{qr}^{\mathtt{OBT}} = t_{l',[o,s]}^{\mathtt{OBT}} + t_{l'',[s,d]}^{\mathtt{OBT}};$$

$$t_{qr}^{\mathtt{TIC}} = t_{[l',l''],s}^{\mathtt{TIC}};$$

and

$$t_{qr}^{\mathtt{WIC}} = \frac{t_{l',s,1}^{\mathtt{WIC}} + t_{l',s,2}^{\mathtt{WIC}}}{2}.$$

Note that $t_{qr}^{\mathtt{WFD}}$ and $t_{qr}^{\mathtt{ICW}}$ are considered as possible average waiting times of passenger q on route r, where every passenger may fail to board at his/her first attempt given the rush hour traffic (*cf.* **Section 3.5.1**).

Moreover, denote by $\boldsymbol{\beta} = (\beta^{\mathtt{WLK}}, \beta^{\mathtt{WFD}}, \beta^{\mathtt{OBT}}, \beta^{\mathtt{TIC}}, \beta^{\mathtt{WIC}})$ a vector of the coefficients, with each being associated with $t_{qr}^{\mathtt{WLK}}$, $t_{qr}^{\mathtt{WFD}}$, $t_{qr}^{\mathtt{OBT}}$, $t_{qr}^{\mathtt{TIC}}$ and $t_{qr}^{\mathtt{WIC}}$, respectively, and representing passengers' sensitivities to each of these time variables. In this example, we consider that those variables are linear in parameters. Thus, the observable utility could be specified as $V_{qr} = \boldsymbol{\beta} \cdot \mathbf{t}_{qr}$; and hence the utility function:

$$U_{qr} = \beta^{\mathtt{WLK}} t_{qr}^{\mathtt{WLK}} + \beta^{\mathtt{WFD}} t_{qr}^{\mathtt{WFD}} + \beta^{\mathtt{OBT}} t_{qr}^{\mathtt{OBT}} + \beta^{\mathtt{TIC}} t_{qr}^{\mathtt{TIC}} + \beta^{\mathtt{WIC}} t_{qr}^{\mathtt{WIC}} + \varepsilon_{qr}, \qquad (6\text{-}2)$$

It would be expected that all the coefficients would be negative values. The passengers' sensitivities to those specified travel time variables could then be learnt from further analyses of the estimates of the behavioural coefficients. Note that any transit-line specific constants or any line specific time coefficient is not specified in this case, this is because our data sample is rather small and the correlations between the transit lines and journey times could not be handled by the simple logit model.

6.4.3 Estimation results

On the basis of the utility function specified as formula (6-2), MNL models were then estimated for the three data sets, by using $\boldsymbol{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$ and $\boldsymbol{\Pi}_{\Delta_{5\%}}^{\mathtt{UMM}}$ as input data, respectively.

The difference between those data sets is as follows:

i. the first set involves data for the O-D pairs each having two alternative routes, denoted by $\Delta_{N_R=2}$;

ii. the second set further involve the data for the O-D cases with three routes, denoted by $\Delta_{N_R\leq 3}$; and (*see next page*)

iii. the third data set consider all the seven O-D cases that have been studied in the previous chapters, denoted by $\Delta_{N_R \le 4}$.

The estimation results, including estimates of the coefficients for all the travel time variables as well as their significance levels, are presented in **Table 6.1** below and **Table 6.2** (*see next page*) – the former results estimated using the mixture model estimates $\mathbf{\Pi}^{\text{MIX}}_{\Delta_{5\%}}$ and the latter using the updated choice probabilities $\mathbf{\Pi}^{\text{UMM}}_{\Delta_{5\%}}$.

Table 6.1 Estimation results for MNL models using the posterior probabilities derived from the mixture models

	Using $\mathbf{\Pi}^{\text{MIX}}_{\Delta_{5\%}}$								
N_R	$=2$			≤ 3			≤ 4		
n	334			511			601		
Log–likelihood	-186.67			-385.77			-496.14		
	Est.	***t*-stat.**		**Est.**	***t*-stat.**		**Est.**	***t*-stat.**	
β^{WLK}	-0.42	-1.65		-0.07	-0.60		-0.02	-0.27	
β^{WFD}	-0.12	-1.49		-0.07	-1.07		-0.08	-1.23	
β^{OBT}	-0.84	-5.23		-0.40	-6.97		-0.37	-7.69	
β^{TIC}	-1.97	-5.29		-0.85	-5.37		-0.66	-5.07	
β^{WIC}	0.12	0.37		-0.11	-0.72		-0.29	-2.16	
	Ratio	***t*-ratios**		**Ratio**	***t*-ratios**		**Ratio**	***t*-ratios**	
		(*vs*. 0)	(*vs*. 1)		(*vs*. 0)	(*vs*. 1)		(*vs*. 0)	(*vs*. 1)
$\beta^{\text{WLK}}/\beta^{\text{OBT}}$	0.50	1.66	-1.66	0.16	0.59	-3.01	0.07	0.26	-3.79
$\beta^{\text{WFD}}/\beta^{\text{OBT}}$	0.14	1.37	-8.35	0.18	1.01	-4.51	0.22	1.14	-4.03
$\beta^{\text{TIC}}/\beta^{\text{OBT}}$	2.34	6.94	3.97	2.13	7.13	3.77	1.79	6.26	2.77
$\beta^{\text{WIC}}/\beta^{\text{OBT}}$	-0.15	-0.37	-2.86	0.29	0.71	-1.78	0.80	2.05	-0.51
$\beta^{\text{WLK}}/\beta^{\text{WFD}}$	3.55	1.15	0.82	0.89	0.58	-0.07	0.30	0.27	-0.65
$\beta^{\text{TIC}}/\beta^{\text{WIC}}$	-15.93	-0.38	-0.41	7.42	0.67	0.58	2.24	1.67	0.92
$\beta^{\text{WLK}}/\beta^{\text{TIC}}$	0.21	1.53	-5.63	0.08	0.59	-7.05	0.04	0.26	-6.96
$\beta^{\text{WFD}}/\beta^{\text{WIC}}$	-0.96	-0.35	-0.72	0.64	0.68	-0.38	0.28	1.22	-3.19

Table 6.2 Estimation results for MNL models using the updated posterior probabilities

	Using $\mathbf{\Pi}^{\text{UMM}}_{\Delta_{5\%}}$								
N_R	$= 2$			≤ 3			≤ 4		
n	334			511			601		
Log–likelihood	$-$ 164.90			$-$ 343.09			$-$ 430.95		
	Est.	***t*-stat.**		**Est.**	***t*-stat.**		**Est.**	***t*-stat.**	
β^{WLK}	-0.20	-0.28		-0.53	-3.81		-0.50	-4.17	
β^{WFD}	-0.61	-4.01		-0.45	-4.61		-0.43	-4.71	
β^{OBT}	-1.14	-1.91		-0.52	-6.98		-0.49	-8.15	
β^{TIC}	-2.92	-1.97		-0.94	-5.26		-0.89	-5.91	
β^{WIC}	-0.20	-0.44		-0.58	-2.81		-0.53	-3.05	
	Ratio	***t*-ratios**		**Ratio**	***t*-ratios**		**Ratio**	***t*-ratios**	
		(*vs.* 0)	(*vs.* 1)		(*vs.* 0)	(*vs.* 1)		(*vs.* 0)	(*vs.* 1)
$\beta^{\text{WLK}}/\beta^{\text{OBT}}$	0.18	0.25	-1.15	1.02	3.37	0.07	1.02	3.61	0.08
$\beta^{\text{WFD}}/\beta^{\text{OBT}}$	0.53	1.86	-1.62	0.87	3.50	-0.50	0.89	3.58	-0.44
$\beta^{\text{TIC}}/\beta^{\text{OBT}}$	2.56	6.47	3.94	1.82	6.61	2.98	1.82	6.99	3.15
$\beta^{\text{WIC}}/\beta^{\text{OBT}}$	0.18	0.44	-2.06	1.13	2.73	0.31	1.08	2.91	0.21
$\beta^{\text{WLK}}/\beta^{\text{WFD}}$	0.33	0.27	-0.55	1.17	3.44	0.50	1.15	3.59	0.47
$\beta^{\text{TIC}}/\beta^{\text{WIC}}$	14.41	0.42	0.39	1.61	2.12	0.81	1.69	2.30	0.94
$\beta^{\text{WLK}}/\beta^{\text{TIC}}$	0.07	0.25	-3.31	0.56	3.28	-2.55	0.56	3.30	-2.57
$\beta^{\text{WFD}}/\beta^{\text{WIC}}$	3.01	0.44	0.29	0.77	2.86	-0.83	0.83	3.08	-0.64

By comparing the two tables of results, it is noticeable that the different sets of choice probabilities led to significantly different estimations in the choice models. This was mainly due to the fact that the updating process substantially altered $\mathbf{\Pi}^{\text{MIX}}_{\Delta_{5\%}}$ (though the aggregate measures, *e.g.* ω_r^{INF} relatively remained unchanged).

For the case of $\Delta_{N_R=2}$ in **Table 6.1**, the coefficient of waiting time at interchange stations were positive, which, though not significantly, might be unreasonable. Commonly, though not always, both β^{TIC} and β^{WIC} are expected to be negative (*cf.* Wardman *et al.*, 2001b). Also, the results indicated that passengers were much more sensitive to the walking time for transferring from one line to another, as well as the on-board travel time, whereas the disutility of waiting time at the origin station and that of the walking time for access and egress were insignificant.

For each of the data sets (given $N_R = 2$, $N_R \le 3$ and $N_R \le 4$), we could draw a comparison between the final log-likelihoods for two models estimated using the same data set (*i.e*. same amount of data). In general, it is shown that the model using the updated posterior probabilities, $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$, was achieved relatively better results than that using the mixture model estimates, $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{MIX}}$. This essentially proves that $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ is believed to be relatively more credible, hence more realistic, than $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{MIX}}$ derived from GM or LNM mixture models in our case, especially at the individual level. Therefore, the proposed updating approach in **Chapter 5** is validated.

In view of the estimation results of all the three settings presented in **Table 6.2**, the model fitted for the data set, $\Delta_{N_R \le 4}$, which examined all the seven O-D pairs, provided with the most significant estimates of coefficients. It is therefore regarded as the most suitable model among all being tested. This is what we expected because $\Delta_{N_R \le 4}$ contains more data and also has more variability that would help the model estimation. Now we focus only on the estimation results of this model. Firstly, it is noticeable, from the ratio $\beta^{\mathrm{WFD}} / \beta^{\mathrm{OBT}}$ that passengers are more sensitive to being travelling on-board than to waiting at the origin station. This may be explained by the fact that the trains might be often over-crowded during the morning with the rush hour passenger traffic for work. Whereas at the interchange stations, it showed the opposite, from the ratio, $\beta^{\mathrm{WIC}} / \beta^{\mathrm{OBT}}$, that waiting time for a connecting line is more undesirable than on-board travelling. This may largely due to the negative effect of the interchange per se. Such results notwithstanding, both of these differences are not significant given the sample being used for the model estimation, as the results of t-ratios (against 1) are rather small.

Besides, the results also show that the walking time between gatelines and platforms (including access and egress) and the on-board travel time of both journey legs have practically the same coefficient. Furthermore, the disutility associated with the platform-to-platform walking time at the interchange stations is nearly double that of on-board travel time (*see* $\beta^{\mathrm{TIC}} / \beta^{\mathrm{OBT}}$), where one-minute walking for transfer is the equivalent in disutility to 1.82 minutes of time spent travelling on board. This difference is highly significant as the t-ratio against 1 reaches about 3.15. Moreover, from the ratios of the coefficients of walking time to that of waiting time, *i.e.* $\beta^{\mathrm{WLK}}/\beta^{\mathrm{WFD}}$ and $\beta^{\mathrm{TIC}}/\beta^{\mathrm{WIC}}$, it is shown that one-minute of walking time is about 1.15 and 1.69 times as unpleasant as the time of waiting at the stations of origin and interchange, respectively. These, however, are not significant. As such, it indicates that passengers are more sensitive to walking than to waiting, especially for interchanges.

What is more, regarding the difference of passengers' sensitivities to the walking times between interchange and other journey segments (*see* $\beta^{\mathrm{WLK}}/\beta^{\mathrm{TIC}}$), the disutility for interchange walking is more or less twice as much as that for access/egress. Its t-ratio against 1 shows this difference is relatively significant. Similarly, as can be seen from the last row of **Table 6.2** (*see* $\beta^{\mathrm{WFD}}/\beta^{\mathrm{WIC}}$), it shows that one-minute of waiting time for transfer is more or less 1.2 times as much as the disutility equivalent of waiting at the start of the journeys, but this is not very significant.

On the whole, the estimation results are all interpretable; and this in turn demonstrates that the proposed latent route choice model is also applicable, by using the posterior probabilities instead of actual observations.

6.4.4 An extended example

Following the previous example, we illustrate, in this subsection, another MNL model given a different specification of the representative utility. Because of the availability of a direct route (*see **Case-3*** described in **Section 4.3.1.3**) in respect of our sample data, we included a dummy variable that indicates whether an alternative route involves an interchange or not. Let it be denoted by v_r. That is, $v_r = 1$ if r is an indirect route, and $v_{qr} = 0$ otherwise. Note that it is essentially equivalent to the variable v_h defined for formula (3-44) (*see* **Section 3.5.1**). Further to this, in addition to the total on-board travel time variable, we then considered only the total walking time and the total waiting time for each individual, which we represent, respectively, by t_{qr}^{AEI} and t_{qr}^{WTT}. More specifically,

$$t_{qr}^{\mathrm{AEI}} = t_{qr}^{\mathrm{ACC}} + t_{qr}^{\mathrm{EGR}} + t_{qr}^{\mathrm{TIC}},$$

with the superscript 'AEI' being short for '**A**ccess, **E**gress and **I**nterchange'; and

$$t_{qr}^{\mathrm{WTT}} = t_{qr}^{\mathrm{WFD}} + t_{qr}^{\mathrm{WIC}},$$ (*see next page*)

with ‘`WTT`’ being short for ‘**wait**ing **t**ime’.

Based on the above, the utility function in this case is as follows:

$$U_{qr} = \beta^{\mathtt{AEI}} t_{qr}^{\mathtt{AEI}} + \beta^{\mathtt{WTT}} t_{qr}^{\mathtt{WTT}} + \beta^{\mathtt{OBT}} t_{qr}^{\mathtt{OBT}} + \beta^{\mathtt{I/C}} v_r + \varepsilon_{qr} ,$$

where $\beta^{\mathtt{AEI}}$, $\beta^{\mathtt{WTT}}$, $\beta^{\mathtt{OBT}}$ and $\beta^{\mathtt{I/C}}$ represent, respectively, the coefficients of the variables $t_{qr}^{\mathtt{AEI}}$, $t_{qr}^{\mathtt{WTT}}$, $t_{qr}^{\mathtt{OBT}}$ and v_r.

By using each set of the choice probabilities, $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$ and $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{UMM}}$, we estimated these parameters, $\boldsymbol{\beta} = (\beta^{\mathtt{AEI}}, \beta^{\mathtt{WTT}}, \beta^{\mathtt{OBT}}, \beta^{\mathtt{I/C}})$, for each of the three samples, $\Delta_{N_R=2}$, $\Delta_{N_R\le 3}$ and $\Delta_{N_R\le 4}$ (as described in the previous example in **Section 6.4.3**).

The estimation results are presented in **Table 6.3** below and **Table 6.4** (*see next page*), respectively, in the same manner as the previous example (*cf.* **Table 6.1** as well as **Table 6.2**, pp.162-163).

Table 6.3 Estimation results for an additional example of MNL models using the posterior probabilities derived from the mixture models

	Using $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathtt{MIX}}$								
N_R	$=2$			≤ 3			≤ 4		
n	334			511			601		
Log–likelihood	− 186.94			− 385.14			− 491.50		
	Est.	***t*-stat.**		**Est.**	***t*-stat.**		**Est.**	***t*-stat.**	
$\beta^{\mathtt{AEI}}$	− 0.63	− 2.81		− 0.13	− 1.50		− 0.12	− 1.84	
$\beta^{\mathtt{WTT}}$	− 0.11	− 1.42		− 0.14	− 2.07		− 0.16	− 2.36	
$\beta^{\mathtt{OBT}}$	− 1.05	− 5.44		− 0.41	− 6.68		− 0.41	− 7.62	
$\beta^{\mathtt{I/C}}$	− 5.13	− 3.36		− 2.79	− 4.37		− 2.82	− 5.42	
	Ratio	***t*-ratios**		**Ratio**	***t*-ratios**		**Ratio**	***t*-ratios**	
		(*vs.* 0)	(*vs.* 1)		(*vs.* 0)	(*vs.* 1)		(*vs.* 0)	(*vs.* 1)
$\beta^{\mathtt{I/C}} / \beta^{\mathtt{OBT}}$	4.86	5.20	4.13	6.77	6.55	5.58	6.91	7.93	6.78
$\beta^{\mathtt{AEI}} / \beta^{\mathtt{OBT}}$	0.60	2.68	− 1.78	0.31	1.41	-3.16	0.29	1.75	− 4.37
$\beta^{\mathtt{WTT}} / \beta^{\mathtt{OBT}}$	0.11	1.34	−11.29	0.35	1.87	-3.47	0.39	2.09	− 3.33

For convenience, here we code-name the previous example, ‘***Test-1***’, and the current case, ‘***Test-2***’. Firstly, we compare each pair of the models fitted for the same data.

Table 6.4 Estimation results for an additional example of MNL models using the updated posterior probabilities

	Using $\mathbf{\Pi}^{\text{UMM}}_{\Delta_{5\%}}$								
N_R	$=2$			≤ 3			≤ 4		
n	334			511			601		
Log−likelihood	− 165.26			− 331.65			− 418.56		
	Est.	***t*-stat.**		**Est.**	***t*-stat.**		**Est.**	***t*-stat.**	
β^{AEI}	−0.60	−1.39		−0.26	−2.03		−0.36	−4.20	
β^{WTT}	−0.58	−4.15		−0.51	−4.64		−0.48	−4.80	
β^{OBT}	−1.53	−2.05		−0.71	−7.21		−0.68	−7.73	
$\beta^{\text{I/C}}$	−9.00	−1.36		−4.57	−4.38		−3.87	−4.73	
	Ratio	***t*-ratios**		**Ratio**	***t*-ratios**		**Ratio**	***t*-ratios**	
		(*vs*. 0)	(*vs*. 1)		(*vs*. 0)	(*vs*. 1)		(*vs*. 0)	(*vs*. 1)
$\beta^{\text{I/C}}/\beta^{\text{OBT}}$	5.89	3.60	2.99	6.45	7.23	6.11	5.73	7.86	6.49
$\beta^{\text{AEI}}/\beta^{\text{OBT}}$	0.39	0.90	−1.41	0.36	1.79	−3.15	0.53	3.35	−2.96
$\beta^{\text{WTT}}/\beta^{\text{OBT}}$	0.38	2.02	−3.28	0.72	3.76	−1.49	0.71	3.79	−1.52

Similar to ***Test-1***, for each sample, the model using the updated choice probabilities $\mathbf{\Pi}^{\text{UMM}}_{\Delta_{5\%}}$ also outperformed that using the original mixture model estimates $\mathbf{\Pi}^{\text{MIX}}_{\Delta_{5\%}}$ in ***Test-2***. This could again be due to the improvement in $\mathbf{\Pi}^{\text{UMM}}_{\Delta_{5\%}}$ compared to $\mathbf{\Pi}^{\text{MIX}}_{\Delta_{5\%}}$.

Furthermore, values of the AIC and BIC were calculated to demonstrate and compare the goodness of fits of the preferable models in ***Test-1*** and ***Test-2*** for each data set (*cf*. **Section 3.5.2**). The results are presented in **Table 6.5** below.

Table 6.5 Goodness of fit of models of ***Test-1*** and ***Test-2***

	Test-1	***Test-2***	***Test-1***	***Test-2***	***Test-1***	***Test-2***
N_R	$=2$		≤ 3		≤ 4	
n	334		511		601	
Dimension of $\boldsymbol{\beta}$	5	4	5	4	5	4
Log−likelihood	−164.90	−165.26	−343.09	−331.64	−430.95	−418.56
AIC	339.80	338.52	696.18	671.28	871.90	845.12
BIC	358.86	353.76	717.36	688.23	893.89	862.71

As illustrated in **Table 6.5**, the models of ***Test-2*** could always achieve a relatively lower value of AIC/BIC and hence have better fits than those of ***Test-1***. In the following discussions, we consider the case of the current example model being fitted for $\Delta_{N_R \leq 4}$.

As can be seen from **Table 6.4** (*see previous page*), the coefficients of all variables are negative as expected, which are also statistically significant. It should be noted that the coefficient $\beta^{\mathrm{I/C}}$ on the interchange/non-interchange dummy variable is actually independent of the amount of time spent interchanging (*cf.* Wardman *et al.*, 2001b). We may express the interchange and the other travel time variables as equivalent amounts of on-board travel time (*cf.* Wardman *et al.*, 2001a), where the ratio $\beta^{\mathrm{I/C}} \big/ \beta^{\mathrm{OBT}}$ is also termed 'interchange penalty'.

Early studies carried out by London Regional Transport (1988); and London Regional Transport (1995)[2], which were quoted by Wardman *et al.* (2001b), have already analysed the interchange penalties on the LU. Their analyses relied particularly on passengers' actual choices between direct and indirect routes. In both of the studies, an interchange penalty was considered, without including the walking and waiting components. Based on a data set in 1980, their initial finding showed that an average interchange penalty was 5.7 minutes. That is, one interchange would be perceived by a typical passenger as equivalent to 5.7 minutes of on-board travel time. In the later analysis, the value was updated to 3.7 minutes, given another data set available from 1990. Furthermore, given that the walking and waiting time variable is not involved in the utility function, the research conducted by Guo and Wilson (2011) showed the interchange penalty would be equivalent to 4.9 minutes of on-board travel time. More recently, the value for the LU published by the Transport for London (2013a) was 3.5 minutes; while the report by the Department for Transport (2014) indicated that the interchange penalty on wider public transport is 5 to 10 minutes of on-board travel time per interchange. As shown in **Table 6.4**, the ratio $\beta^{\mathrm{I/C}} \big/ \beta^{\mathrm{OBT}}$ obtained from our model suggests that the time value of the interchange penalty would be 5.73 minutes of on-board travel time, which appears plausible given the above values as references.

[2] London Regional Transport (1984–2000) is the predecessor to TfL.

It should also be noted, in the first two cases mentioned above, that the walking and waiting time were constrained to be weighted twice on-board travel time (*see also* Guo, 2008, p.47). Similarly, for the calculation of the generalised journey time on the LU (Transport for London, 2013a), walking time along congested passageways at an origin or a destination should also be weighted by 2.0, while at interchange stations the walking time would be weighted slightly higher, by 2.08. As for the time spent waiting for a train on a fairly crowded platform and standing in a crowded train, the weights could be as large as 4.0 and 2.03, respectively. Only being seated in an uncrowded train is not weighted. For this reason, during rush hour (or in a congested environment), one-minute walking and 1-minute waiting are roughly the equivalents in disutility to one minute and 1.97 minute of in-train time respectively. This is analogous to, though slightly different from, the estimation results in ***Test-1***, given $\mathbf{\Pi}_{\Delta_{5\%}}^{\mathrm{UMM}}$ and $\Delta_{N_R \leq 4}$ (*cf.* **Table 6.2**, p.163).

By comparison, in the current model, the estimation results showed quite the opposite. The coefficient of on-board travel time, β^{OBT}, was approximately twice (practically 1.9 times) that of the total walking time, β^{AEI}, and about 1.4 times that of the total waiting during a journey, β^{WTT}, though the latter difference was not significant. That is, the disutility associated with travelling aboard was nearly double that of walking. One possible explanation could be that passengers might be practically indifferent to the inevitable walk for access/egress, but may be more concerned with on-train delays and congestion. A train might possibly be stuck in a tunnel and/or take longer time than scheduled for loading/unloading passengers. Analogous situations could be found from the research conducted by Guo and Wilson (2011) and Raveau *et al.* (2014), where the walking and waiting time spent interchanging were both modelled as explanatory variables in the utility function.

6.5 Summary and conclusions

This chapter has presented a new approach to modelling passengers' route choice behaviour in a situation that each individual's actual chosen route is unobserved (or simply unobservable) but their probabilistic route choices are considered. That is, each passenger's route choice is only learnt and hence described by a set of posterior probabilities. All these posterior probabilities were estimated in line with Bayes' theorem, with each, as a conditional probability, expressing the probability of one alternative route being chosen by an individual, given the knowledge about his/her journey time. These posterior probabilities vary across respondents; and they are used instead of the simple deterministic 0-1 indicators typically used in a choice model. In other words, the numerator in, say, a multinomial logit choice probability would no longer just be the exponential of the utility of the chosen alternative, but would be a weighted average of exponentials of such utilities, where the weights are given by the posterior probabilities.

Testing of the proposed approach was conducted based on the MNL model. The estimation results, based on the posterior probabilities as inputs into choice model, have shown that we could estimate meaningful relative sensitivities to the different journey time segments, thus allowing us to obtain an understanding of the passengers of route choice even in the absence of observations of the actual chosen routes. This is a key step forward to overcome the shortage of **r**evealed **p**reference (RP) data for discrete choice analyses.

It must be pointed out, however, that the estimation results of the discrete choice models described and discussed in this chapter would depend crucially on the feasibility of acquiring credible (posterior) route-choice probabilities of each passenger in a given sample. On the other hand, there is still a need for the validation of the coefficient estimates of the latent route choice model. To this end, ideally, we should compare the results with previous/similar studies that are based on real RP data, where passengers' actual route choices are observed. Additionally, we may use such RP data, if available, to estimate the same, say MNL model. An alternative way could be that we may try to simulate data where passengers' actual route choices and the underlying sensitivities (to the specified variables) that determine both the choice set and the actual choice are known. Then, we also use that simulated data in the modelling framework to see whether it could be retrieved. By doing so, we could then compare the estimated coefficients from the posterior route-choice probabilities, and the estimates from the data of actual/simulated route choices. If they were close enough, then the development of the latent route choice model would ultimately be confirmed.

Moreover, the utility specification could be refined, and take into account *e.g.* the passengers' sensitivities to different transit lines, as some travellers may have strong preferences to a certain line while others may have different tastes. Also, the crowdedness as well as seat availabilities could be considered further. In addition, the testing of this proposed approach could be extended to the other advanced route choice models, such as the path size logit, C-logit, as well as the error components approach allowing for correlation between routes sharing key parts of the network.

Chapter 7
Concluding remarks and future research

7.1 Summary of the thesis

This thesis is devoted to making an attempt to develop a modelling approach towards passengers' route choice behaviour, where their actual route choices are unobserved/unobservable and hence latent. At best, the route choice of a passenger could only be known up to a choice probability. It is thus distinct from standard discrete choice model that requires the actual choice is explicitly known. The study is based on the LU system and the Oyster smart-card travel time data; it focuses on the mechanisms and modelling techniques to cater to the development of a latent route choice model. The work presented in this thesis provides fundamental solutions to the model configuration, whereby the implementation of the latent route choice model has been achieved under a modelling framework combining two building blocks: data mining and discrete choice modelling (*cf.* **Figure 1.1**, p.5). The outputs of the first building block provide the input data for the second one.

For the data mining, we utilised the methods of Bayesian inference in a bid to find out posterior probabilities of passengers' route choices between a given pair of O-D stations. This building block has three modules as follows.

(a) Data processing. It deals with all existing information from different data sources, especially the smart-card data that provides the entry times and journey times of each individual passenger on a given O-D pair.

(b) Finite mixture model. It produces a set of estimates of choice probabilities for each individual passenger, given their journey times being observed and the route-choice set being identified for each O-D. Additionally, proportions of the passenger flow on each alternative route are estimated as well.

(c) Update. It updates the posterior probabilities obtained from the finite mixture model for each individual, by further considering their entry times and trains' timetable, in addition to the actual journey times.

For the data processing, we consider the following additional information:

(1) timetable of each transit service;

(2) average walking time between gatelines and platforms, as well as that between any two platforms within each station; and

(3) historical route-choice data indicating the proportions of passenger-traffic flow among alternative routes.

For the finite mixture model, the passengers' journey times for each O-D are modelled by a finite mixture distribution. The prior knowledge, especially about the component distributions and their mixture weights, are of significant importance, as that could provide ideal initial values for the model estimation. In the case studies on the LU network, the information about the passenger-flow proportions of each route was available from the RODS data; however, it was not used as the initial value, but served as the only reference for validation of the estimated mixture weights (and the inference of passenger-flow proportions).[1] In this respect, we applied the *K*-means clustering method to estimate initial values for the parameters of the mixture model, which were then estimated by applying the EM algorithm. Since the estimation by itself does not show a one-to-one correspondence between an estimated component of the mixture and an alternative route in the real world, we put forward a set of principles for matching a component-label to a real alternative route, in order to interpret as well as validate the model estimates. Note that the interpretation and validation of the model estimates are crucial to determining whether the mixture distribution (or the model) would be suitable. This in turn largely depends on the credibility and accuracy of the expected average journey times of each alternative (calculated based on the information sources (1) and (2)) as well as the existing information about the traffic distribution (based on source (3)).

For the updating of the posterior probabilities of each individual passenger's route choices, we further considered the expected journey times that every one

[1] It was also pointed out that the RODS results were derived from aggregation on a rolling basis and hence may not be accurate.

might have for each alternative route, according to their actual entry times and the information sources (1) and (2) as mentioned above. From that, the posterior probabilities are re-estimated by allowing for two conditions, where the mixture weights estimated from the finite mixture model serve as prior knowledge. We demonstrated that the process was fully complying with Bayes' theorem.

Further, we presented seven case studies on the LU system, where these three modules were implemented in the context of different network scales. Within the scope of this thesis, the applicability of module is confined only to a single O-D network, and only GM and LNM models were tested for the LU case studies. Still, it has been demonstrated that the finite mixture model could offer an effective solution to estimating passengers' route choices at the aggregate level. It was also noted that GM had a relatively greater capacity than LNM in this context since it could always provide feasible estimates. Although LNM might fit the journey time data well in each case study, the model estimates were clearly less interpretable as the size of route-choice set becomes larger. What is more, at the individual level, we illustrated the posterior probabilities, which were estimated before and after the update was applied for each passenger. Then, the two sets of estimates were used as input data for estimation and development of the latent route choice model, which acts as the second building block of the established modelling framework.

Conventionally, a standard choice model relies on having real data of individuals' route choices, which are further coded as binary indicators, 0 and 1, and enter the choice probabilities for the estimation of the model coefficients. Suppose that we can know exactly which alternative route a passenger has actually chosen. The choice probability for the chosen route is equal to 1, while the probabilities for all other alternatives should be 0. It would then be expected that the standard choice model could predict the choice probability for this passenger to be as close to 1 as possible. However, each of the alternative routes may in fact have its own probability of being chosen for reasons. Suppose if all the alternatives are treated as having the same choice probability (*i.e.* $1/N_R$, where N_R is the total number of the alternative routes), the standard choice model would not be able to capture these true probabilities.

We modified the formula of choice probabilities for each individual passenger to be a weighted average of exponentials of the utilities of all alternative routes that might possibly have been chosen, rather than just the exponential of the utility of the chosen alternative. For each of the modified choice probabilities, the estimated posterior probabilities for each individuals served as the weights for the exponentials. In that way, the latent route choice model is expected to retrieve the posterior probabilities as those being put in, whereby the estimated model coefficients should thus present passengers' sensitivities that underlie and determine their actual choices in a more realistic manner.

We implemented the latent choice modelling approach based on a simple MNL model, using the two sets of posterior probability estimates derived from the first building block. Then we compared and analysed the two sets of estimated coefficients of the different journey time segments as we specified for the utility function. The outcome demonstrated that the updated posterior probabilities yielded meaningful coefficients.

It should be noted however that the credibility of the posterior estimates could not be assured because of the fact that the passengers' actual chosen routes are latent. In this sense, not only would the posterior probabilities serve as the input data for estimation of the latent route choice model, but also the estimation results of the latter would in turn serve as evidence to verify whether those are reasonable. Therefore, the first building block is expected to provide a set of posterior probabilities of choosing each of the alternative routes as realistic as it possibly can. The better the posterior probabilities are generated, the less uncertainty there would be in a choice model, and we would then gain a better understanding of passengers' route choice behaviour. In general, this research will have immediate practical implications to the underground network managers, and would be applicable to other cities with major and complex public transport networks.

7.2 Directions for future research

The modelling framework is still an incomplete structure where there is room for more experiments, practices and crucial modules to be included. Several important issues merit future research. On the strength of the established structure in this thesis (*cf.* **Figure 1.1**, p.5), we propose a relatively more robust version for future research, which is illustrated in **Figure 7.1** (*see next page*).

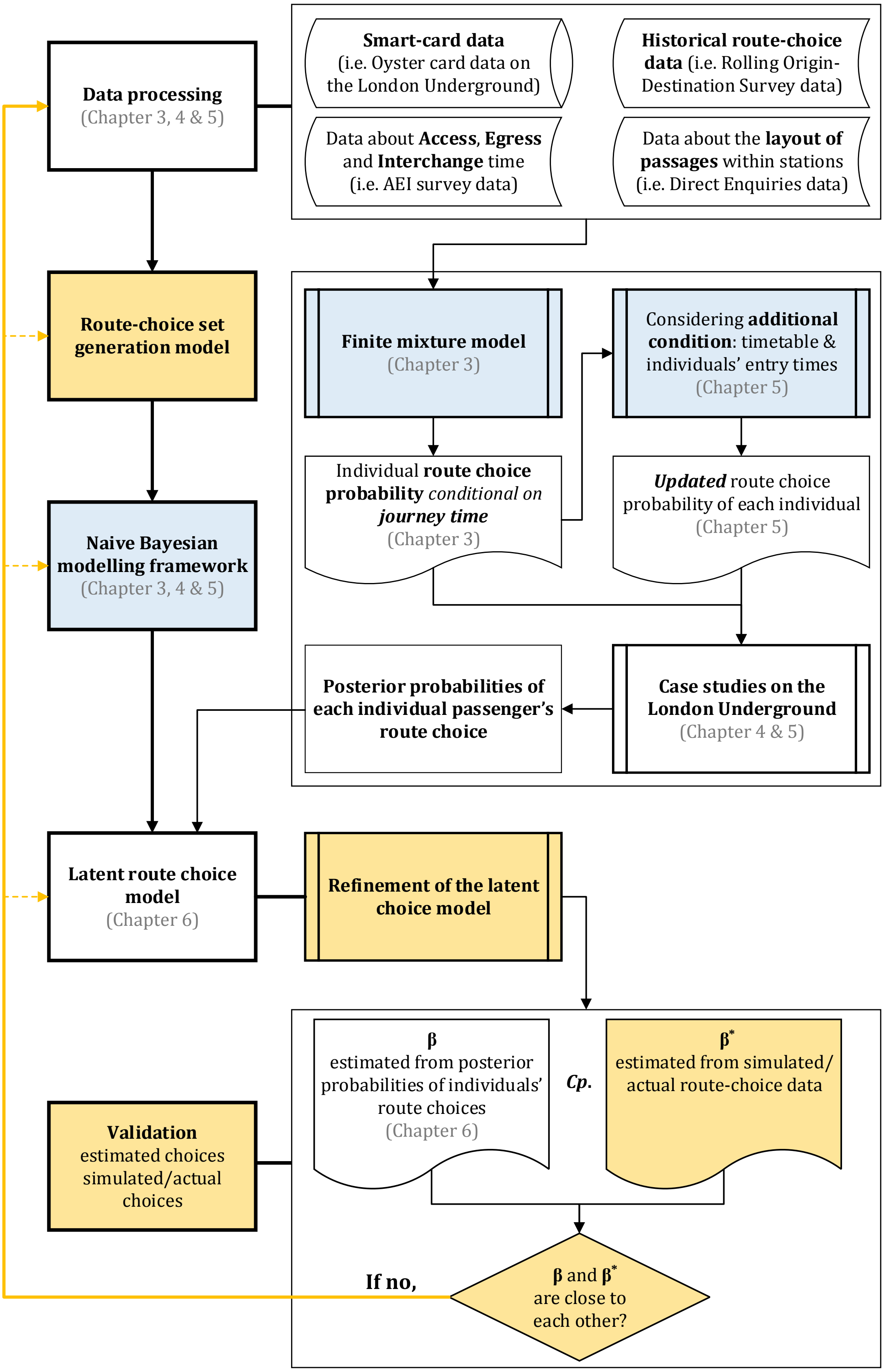

Data processing
(Chapter 3, 4 & 5)
Smart-card data
(i.e. Oyster card data on the London Underground)
Historical route-choice data (i.e. Rolling Origin-Destination Survey data)
Data about Access, Egress and Interchange time
(i.e. AEI survey data)
Data about the layout of passages within stations
(i.e. Direct Enquiries data)
Route-choice set generation model
Finite mixture model
(Chapter 3)
Considering additional condition: timetable & individuals' entry times
(Chapter 5)
Individual route choice probability conditional on journey time
(Chapter 3)
Updated route choice probability of each individual
(Chapter 5)
Naive Bayesian modelling framework
(Chapter 3, 4 & 5)
Posterior probabilities of each individual passenger's route choice
Case studies on the London Underground
(Chapter 4 & 5)
Latent route choice model
(Chapter 6)
Refinement of the latent choice model
β
estimated from posterior probabilities of individuals' route choices
(Chapter 6)
Cp.
β*
estimated from simulated/actual route-choice data
Validation
estimated choices
simulated/actual choices
If no,
β and β*
are close to each other?

Figure 7.1 A modelling framework to be developed for future research.

In this modified framework, improvements will be necessary for the modules coloured in **blue**. Furthermore, the modules coloured in **gold** (including the directed **golden** line) will also need to be built in order to provide robustness to the modelling framework.

7.2.1 Refinement of naive Bayesian modelling framework

The purposes of improving the naive Bayesian modelling framework is intended, on the one hand, to acquire posterior probabilities of passengers' route choices with better credibility; and also, on the other hand, to have a more realistic inference of passenger-traffic distribution. The more accurate the posterior probabilities are estimated to be, the closer such inference is to the truth.

The implementation of this model, in reference to the case studies conducted on the LU system, was predicated on the premise of some simplistic assumptions being made in each module. The major issues and possible solutions are summarised as follows.

The universal route-choice set of a given O-D pair had to be identified through our own judgement. As such, in practical applications, this would largely be dependent on modellers' own senses and perceptions of the O-D network and the possible alternative routes, rather than the passengers' perspective. In this regard, a model for generating a choice set will be indispensable. More details are presented in **Section 7.2.2**.

For testing of the finite mixture model in each of the LU case studies, we considered the component distribution, *i.e.* the journey time distribution of an alternative route, to be either Gaussian or log-normal. This was due to the true component distributions were not known. It would be better to test other, different types of statistical distributions hence different mixture models for fitting the journey time data of a given O-D. Moreover, a simulation-based transit assignment model may be employed, or developed for the estimation of the distribution of journey times. However, note that different behavioural assumptions of how passengers make route choices may have different impacts on the estimation of the latent route choice model. Therefore, we need also check that what assumption is the best or have least impact in this regard. Additionally, the mixture distribution of journey time for a given O-D pair may not be a standard mixture, but may virtually be a mixture of different types of components distributions, where each route has its own distributional form. Such advanced mixture models can also be studied in the future.

Furthermore, setting initial values and a stopping threshold for estimating the mixture model parameters may potentially pose some challenges for future research. Admittedly, there is no guarantee that the general EM algorithm converges to the global optimisation, though, which was not what we pursued either. Given a set of initial values, the estimated parameters might be different. The smaller the threshold value is, the more likely that the estimated results would be the global optimal, and the longer time the estimation would take. In contrast, a larger threshold may achieve a faster convergence, though, which would be more likely to a local optimal. Given a threshold value, different initial values may result in different estimates of the model parameters. The combined impact that the initial values and the threshold may have on the estimation results needs to be further assessed and analysed; and meanwhile, a variety of methods for the generation of initial values for the estimating models could also be tested.

What is more important is that it will be vital to develop a more effective algorithm for matching an estimated component from mixture models to a real-life alternative route. In this thesis, we had only suggested a set of general principles; and its practical application (*e.g.* in the LU case studies) were still indefinite and rather subjective.

Then, for the approach to the updating of the posterior probabilities, we considered that each individual passenger had a set of hypothetical journey time distributions, each of which was based on the premise that he/she had chosen one of the alternative routes. From that, we obtained the likelihood that he/she was travelling on that route, given the actual journey time. However, the same type of hypothetical journey time distribution was assumed and used in the mixture model. As a matter of fact, that was not necessary. It would be useful to consider that such hypothetical distribution could have various types of statistical distributions for different individual passengers; and that each individual could have different types of hypothetical journey time distributions for different alternative routes.

One last major issue is that we were short of detailed individual Oyster data. For that reason, in each of the LU O-D cases, the updated posterior probabilities were derived from a very small sample of OJT^{OBS}, which was neither the same as, nor sourced from, the data sample used for the estimation of the mixture model. For future research, a large sample of detailed individual data will be essential.

7.2.2 Route-choice set generation

As mentioned in the previous section, it would be important to consider further developing a choice-set generation model for the modelling framework (*see* **Figure 7.1**). Once a set of alternatives (*i.e.* route-choice set) has been determined, a passenger, for example, would then be expected to choose one alternative (at a time) within this given choice set. In reality, however, any individual passenger's choice set is unable to be observed. Also as mentioned in **Section 3.2**, not all the available alternative routes are necessarily included in each passenger's choice set. Biased estimates of parameters for the attributes in the choice models might be yielded if it is simply supposed that all the possible alternative routes are considered by the passengers.

The generation of the choice set is regarded as a learning process to dynamically adapt passengers' own perceptions on reasonable alternatives (Richardson, 1982). In a dense public transport network, such as the LU, there might be a large number of different possible routes for some O-D pairs. Obviously, it may not be known or observed that which alternative routes are considered by a passenger.

Ben-Akiva and Boccara (1995) discussed the approaches to modelling a latent process of reproducing a choice set. As such, the choice set is probabilistic and influenced by some random factors that are not observable but varying across decision makers. More detailed discussions on this issue can be referred to Cascetta and Papola (2001) and Bierlaire *et al.* (2010).

7.2.3 Refinement and validation of latent route-choice model

To refine the latent route choice model, as mentioned in **Section 6.5**, it would be interesting to re-specify the utility function by further involving line-specific constants, and other significant attributes given the available data. The path size logit, C-Logit, and error components approach could be tested in future research, so as to capture the effect of routes with overlap.

As has also been mentioned in the previous chapter, the validation module that should come into the work is that: we might need to do either simulations or surveys to acquire data where we could know the actual routes that passengers took during their journeys. Then, the choice model shall be estimated via the conventional procedure and we would obtain a set of coefficients, which could be denoted by $\boldsymbol{\beta}^*$. We would need to make a comparison between $\boldsymbol{\beta}^*$ and the estimates from the latent route, denoted by $\boldsymbol{\beta}$. If there would be a big gap between them, then it would be necessary to follow the 'directed golden line', as shown in **Figure 7.1** (*see* p.177), to check the procedures and every aspect of the modelling framework.

Clearly, the mechanism of the current modelling framework is to sequentially deal with the two building blocks (*i.e.* data mining and estimation of the latent choice model). That is, we estimate the posterior probabilities and then the coefficients of the latent route-choice model. Further to this, we foresee the ultimate goal of future study would be to develop a platform where advanced latent choice model is integrated with a simulation-based transit assignment model and both evolve simultaneously to deliver a more robust modelling framework. As such, the proposed modelling framework could be extended to a broader transit network with multiple O-D pairs; and the prospective integrated framework should then contribute to a more realistic representation of the passengers' route choice behaviour as well as a more accurate prediction of the passenger-traffic over the network. This would provide policy makers with much deeper insight into the passengers' travel behaviour and a valuable asset for effective planning of the public transport.

Bibliography

Bagchi, M. and White, P. 2005. The potential of public transport smart card data. *Transport Policy,* **12**(5), pp.464-474.

Bekhor, S., Ben-Akiva, M. E. and Ramming, M. S. 2002. Adaptation of logit kernel to route choice situation. *Transportation Research Record: Journal of the Transportation Research Board,* **1805**, pp.78-85.

Bekhor, S. and Prashker, J. N. 2001. Stochastic user equilibrium formulation for generalized nested logit model. *Transportation Research Record: Journal of the Transportation Research Board,* **1752**, pp.84-90.

Ben-Akiva, M., Bergman, M. J., Daly, A. J. and Ramaswamy, R. 1984. Modelling inter urban route choice behaviour. *In:* Volmuller, J. and Hamerslag, R., eds. *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory*, *11th-13th July 1984*, Delft, The Netherlands. Utrecht, The Netherlands: VNU Science Press, 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; London: Kluwer Academic, pp.5-33.

Ben-Akiva, M. and Boccara, B. 1995. Discrete choice models with latent choice sets. *International Journal of Research in Marketing,* **12**(1), pp.9-24.

Bierlaire, M., Chen, J. and Newman, J. 2013. A probabilistic map matching method for smartphone GPS data. *Transportation Research Part C: Emerging Technologies,* **26**, pp.78-98.

Bierlaire, M. and Frejinger, E. 2008. Route choice modeling with network-free data. *Transportation Research Part C: Emerging Technologies,* **16**(2), pp.187-198.

Bierlaire, M., Frejinger, E. and Stojanovic, J. 2006. A latent route choice model in Switzerland. *In: Proceedings of the European Transport Conference 2006*, *18th-20th September 2006*, Strasbourg, France. Association for European Transport.

Bierlaire, M., Hurtubia, R. and Flötteröd, G. 2010. An analysis of implicit choice set generation using a constrained multinomial Logit model. *Transportation Research Record: Journal of the Transportation Research Board,* **2175**, pp.92-97.

Billi, C., Gentile, G., Nguyen, S. and Pallottino, S. 2004. Rethinking the wait model at transit stops. *In: Proceedings of the Fifth Triennial Symposium on Transportation Analysis (TRISTAN V)*, *13th-18th June 2004*, Le Gosier, Guadeloupe, French West Indies.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. Information science and statistics. New York: Springer-Verlag.

Blömer, J. and Bujna, K. 2013. Simple methods for initializing the EM algorithm for Gaussian mixture models. *arXiv:1312.5946 [cs.LG]* [online]. [Last accessed on 30th September 2014]. Available from: http://arxiv.org/abs/1312.5946.

Bousquet, O., Von Luxburg, U. and Rätsch, G. eds. 2004. *Advanced lectures on machine learning, Machine Learning Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, revised lectures*. Berlin: Springer-Verlag Berlin Heidelberg.

Bouzaïene-Ayari, B., Gendreau, M. and Nguyen, S. 1998. Passenger assignment in congested transit networks: a historical perspective. *In:* Marcotte, P. and Nguyen, S., eds. *Equilibrium and advanced transportation modelling*. Boston, Massachusetts: Kluwer Academic Publishers, pp.47-71.

Bouzaïene-Ayari, B., Gendreau, M. and Nguyen, S. 2001. Modeling bus stops in transit networks: a survey and new formulations. *Transportation Science,* **35**(3), pp.304-321.

Bovy, P. H. L., Bekhor, S. and Prato, C. 2008. The factor of revisited path size: alternative derivation. *Transportation Research Record: Journal of the Transportation Research Board,* **2076**, pp.132-140.

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:* Lesort, J.-B., ed. *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, *24th-26th July 1996*, Lyon, France. Oxford, New York, USA: Pergamon, 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.

Cats, O. 2011. *Dynamic modelling of transit operations and passenger decisions*. Ph.D. thesis, KTH - Royal Institute of Technology.

Cats, O., Koutsopoulos, H., Burghout, W. and Toledo, T. 2011. Effect of real-time transit information on dynamic path choice of passengers. *Transportation Research Record: Journal of the Transportation Research Board,* **2217**, pp.46-54.

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.

Chakirov, A. and Erath, A. 2011. Use of public transport smart card fare payment data for travel behaviour analysis in Singapore. *In:* Szeto, W. Y., Wong, S. C. and Sze, N. N., eds. *Proceedings of the 16th International Conference of Hong Kong Society for Transportation Studies*, *17th-20th December 2011*, Hong Kong, China. Hong Kong Society for Transportation Studies.

Chan, A. H. Y. and Cannon, P. S. 2002. Nonlinear forecasts of *fo*F2: variation of model predictive accuracy over time. *Annales Geophysicae,* **20**(7), pp.1031-1038.

Chan, J. 2007. *Rail transit OD matrix estimation and journey time reliability metrics using automated fare data*. M.Sc. dissertation, Massachusetts Institute of Technology.

Chen, J. 2013. *Modeling route choice behavior using smartphone data*. Ph.D. thesis, École polytechnique fédérale de Lausanne.

Cheung, C.-Y. 1998. *Pedestrian flow characteristics in the Hong Kong mass transit railway stations*. M.Phil. dissertation, The Hong Kong Polytechnic University.

Chriqui, C. and Robillard, P. 1975. Common bus lines. *Transportation Science,* **9**(2), pp.115-121.

Cominetti, R. and Correa, J. 2001. Common-lines and passenger assignment in congested transit networks. *Transportation Science,* **35**(3), pp.250-267.

Daganzo, C. F. and Sheffi, Y. 1977. On stochastic models of traffic assignment. *Transportation Science,* **11**(3), pp.253-274.

Daly, P. N., Mcgrath, F. and Annesley, T. J. 1991. Pedestrian speed/flow relationships for underground stations. *Traffic Engineering & Control,* **32**(2), pp.75-78.

Davis, P. and Dutta, G. 2002. *Estimation of capacity of escalators in London Underground*. London: London School of Economics and Political Sciences.

De Cea, J. and Fernández, E. 1993. Transit assignment for congested public transport systems: An equilibrium model. *Transportation Science,* **27**(2), pp.133-147.

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.

Department for Transport. 2014. *Public transport assignment.* Transport Analysis Guidance (TAG) Unit M3.2.

Dial, R. B. 1967. Transit pathfinder algorithm. *Highway Research Record,* **205**(67), pp.67-85.

Dziak, J. J., Coffman, D. L., Lanza, S. T. and Li, R. 2012. *Sensitivity and specificity of information criteria*. (Technical Report Series #12-119). University Park, PA: College of Health and Human Development, The Pennsylvania State University.

Farmer, J. D. and Sidorowich, J. J. 1987. Predicting chaotic time series. *Physical Review Letters,* **59**(8), pp.845-848.

Fearnside, K. and Draper, D. P. 1971. Public transport assignment - a new approach. *Traffic Engineering and Control,* **13**(7), pp.298-299.

Florian, M. 1999. Deterministic time table transit assignment. *In: Proceedings of the First Asian EMME/2 Users Group Meeting*, *23th-24th August 1999*, Shanghai, China.

Forgy, E. W. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics,* **21**, pp.768-769.

Freeman, J. V., Walters, S. J. and Campbell, M. J. 2008. *How to display data*. Malden, Massachusetts; Oxford: Blackwell.

Frejinger, E. and Bierlaire, M. 2007. Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological,* **41**(3), pp.363-378.

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 series in statistics. New York: Springer-Verlag.

Fu, Q. 2012a. Bayesian inference of passengers' path choices with incomplete data - an application for London Underground. *In:* Mak, H.-Y. and Lo, H. K., eds. *Proceedings of the 17th International Conference of Hong Kong Society for Transportation Studies*, *15th-17th December 2012*, Hong Kong, China. Hong Kong Society for Transportation Studies, pp.39-46.

Fu, Q. 2012b. Understanding the travel patterns on London rail network: the use of Oyster card data. *In: Proceedings of the 44th Annual Conference of Universities' Transport Study Group*, *4th-6th January 2012*, Aberdeen, UK.

Fu, Q. 2014. A Bayesian modelling framework for individual passenger's probabilistic route choices: a case study on the London Underground. *In: Proceedings of the 46th Annual Conference of Universities' Transport Study Group*, *6th-8th January 2014*, Newcastle upon Tyne, UK.

Fu, Q., Liu, R. and Hess, S. 2012a. On considering journey time variability and passengers' path choice: an empirical study using Oyster card data on London Underground. *In:* Lam, W. H. K., Lo, H. K. and Wong, S. C., eds. *Proceedings of the Fifth International Symposium on Transportation Network Reliability*, *18th-19th December 2012*, Hong Kong, China. pp.619-633.

Fu, Q., Liu, R. and Hess, S. 2012b. A review on transit assignment modelling approaches to congested networks: a new perspective. *In:* Aguiléra, V., Bhouri, N., Farhi, N., Leurent, F. and Seidowsky, R., eds. *Proceedings of the 15th Meeting of the EURO Working Group on Transportation*, *10th-13th September 2012*, Paris, France. Procedia - Social and Behavioral Sciences, Vol.54, pp.1145-1155.

Fu, Q., Liu, R. and Hess, S. 2014. A Bayesian modelling framework for individual passenger's probabilistic route choices: a case study on the London Underground. *In: Proceedings of the Transportation Research Board 93rd Annual Meeting*, *12th-16th January 2014*, Washington D.C., USA. Transportation Research Board of the National Academies.

Gentile, G., Nguyen, S. and Pallottino, S. 2005. Route choice on transit networks with online information at stops. *Transportation Science,* **39**(3), pp.289-297.

Guo, Z. 2008. *Transfers and path choice in urban public transport systems*. Ph.D. 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.

Guo, Z. and Wilson, N. H. M. 2011. Assessing the cost of transfer inconvenience in public transport systems: a case study of the London Underground. *Transportation Research Part A: Policy and Practice,* **45**(2), pp.91-104.

Guyon, I. and Elisseeff, A. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research,* **3**, pp.1157-1182.

Guyon, I., Saffari, A., Dror, G. and Cawley, G. 2010. Model selection: beyond the bayesian/frequentist divide. *The Journal of Machine Learning Research,* **11**, pp.61-87.

Hamdouch, Y., Ho, H. W., Sumalee, A. and Wang, G. 2011. Schedule-based transit assignment model with vehicle capacity and seat availability. *Transportation Research Part B: Methodological,* **45**(10), pp.1805-1830.

Hamdouch, Y. and Lawphongpanich, S. 2008. Schedule-based transit assignment model with travel strategies and capacity constraints. *Transportation Research Part B: Methodological,* **42**(7-8), pp.663-684.

Hamdouch, Y., Marcotte, P. and Nguyen, S. 2004. Capacitated transit assignment with loading priorities. *Mathematical Programming,* **101**(1), pp.205-230.

Hankin, B. D. and Wright, R. A. 1958. Passenger flow in subways. *OR,* **9**(2), pp.81-88.

Harris, N. G. 1991. Modelling walk link congestion and the prioritisation of congestion relief. *Traffic engineering & control,* **32**(2), pp.78-80.

Heckerman, D. 1997. Bayesian networks for data mining. *Data Mining and Knowledge Discovery,* **1**(1), pp.79-119.

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.

Hickman, M. D. and Wilson, N. H. M. 1995. Passenger travel time and path choice implications of real-time transit information. *Transportation Research Part C: Emerging Technologies,* **3**(4), pp.211-226.

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An introduction to statistical learning: with applications in R*. Springer texts in statistics. New York: Springer-Verlag.

Jang, W. 2010. Travel time and transfer analysis using transit smart card data. *Transportation Research Record: Journal of the Transportation Research Board,* **2144**, pp.142-149.

Johnson, R. A. and Bhattacharyya, G. K. 2009. *Statistics: principles and methods. (6th ed.)*. Hoboken, New Jersey: Wiley.

Karlis, D. and Xekalaki, E. 2003. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis,* **41**(3–4), pp.577-590.

Kleinrock, L. 1975. *Queueing systems - Volume 1: theory*. New York: Wiley.

Kuha, J. 2004. AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research,* **33**(2), pp.188-229.

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.

Kurauchi, F., Schmöcker, J.-D., Shimamoto, H. and Hassan, S. M. 2012. Empirical analysis on passengers' hyperpath construction by smart card data. *In: Proceedings of the 12th Conference on Advanced Systems for Public Transport, 23rd-27th July, 2012*, Santiago, Chile.

Lam, W. H. K. and Cheung, C.-Y. 2000. Pedestrian speed/flow relationships for walking facilities in Hong Kong. *Journal of Transportation Engineering,* **126**(4), pp.343-349.

Lam, W. H. K., Gao, Z. Y., Chan, K. S. and Yang, H. 1999. A stochastic user equilibrium assignment model for congested transit networks. *Transportation Research Part B: Methodological,* **33**(5), pp.351-368.

Lam, W. H. K., Zhou, J. and Sheng, Z.-H. 2002. A capacity restraint transit assignment with elastic line frequency. *Transportation Research Part B: Methodological,* **36**(10), pp.919-938.

Laplace, P.-S. 1995. *Philosophical essay on probabilities; translated from the fifth French edition of 1825, with notes by the translator, Andrew I. Dale.* Sources in the history of mathematics and physical sciences. New York: Springer-Verlag.

Le Clercq, F. 1972. A public transport assignment method. *Traffic Engineering and Control,* **14**(2), pp.91-96.

Lee, D.-H., Sun, L. and Erath, A. 2012. Study of bus service reliability in Singapore using fare card data. *In: Proceedings of the 12th Asia-Pacific ITS Forum, 16th-18th April 2012*, Kuala Lumpur, Malaysia.

Leonard, T., Hsu, J. S. J. and Tsui, K.-W. 1989. Bayesian marginal inference. *Journal of the American Statistical Association,* **84**(408), pp.1051-1058.

Leurent, F. 2010. On seat capacity in traffic assignment to a transit network. *Journal of Advanced Transportation,* **46**(2), pp.112-138.

Liu, Y., Bunker, J. and Ferreira, L. 2010. Transit users' route-choice modelling in transit assignment: a review. *Transport Reviews,* **30**(6), pp.753-769.

London Regional Transport. 1988. *Research into interchange penalties on the London Underground*. Operational Research Note 88/24, London, United Kingdom.

London Regional Transport. 1995. *Interchange penalty*. Operational Research Report OR95/5, London, UK.

Lowd, D. and Domingos, P. 2005. Naive Bayes models for probability estimation. *In:* De Raedt, L. and Wrobel, S., eds. *Proceedings of the 22nd International Conference on Machine learning*, *7th-11th August 2005*, Bonn, Germany. New York: ACM, pp.529-536.

Macqueen, J. 1967. Some methods for classification and analysis of multivariate observations. *In:* Le Cam, L. M. and Neyman, J., eds. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *21st June – 18th July 1965, and 27th December 1965 – 7th January 1966*, Statistical Laboratory of the University of California, Berkeley. California, USA, Vol.1, pp.281-297.

Marron, J. S. and Wand, M. P. 1992. Exact mean integrated squared error. *The Annals of Statistics,* **20**(2), pp.712-736.

Mcfadden, D. 2000. Disaggregate behavioral travel demand's RUM side - a 30-year retrospective. *Travel Behaviour Research*, pp.17-63.

Mclachlan, G. J. 1988. On the choice of starting values for the EM algorithm in fitting mixture models. *Journal of the Royal Statistical Society. Series D (The Statistician),* **37**(4/5), pp.417-425.

Mclachlan, G. J. and Peel, D. 2000. *Finite mixture models*. Wiley series in probability and statistics. Applied probability and statistics. New York; Chichester: Wiley.

Melnykov, V. and Melnykov, I. 2012. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis,* **56**(6), pp.1381-1395.

Mood, A. M., Graybill, F. A. and Boes, D. C. 1974. *Introduction to the theory of statistics*. McGraw-Hill series in probability and statistics. *(3rd ed.)*. New York; London: McGraw-Hill.

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.

Nagele, P. 2003. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *British Journal of Anaesthesia,* **90**(4), pp.514-516.

Nguyen, S. and Pallottino, S. 1988. Equilibrium traffic assignment for large scale transit networks. *European Journal of Operational Research,* **37**(2), pp.176-186.

Nguyen, S., Pallottino, S. and Malucelli, F. 2001. A modeling framework for the passenger assignment on a transport network with timetables. *Transportation Science,* **35**(3), pp.238-249.

Nielsen, O. A. 2000. A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research Part B: Methodological,* **34**(5), pp.377-402.

Nökel, K. and Wekeck, S. 2009. Boarding and alighting in frequency-based transit assignment. *Transportation Research Record: Journal of the Transportation Research Board,* **2111**(1), pp.60-67.

Nuzzolo, A. and Crisalli, U. 2004. The schedule-based approach in dynamic transit modeling: a general overview. *In:* Wilson, N. H. M. and Nuzzolo, A., eds. *Schedule-based dynamic transit modeling: theory and applications (1st ed.)*. Dordrecht; London: Kluwer Academic, pp.1-24.

Nuzzolo, A. and Crisalli, U. 2009. The schedule-based modeling of transportation systems: recent developments. *In:* Wilson, N. H. M. and Nuzzolo, A., eds. *Schedule-based modeling of transportation networks: theory and applications*. New York: Springer, pp.1-26.

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.

Nuzzolo, A., Russo, F. and Crisalli, U. 2003. *Transit network modelling: the schedule-based dynamic approach*. Collana trasporti. Milano: Franco Angeli.

Ortúzar, J. D. D. and Willumsen, L. G. 2011. *Modelling transport*. *(4th ed.)*. Chichester, West Sussex, UK: John Wiley & Sons.

Park, J. Y., Kim, D.-J. and Lim, Y. 2008. Use of smart card data to define public transit use in Seoul, South Korea. *Transportation Research Record: Journal of the Transportation Research Board,* **2063**, pp.3-9.

Parkin, T. B., Chester, S. T. and Robinson, J. A. 1990. Calculating confidence intervals for the mean of a lognormally distributed variable. *Soil Science Society of America Journal,* **54**(2), pp.321-326.

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.

Poon, M. H., Wong, S. C. and Tong, C. O. 2004. A dynamic schedule-based model for congested transit networks. *Transportation Research Part B: Methodological,* **38**(4), pp.343-368.

Prashker, J. N. and Bekhor, S. 1998. Investigation of stochastic network loading procedures. *Transportation Research Record: Journal of the Transportation Research Board,* **1645**, pp.94-102.

Prato, C. G. 2009. Route choice modeling: past, present and future research directions. *Journal of Choice Modelling,* **2**(1), pp.65-100.

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.

Redner, R. A. and Walker, H. F. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review,* **26**(2), pp.195-239.

Richardson, A. 1982. Search models and choice set generation. *Transportation Research Part A: General,* **16**(5-6), pp.403-419.

Riley, T. and Goucher, A. eds. 2009. *Beautiful testing: leading professionals reveal how they improve software*. Sebastopol, California: O'Reilly.

Russell, S. J. and Norvig, P. 2010. *Artificial intelligence: a modern approach*. Prentice Hall series in artificial intelligence. *(3rd ed.)*. Upper Saddle River, New Jersey: Prentice Hall.

Schmöcker, J.-D. 2006. *Dynamic capacity constrained transit assignment*. Ph.D. thesis, Imperial College London.

Schmöcker, J.-D. and Bell, M. G. H. 2009. The build-up of capacity problems during the peak hour. *In:* Wilson, N. H. M. and Nuzzolo, A., eds. *Schedule-based modeling of transportation networks: theory and applications*. New York: Springer, pp.217-239.

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.

Schmöcker, J.-D., Shimamoto, H. and Kurauchi, F. 2013. Generation and calibration of transit hyperpaths. *Transportation Research Part C: Emerging Technologies,* **36**, pp.406-418.

Seidel, W., Mosler, K. and Alker, M. 2000. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics,* **52**(3), pp.481-487.

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.

Sumalee, A., Tan, Z. and Lam, W. H. K. 2009. Dynamic stochastic transit assignment with explicit seat allocation model. *Transportation Research Part B: Methodological,* **43**(8-9), pp.895-912.

Szeto, W. Y., Jiang, Y., Wong, K. I. and Solayappan, M. 2013. Reliability-based stochastic transit assignment with capacity constraints: formulation and solution method. *Transportation Research Part C: Emerging Technologies,* **35**, pp.286–304.

Szeto, W. Y., Solayappan, M. and Jiang, Y. 2011. Reliability-based transit assignment for congested stochastic transit networks. *Computer-Aided Civil and Infrastructure Engineering,* **26**(4), pp.311-326.

Teklu, F. 2008a. *A Markov process model for capacity-constrained transit assignment*. Ph.D. thesis, University of Leeds.

Teklu, F. 2008b. A stochastic process approach for frequency-based transit assignment with strict capacity constraints. *Networks and Spatial Economics,* **8**(2), pp.225-240.

Teklu, F., Watling, D. P. and Connors, R. D. 2007. A Markov process model for capacity constrained transit assignment. *In:* Allsop, R. E., Bell, M. G. H. and Heydecker, B. G., eds. *Transportation and traffic theory 2007: papers selected for presentation at ISTTT17, a peer reviewed series since 1959*. Amsterdam; Oxford: Elsevier, pp.483-505.

The Office of Rail Regulation. 2014. *Underground railways* [online]. [Last accessed on 30th September 2014]. Available from: http://orr.gov.uk/about-orr/who-we-work-with/rail-infrastructure/underground-railways.

Tian, Q., Huang, H. and Yang, H. 2007. Commuting Equilibria on a mass transit system with capacity constraints. *In:* Allsop, R. E., Bell, M. G. H. and Heydecker, B. G., eds. *Transportation and traffic theory 2007: papers selected for presentation at ISTTT17, a peer reviewed series since 1959*. Amsterdam; Oxford: Elsevier, pp.360-383.

Tong, C. O. 1986. *A schedule-based transit network model*. Ph.D. thesis, Monash University.

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.

Train, K. E. 2009. *Discrete choice methods with simulation*. *(2nd ed.)*. Cambridge; New York: Cambridge University Press.

Transport for London. 2010. *Travel in London, report 3*. London, UK: Transport for London. [Online]. [Last accessed on 30th September 2014]. Available from: http://www.tfl.gov.uk/corporate/publications-and-reports/travel-in-london-reports.

Transport for London. 2012. *Transport for London factsheet*. London, UK: Transport for London. [Online]. [Last accessed on 9th August 2012]. Available from: http://www.tfl.gov.uk/corporate/publications-and-reports/factsheets.

Transport for London. 2013a. *Business Case Development Manual.* London, UK: TfL Programme Management Office.

Transport for London. 2013b. *Large print Tube map*. London, UK: Transport for London. [Online]. [Last accessed on 22nd July 2013]. Available from: https://www.tfl.gov.uk/maps/track/tube.

Transport for London. 2013c. *Tube map*. London, UK: Transport for London. [Online]. [Last accessed on 13th May 2013]. Available from: https://www.tfl.gov.uk/maps/track/tube.

Trépanier, M., Morency, C. and Agard, B. 2009. Calculation of transit performance measures using smartcard data. *Journal of Public Transportation,* **12**(1), pp.76–96.

Uniman, D. L. 2009. *Service reliability measurement framework using smart card data: application to the London Underground*. M.Sc. dissertation, Massachusetts Institute of Technology.

Uniman, D. L., Attanucci, J., Mishalani, R. and Wilson, N. H. M. 2010. Service reliability measurement using automated fare card data. *Transportation Research Record: Journal of the Transportation Research Board,* **2143**, pp.92-99.

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.

Walck, C. 1996. *Hand-book on statistical distributions for experimentalists*. (Internal Report SUF–PFY/96–01). Stockholm: Particle Physics Group, Fysikum, University of Stockholm, 11 December 1996.

Wardman, M., Hine, J. and Stradling, S. 2001a. *Interchange and travel choice: volume 1 - Report for the Scottish Executive by the Institute for Transport Studies at the University of Leeds and the Transport Research Institute at Napier University*. UK.

Wardman, M., Hine, J. and Stradling, S. 2001b. *Interchange and travel choice: volume 2 - Report for the Scottish Executive by the Institute for Transport Studies at the University of Leeds and the Transport Research Institute at Napier University*. United Kingdom.

White, P., Bagchi, M. and Bataille, H. 2010. The role of smartcard data in public transport. *In: Proceedings of the 12th World Conference on Transport Research, 11th-15th July 2010*, Lisbon, Portugal.

Wikipedia. 2014. *List of metro systems* [online]. [Last accessed on 30th September 2014]. Available from:
https://en.wikipedia.org/wiki/List_of_metro_systems.

Wilson, N. H. M., Zhao, J. and Rahbee, A. 2009. The potential impact of automated data collection systems on urban public transport planning. *In:* Wilson, N. H. M. and Nuzzolo, A., eds. *Schedule-based modeling of transportation networks: theory and applications*. New York: Springer, pp.1-25.

Wu, J. H., Florian, M. and Marcotte, P. 1994. Transit equilibrium assignment: a model and solution algorithms. *Transportation Science,* **28**(3), pp.193-203.

Yin, Y., Lam, W. H. K. and Miller, M. A. 2004. A simulation-based reliability assessment approach for congested transit network. *Journal of Advanced Transportation,* **38**(1), pp.27-44.

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.

Zwillinger, D. and Kokoska, S. 1999. *CRC standard probability and statistics tables and formulae*. Boca Raton: Chapman & Hall/CRC Press.

Appendix A
Explanatory notes on logical dependency and relationship between route choice and journey time

Suppose there are two alternative routes (hereinafter referred to as Route1 and Route2) connecting a given pair of **o**rigin and a **d**estination (O-D) stations. We let z denote the travel demand (*i.e.* total number of passengers) on this O-D; and

$$z = z_1 + z_2 , \tag{A-1}$$

where z_1 represents the total number of passengers who choose Route1, and z_2 represents that on Route2.

For any individual passenger, the probability of choosing Route1 is

$$\Pr(choice_1) = \frac{z_1}{z} ; \tag{A-2}$$

and that the probability that an individual chooses Route2 is

$$\Pr(choice_1) = \frac{z_2}{z} = 1 - \frac{z_1}{z} . \tag{A-3}$$

Let us adapt an example from the Wikipedia[1] to imitate the case of passengers' route choices. Suppose that $A\%$ of the passenger population chose Route1, and $B\%$ chose Route2. Let δ_1^* and δ_2^* be the average journey times of travelling by Route1 and Route2, respectively. Further, we can observe each passenger's journey time, but without knowing his/her route choice.

If we have known that a passenger's journey time is δ^* minutes, $X\%$ of passengers on Route1 spent δ^* minutes, and $Y\%$ of passengers on Route2 spent δ^* minutes as well, but can we infer the probability that the passenger chose Route1 (or Route2)?

According to equations (A-1) to (A-3), we may have (*see next page*)

[1] Available online at http://en.wikipedia.org/wiki/Posterior_probability; last accessed on 30 September 2014.

$$\Pr(choice_1) = A\%\,;$$

$$\Pr(choice_2) = B\%\,;$$

$$\Pr(\delta^* \mid choice_1) = X\%\,;$$

and

$$\Pr(\delta^* \mid choice_2) = Y\%\,.$$

The total number of passengers who chose Route1 and took δ^* minutes to complete his/her journey is $z \cdot \Pr(choice_1) \cdot \Pr(\delta^* \mid choice_1)$; and the amount of passengers who chose Route2 and took δ^* minutes to complete his/her journey is $z \cdot \Pr(choice_2) \cdot \Pr(\delta^* \mid choice_2)$. Therefore, the probability that the passenger chose Route1 is calculated as

$$\Pr(choice_1 \mid \delta^*) = \frac{z \cdot \Pr(choice_1) \cdot \Pr(\delta^* \mid choice_1)}{z \cdot \Pr(choice_1) \cdot \Pr(\delta^* \mid choice_1) + z \cdot \Pr(choice_2) \cdot \Pr(\delta^* \mid choice_2)}\,;$$

and so

$$\Pr(choice_1 \mid \delta^*) = \frac{\Pr(choice_1) \cdot \Pr(\delta^* \mid choice_1)}{\Pr(choice_1) \cdot \Pr(\delta^* \mid choice_1) + \Pr(choice_2) \cdot \Pr(\delta^* \mid choice_2)}\,. \quad \text{(A-4)}$$

That is,

$$\Pr(choice_1 \mid \delta^*) = \frac{A\% \cdot X\%}{A\% \cdot X\% + B\% \cdot Y\%}\,.$$

Likewise,

$$\Pr(choice_2 \mid \delta^*) = \frac{B\% \cdot X\%}{A\% \cdot X\% + B\% \cdot Y\%}\,.$$

In equation (A-4), the denominator is in fact the proportion of passengers who spent δ^* minutes in travelling between the O-D, which we represent by $\Pr(\delta^*)$. The numerator is equivalent to $\Pr(choice_1, \delta^*)$. The conditional probability of route choice (say, the choice of Route-j), $\Pr(choice_j, \delta^*)$, is the probability of a passenger choosing Route-j given that his/her journey time δ^* has been already observed. We have

$$\Pr(choice_1 \mid \delta^*) = \frac{\Pr(choice_1, \delta^*)}{\Pr(\delta^*)}\,, \quad \text{(A-5)}$$

or

$$\Pr(choice_1, \delta^*) = \Pr(choice_1 \mid \delta^*) \cdot \Pr(\delta^*)\,, \tag{A-6}$$

or

$$\Pr(choice_1 \mid \delta^*) = \frac{\Pr(choice_1) \cdot \Pr(\delta^* \mid choice_1)}{\Pr(\delta^*)}\,. \tag{A-7}$$

If the journey time δ^* has no correlation with the route choice, then

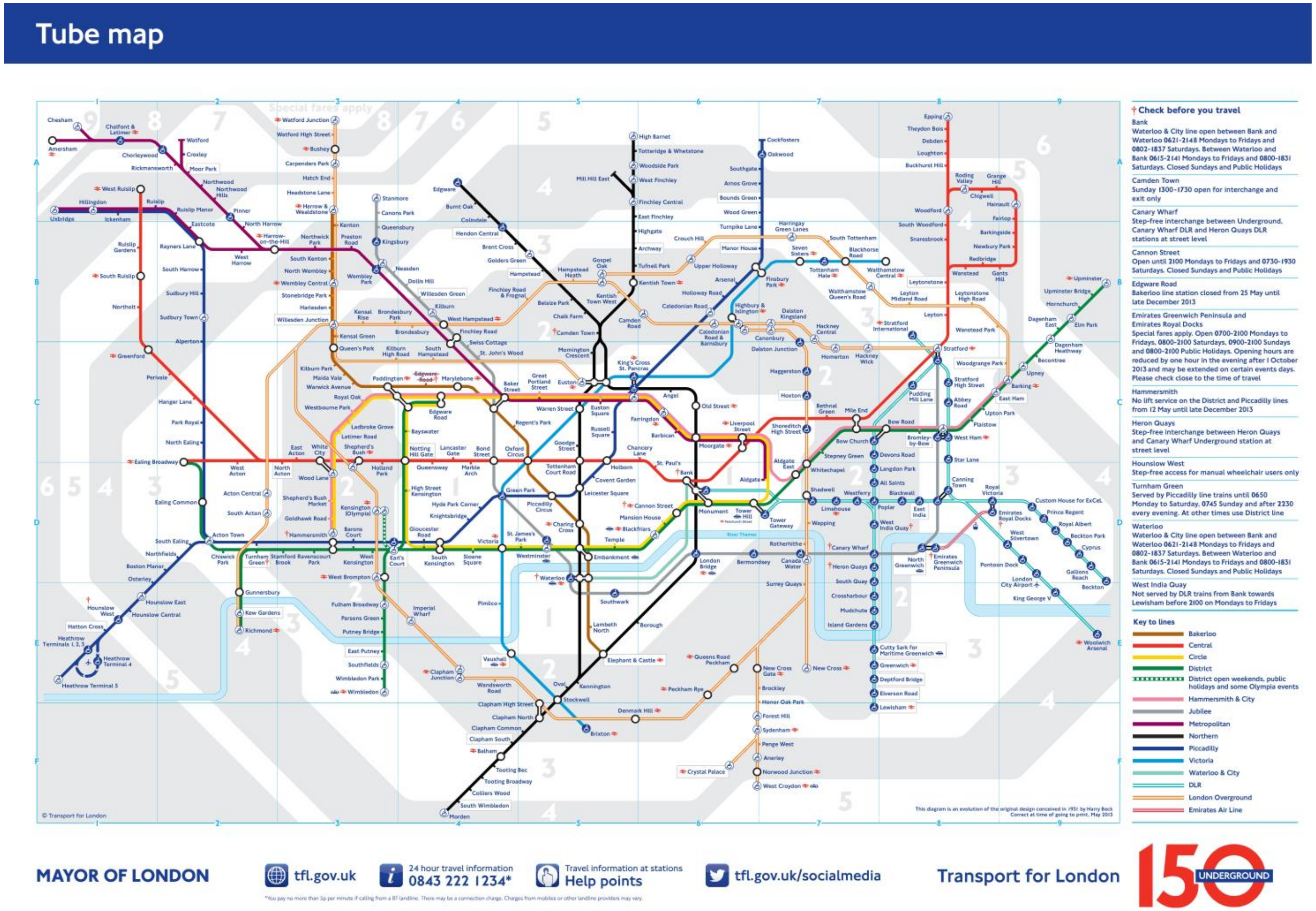
$$\Pr(choice_j \mid \delta^*) = \Pr(choice_j)\,.$$

Intuitively, we believe that, for an individual passenger, he/she may have different journey times when travelling by different routes. Note that $choice_j$ and δ^* are not independent of each other. The probability of both events occurring at the same time is defined by equation (A-6) and the conditional probability is obtained by equation (A-5).

Given data of journey time observations, $\Pr(\delta^*)$ is certainly greater than 0; otherwise, journey time δ^* is not observed and $\Pr(\delta^*) = 0$.

Appendix B

Standard Tube map (© Transport for London)



Tube map

† Check before you travel

Bank
Waterloo & City line open between Bank and Waterloo 0621-2148 Mondays to Fridays and 0802-1837 Saturdays. Between Waterloo and Bank 0615-2141 Mondays to Fridays and 0800-1831 Saturdays. Closed Sundays and Public Holidays

Camden Town
Sunday 1300-1730 open for interchange and exit only

Canary Wharf
Step-free interchange between Underground, Canary Wharf DLR and Heron Quays DLR stations at street level

Cannon Street
Open until 2100 Mondays to Fridays and 0730-1930 Saturdays. Closed Sundays and Public Holidays

Edgware Road
Bakerloo line station closed from 25 May until late December 2013

Emirates Greenwich Peninsula and Emirates Royal Docks
Special fares apply. Open 0700-2100 Mondays to Fridays, 0800-2100 Saturdays, 0900-2100 Sundays and 0800-2100 Public Holidays. Opening hours are reduced by one hour in the evening after 1 October 2013 and may be extended on certain events days. Please check close to the time of travel

Hammersmith
No lift service on the District and Piccadilly lines from 12 May until late December 2013

Heron Quays
Step-free interchange between Heron Quays and Canary Wharf Underground station at street level

Hounslow West
Step-free access for manual wheelchair users only

Turnham Green
Served by Piccadilly line trains until 0650 Monday to Saturday, 0745 Sunday and after 2230 every evening. At other times use District line

Waterloo
Waterloo & City line open between Bank and Waterloo 0621-2148 Mondays to Fridays and 0802-1837 Saturdays. Between Waterloo and Bank 0615-2141 Mondays to Fridays and 0800-1831 Saturdays. Closed Sundays and Public Holidays

West India Quay
Not served by DLR trains from Bank towards Lewisham before 2100 on Mondays to Fridays

Key to lines

- Bakerloo
- Central
- Circle
- District
- District open weekends, public holidays and some Olympia events
- Hammersmith & City
- Jubilee
- Metropolitan
- Northern
- Piccadilly
- Victoria
- Waterloo & City
- DLR
- London Overground
- Emirates Air Line

© Transport for London

This diagram is an evolution of the original design conceived in 1931 by Harry Beck
Correct at time of going to print, May 2013

MAYOR OF LONDON

tfl.gov.uk

24 hour travel information 0843 222 1234*

Travel information at stations Help points

tfl.gov.uk/socialmedia

Transport for London

150 UNDERGROUND

*You pay no more than 5p per minute if calling from a BT landline. There may be a connection charge. Charges from mobiles or other landline providers may vary.

Appendix C
Application of mixture models: *Case-5* and *Case-7*

In addition to the five case studies described in **Section 4.3**, two more cases are showcased in this appendix, which exhibits only the estimation results obtained from GM and LNM models.

Further to ***Case-4*** (*see* **Section 4.3.2.1**), **Appendix C.1** shows a case study of a pair of O-D stations that are connected by three alternative routes. We code-name this case study '***Case-5***'. And in addition to ***Case-6*** (*see* **Section 4.3.2.2**), the results of another case of four alternative routes, code-named '***Case-7***', are presented in **Appendix C.2**.

C.1 *Case-5*: Liverpool Street – Green Park

This section shows a case study of the mixture models applied on another O-D pair with three alternative routes. Its network is illustrated in **Figure C.1** (*see next page*).

Any passenger starting his/her journey at the origin, **Liverpool Street** station, may take a westbound train on the Central line (as the only option) for the first leg of his/her journey. In order to reach the destination, *i.e.* **Green Park** station, alternative interchange stations include **Holborn** (transferring to a westbound train on the Piccadilly line, for the shortest first journey leg among all of the three alternatives); **Oxford Circus** (transferring to a southbound train on the Victoria line train; and **Bond Street** (transferring to a southbound train on the Jubilee line, for the longest first journey leg).

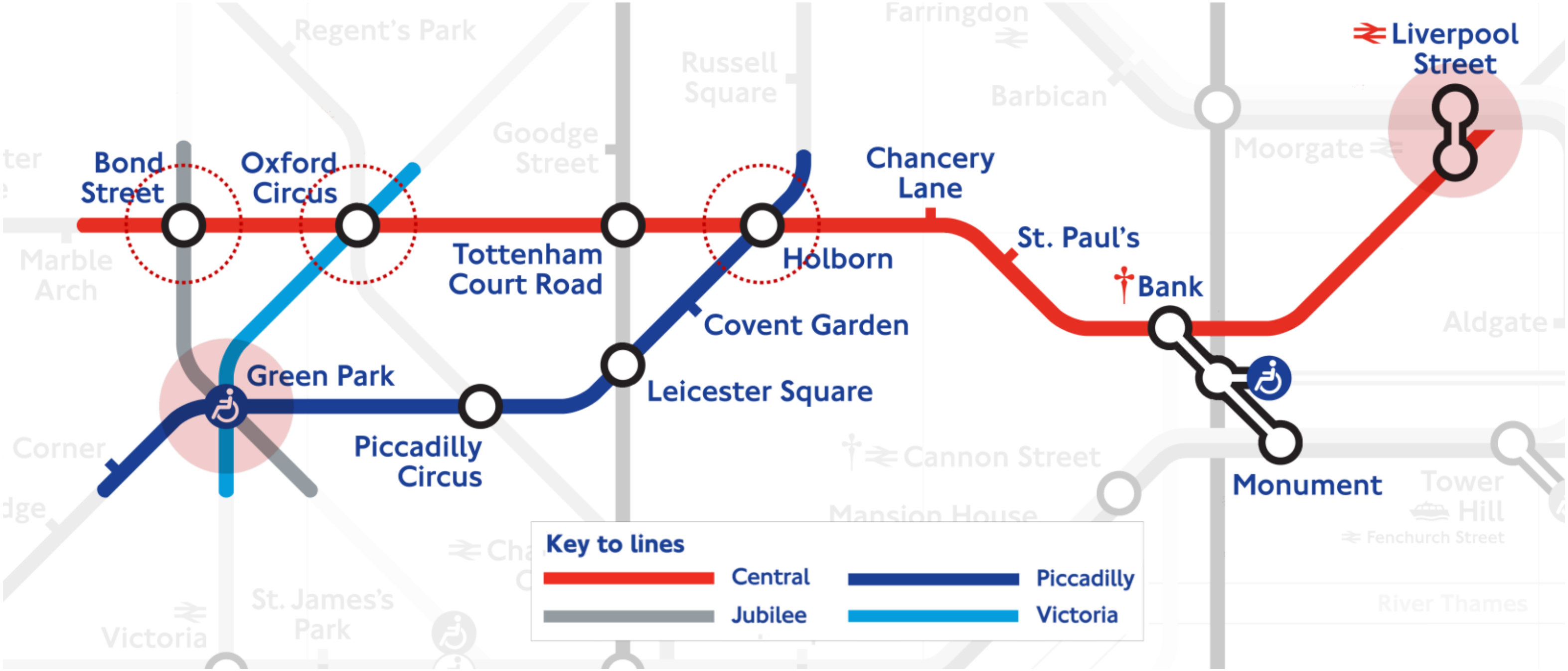


Figure C.1 The LU network that connects the O-D pair: Liverpool Street – Green Park.

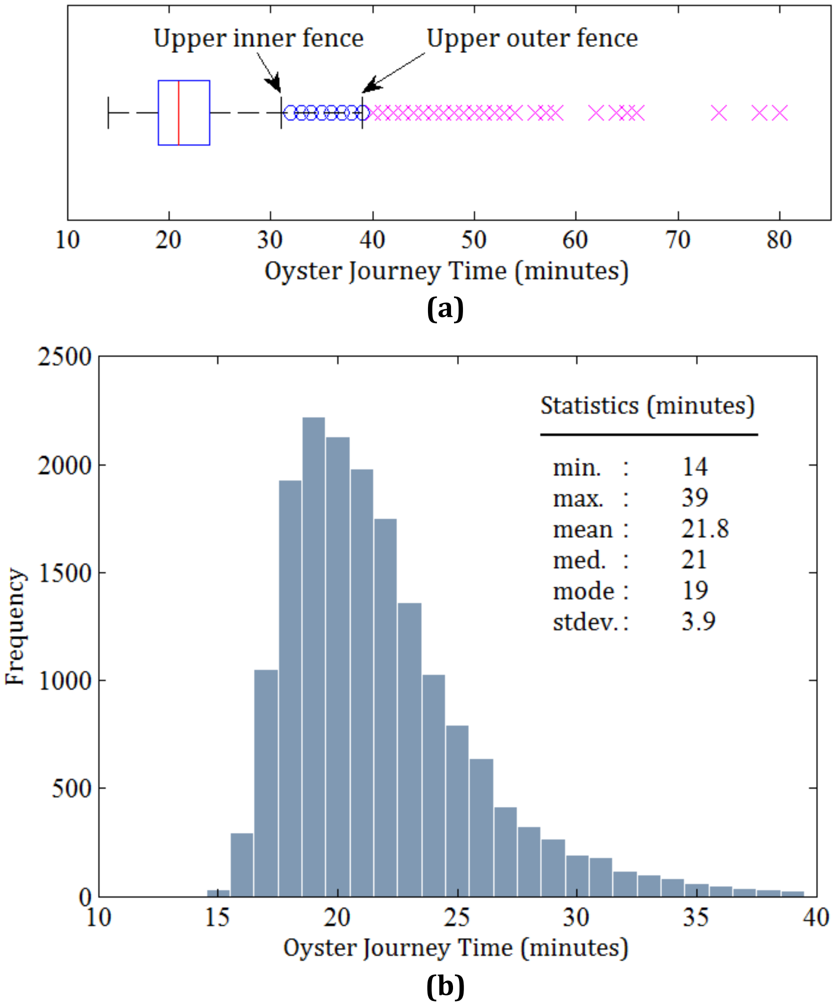


Figure C.2 Summary of OJT^{OBS} data for **Liverpool Street – Green Park**:

(a) a box-and-whisker plot of the raw data $(n_0 = 17{,}423)$; and
(b) a histogram of the valid data $(n = 17{,}102)$.

Table C.1 Parameter estimates of GM and LNM models based on OJT^{OBS} data for **Liverpool Street – Green Park**

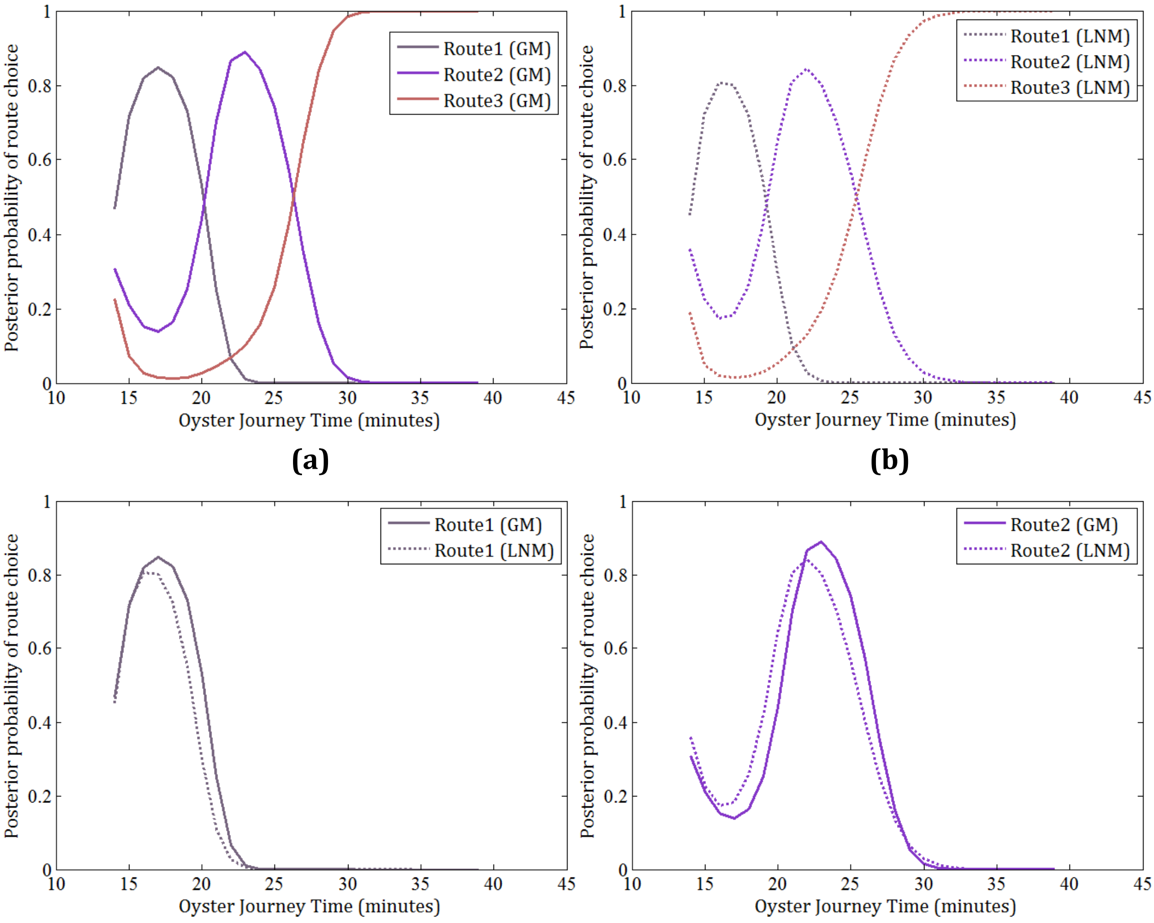
The initial values and the model parameters were estimated using the *K*-means clustering and the EM algorithm, respectively. $n = 17{,}102$.

	GM			LNM		
Component-label	$r=1$	$r=2$	$r=3$	$r=1$	$r=2$	$r=3$
Initial values						
η_r^{KMS} (minute)	19.0	22.0	27.0	18.0	21.0	26.2
σ_r^{KMS} (minute)	1.2	1.1	3.3	0.9	1.1	3.0
ω_r^{KMS} (%)	44.8	35.8	19.4	32.3	42.2	25.4

(*Continued*)

Table C.1 (*Continued.*)

		GM			LNM		
Component-label		$r=1$	$r=2$	$r=3$	$r=1$	$r=2$	$r=3$
Parameter estimates							
$\hat{\mu}_r$	(minute)	18.7	22.0	27.6	18.4	21.5	26.6
$\hat{\sigma}_r$	(minute)	1.4	2.3	4.5	1.3	2.3	4.4
$\hat{\omega}_r$	(%)	35.9	47.7	16.4	27.1	51.6	21.3



(a)

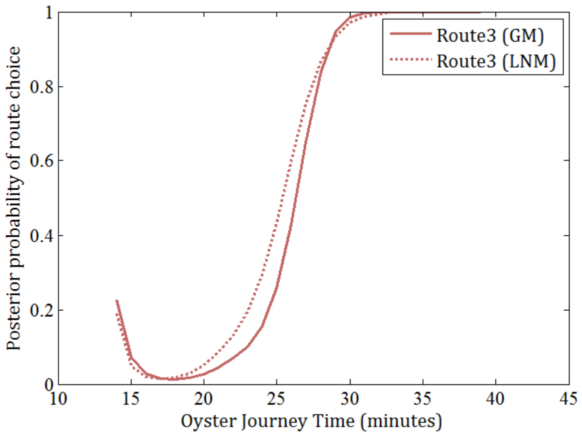
(b)

(c)

(d)

Figure C.3 Posterior probabilities of route choices given OJT^{OBS} for **Liverpool Street – Green Park** $(n=17{,}102)$:

(a) for all alternatives, based on GM; **(b)** for all alternatives, based on LNM; **(c)** for Route1, based on GM and LNM; **(d)** for Route2, based on GM and LNM; and **(e)** for Route3, based on GM and LNM (*see next page*).



(e)

Figure C.3 (*Continued.*)

Table C.2 Inferences of proportion of passenger traffic on each alternative route connecting **Liverpool Street** to **Green Park** ($n = 17{,}102$)

	GM			LNM		
Component-label	$r=1$	$r=2$	$r=3$	$r=1$	$r=2$	$r=3$
$\hat{\omega}_r$ (%)	35.9	47.7	16.4	27.1	51.6	21.3
$n_r^{\mathrm{INF}_0}$	7,660	7,553	1,889	5,528	9,047	2,527
$\omega_r^{\mathrm{INF}_0}$ (%)	44.8	44.2	11.0	32.3	52.9	14.8
n_r^{INF}	6,060	8,257	2,785	4,672	8,801	3,629
ω_r^{INF} (%)	35.4	48.3	16.3	27.3	51.5	21.2

Table C.3 Expected journey times of simulated samples for each alternative route connecting **Liverpool Street** to **Green Park**

	Calculated average travel time (minutes)		
l' – l'' s	Central – Victoria Oxford Circus	Central – Piccadilly Holborn	Central – Jubilee Bond Street
Journey segment			
$t^{\mathrm{ACC}}_{l',o}$	2.1	2.1	2.1
$t^{\mathrm{WTD}}_{l',o,1}$ / $t^{\mathrm{WTD}}_{l',o,2}$	1.5 / 3.7	1.5 / 3.7	1.5 / 3.7
$t^{\mathrm{OBT}}_{l',[o,s]}$	10.0	7.0	11.0
$t^{\mathrm{ICT}}_{[l',l''],s}$	2.0	3.4	3.1
$t^{\mathrm{ICW}}_{l'',s,1}$ / $t^{\mathrm{ICW}}_{l'',s,2}$	0.7 / 2.7	1.4 / 3.8	1.4 / 3.6
$t^{\mathrm{OBT}}_{l'',[s,d]}$	1.0	6.0	2.0
$t^{\mathrm{EGR}}_{l'',d}$	2.1	2.6	3.8

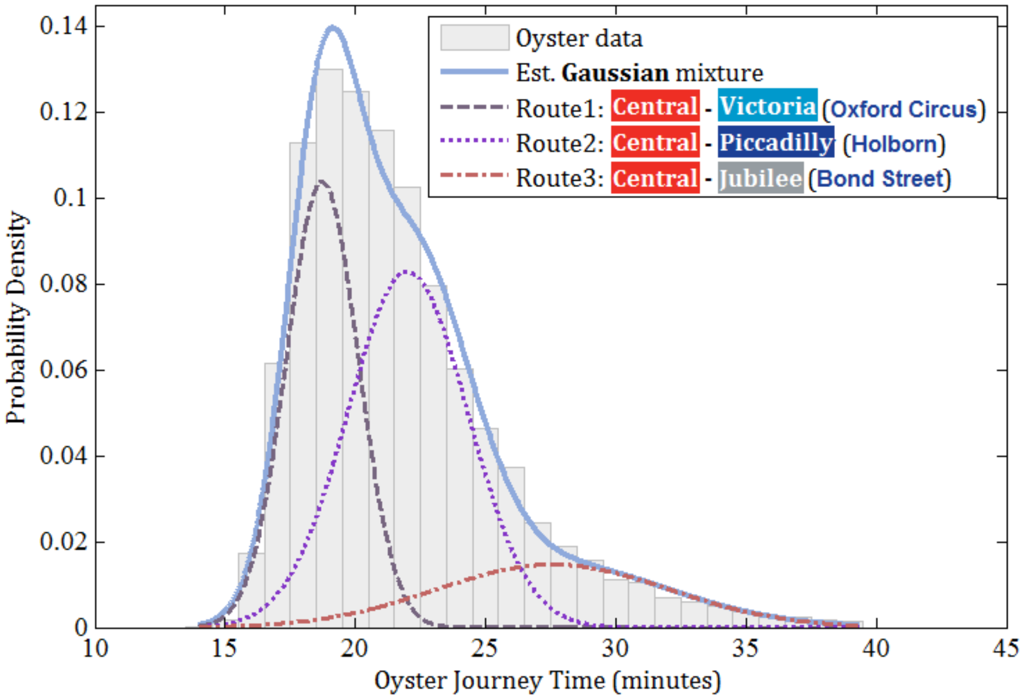
(*Continued*)

Table C.3 (*Continued.*)

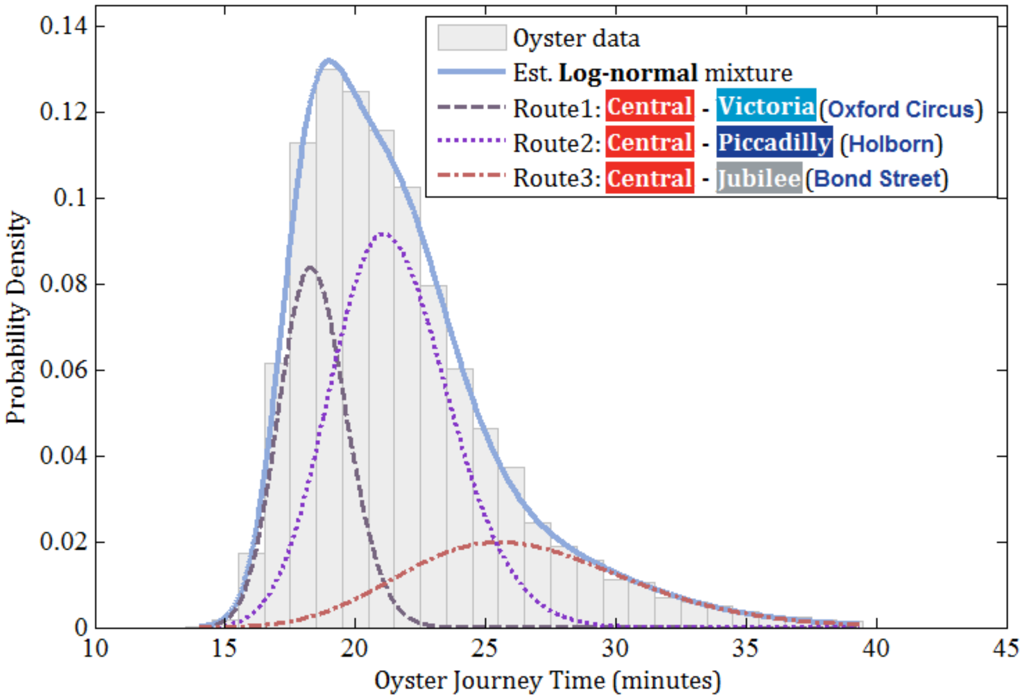
	Calculated average travel time (minutes)		
l' – l'' s	Central – Victoria Oxford Circus	Central – Piccadilly Holborn	Central – Jubilee Bond Street
Route-labels	$h=1$	$h=2$	$h=3$
Total average			
$t_h(1, 1)$	19.4	24.0	24.9
$t_h(2, 1)$	21.6	26.2	27.1
$t_h(1, 2)$	21.4	26.5	27.0
$t_h(2,2)$	23.6	28.6	29.3
t_h^{REF}	21.5	26.3	27.1

Table C.4 Matching the estimated mixture components with the real-world routes for Liverpool Street – Green Park

		r **matches** h		
Component-label r		$r=1$	$r=2$	$r=3$
Journey time (minutes)				
$\hat{\mu}_r$	GM	18.7	22.0	27.6
	LNM	18.4	21.5	26.6
t_h^{REF} $(\hat{\sigma}_h^{\text{SEM}})$		21.5 (0.9)	26.3 (0.9)	27.1 (0.9)
CI for h	95% CL	[18.7, 24.3]	[23.4, 29.2]	[24.2, 29.9]
Traffic distribution (%)				
$\hat{\omega}_r$	GM	35.9	47.7	16.4
	LNM	27.1	51.6	21.3
ω_h^{ROD} (n_h^{ROD})	AM Peak	71.3 (141)	17.9 (35)	10.2 (20)
	A weekday	45.5 (298)	41.4 (271)	13.1 (86)
Route-label h		$h=1$	$h=2$	$h=3$
		Central – Victoria Oxford Circus	Central – Piccadilly Holborn	Central – Jubilee Bond Street



(a)



(b)

Figure C.4 Estimated mixture distributions, and weighted components thereof, of OJT for **Liverpool Street – Green Park** $(n = 17{,}102)$:

(a) estimated GM model; and
(b) estimated LNM model (*see next page*).

Table C.5 Goodness-of-fit test result for Liverpool Street – Green Park

The calculation of *gof* was repeated 1,000 times for each model.

	GM	LNM
Rate of obtaining lower *gof* (%)	51.4	48.6
Average *gof*	0.0834	0.0838

C.2 *Case-7*: Victoria – Waterloo

This section shows the results of a case study of another single O-D connected by four alternative routes. Its network is illustrated with **Figure C.5** below, which involves six lines – the most among all the seven cases in this thesis.

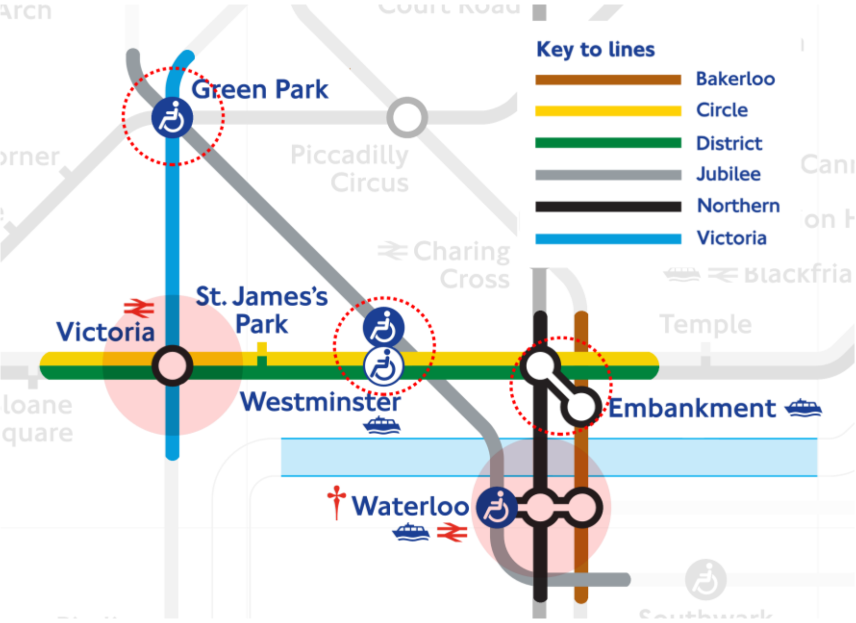


Figure C.5 The LU network that connects the O-D: Victoria – Waterloo.

In this case, passengers starting from Victoria may choose to take an eastbound Circle/District line train and transfer to a southbound train on the Jubilee line at Westminster. Alternatively, they may travel further on the same line/train to Embankment, where they could choose to change onto a southbound train on either the Northern line or the Bakerloo line. The fourth option for the passengers is to take a northbound Victoria line train at the origin station and transfer at Green Park, where they may take a southbound Jubilee line train running towards the destination, Waterloo.

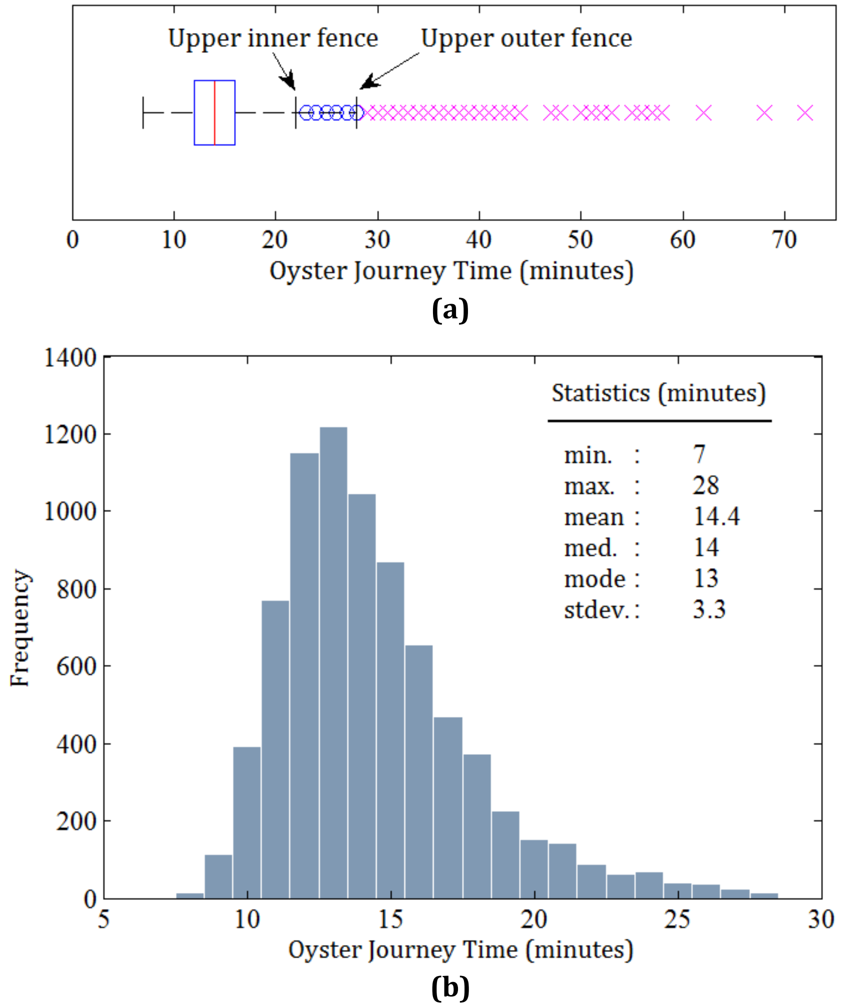


Figure C.6 Summary of OJT^{OBS} data for **Victoria – Waterloo**:

(a) a box-and-whisker plot of the raw data $(n_0 = 8{,}140)$; and
(b) a histogram of the valid data $(n = 7{,}935)$.

Table C.6 Parameter estimates of GM and LNM models based on OJT^{OBS} data for **Victoria – Waterloo**

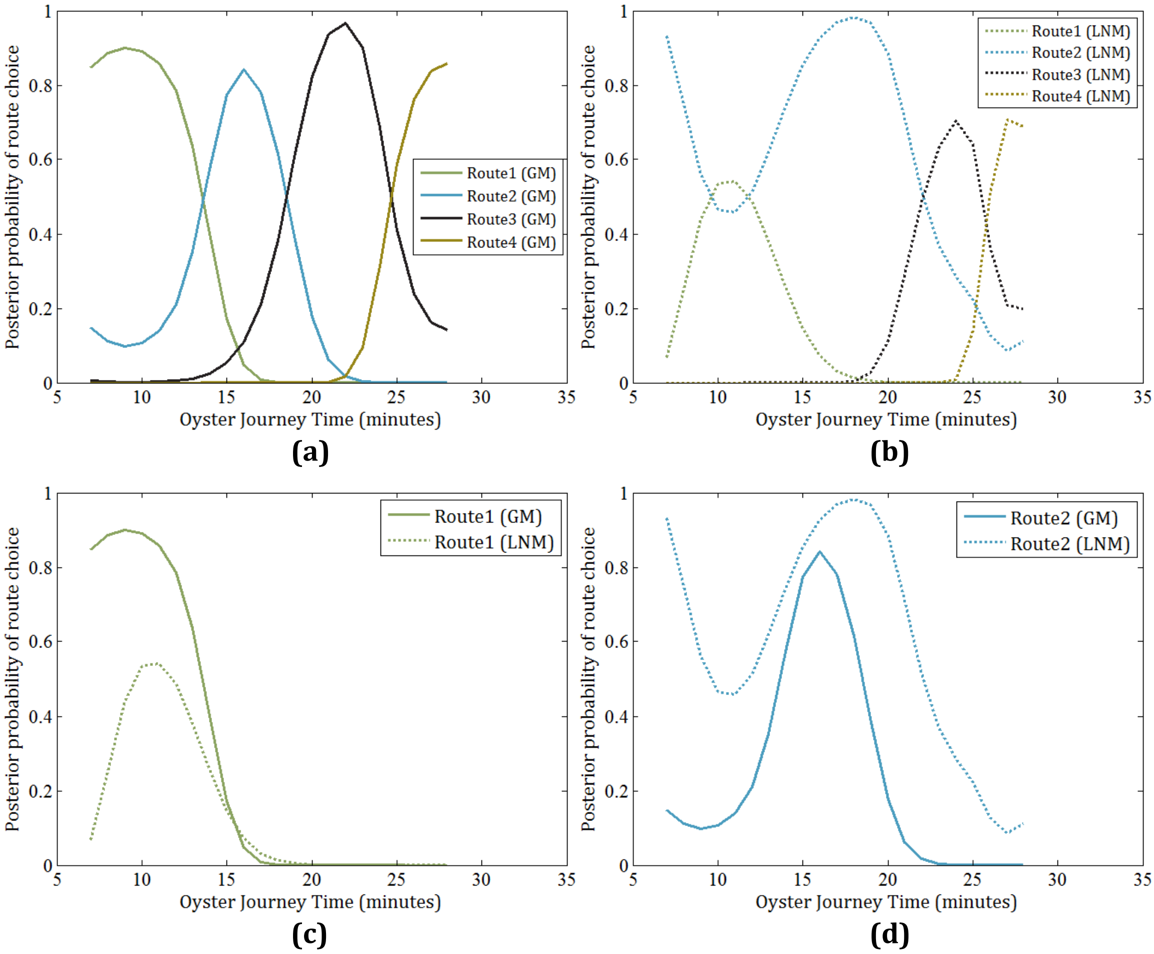
The initial values and the model parameters were estimated using the *K*-means clustering and the EM algorithm, respectively. $n = 7{,}935$.

	GM				**LNM**			
Component-label	$r=1$	$r=2$	$r=3$	$r=4$	$r=1$	$r=2$	$r=3$	$r=4$
Initial values								
η_r^{KMS} (minute)	11.0	14.0	17.0	21.0	11.0	13.0	16.0	21.1
σ_r^{KMS} (minute)	0.9	0.8	0.8	2.4	0.8	0.8	1.1	2.3
ω_r^{KMS} (%)	30.8	39.5	18.9	10.8	16.3	43.1	29.8	10.8

(*Continued*)

Table C.6 (*Continued.*)

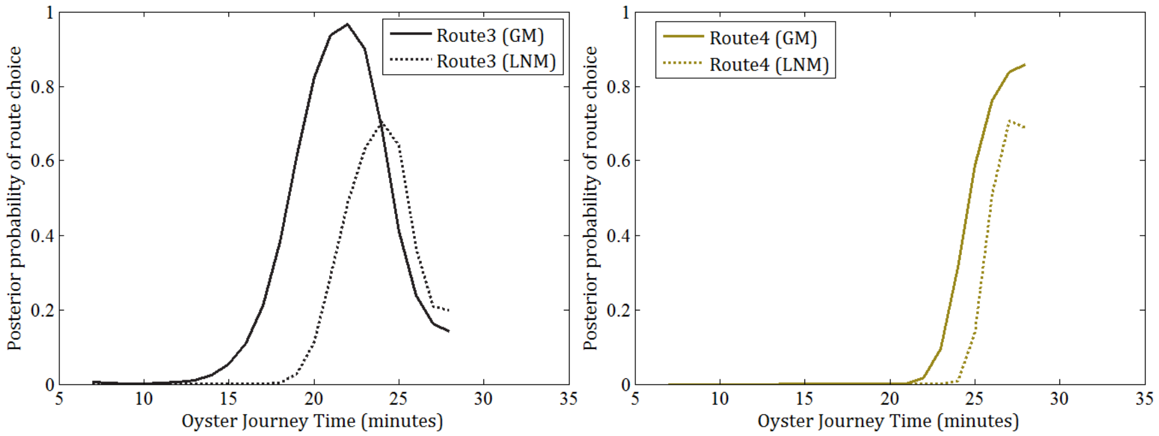
		GM				LNM			
Component-label		$r=1$	$r=2$	$r=3$	$r=4$	$r=1$	$r=2$	$r=3$	$r=4$
Parameter estimates									
$\hat{\mu}_r$	(minute)	12.1	14.9	19.3	25.6	12.3	14.8	22.8	26.6
$\hat{\sigma}_r$	(minute)	1.5	2.1	2.9	1.5	1.6	2.9	2.0	1.0
$\hat{\omega}_r$	(%)	43.0	42.6	13.0	1.4	27.3	68.9	3.1	0.7



(a) **(b)** **(c)** **(d)**

Figure C.7 Posterior probabilities of route choices given OJT^{OBS} for **Victoria – Waterloo** $(n = 7,935)$:

(a) for all alternatives, based on GM; **(b)** for all alternatives, based on LNM;
(c) for Route1, based on GM and LNM; **(d)** for Route2, based on GM and LNM;
(e) for Route3, based on GM and LNM (*see next page*); and
(f) for Route4, based on GM and LNM (*see next page*).



(e) **(f)**

Figure C.7 (*Continued.*)

Table C.7 Inferences of proportion of passenger traffic on each alternative route connecting Victoria to Waterloo ($n = 7{,}935$)

	GM				LNM			
Component-label	$r=1$	$r=2$	$r=3$	$r=4$	$r=1$	$r=2$	$r=3$	$r=4$
$\hat{\omega}_r$ (%)	43.0	42.6	13.0	1.4	27.3	68.9	3.1	0.7
$n_r^{\mathrm{INF}_0}$	3,665	3,412	741	117	1,164	6,522	172	77
$\omega_r^{\mathrm{INF}_0}$ (%)	46.2	43.0	9.3	1.5	14.7	82.2	2.2	1.0
n_r^{INF}	3,380	3,377	1,070	108	2,237	5,412	229	57
ω_r^{INF} (%)	42.6	42.6	13.5	1.4	28.2	68.2	2.9	0.7

Table C.8 Expected journey times of simulated samples for each alternative route connecting Victoria to Waterloo

	Calculated average travel times (minutes)			
l' $-\ l''$ s	Circle/District – Bakerloo Embankment	Circle/District – Northern Embankment	Circle/District – Jubilee Westminster	Victoria – Jubilee Green Park
Journey segment				
$t_{l',o}^{\mathrm{ACC}}$	2.1	2.1	2.1	2.7
$t_{l',o,1}^{\mathrm{WTD}}$ / $t_{l',o,2}^{\mathrm{WTD}}$	1.5 / 3.7	1.5 / 3.7	1.5 / 3.7	0.8 / 2.8
$t_{l',[o,s]}^{\mathrm{OBT}}$	5.0	5.0	3.0	1.0
$t_{[l',l''],s}^{\mathrm{ICT}}$	2.7	1.8	1.9	3.0
$t_{l'',s,1}^{\mathrm{ICW}}$ / $t_{l'',s,2}^{\mathrm{ICW}}$	1.3 / 4.1	1.5 / 4.8	0.5 / 2.7	0.7 / 2.9
$t_{l'',[s,d]}^{\mathrm{OBT}}$	1.0	1.0	1.0	3.0
$t_{l'',d}^{\mathrm{EGR}}$	4.8	2.5	3.1	3.1

(*Continued*)

Table C.8 (*Continued.*)

	Calculated average travel times (minutes)			
l' – l'' s	Circle/District – Bakerloo Embankment	Circle/District – Northern Embankment	Circle/District – Jubilee Westminster	Victoria – Jubilee Green Park
Route-labels	$h=1$	$h=2$	$h=3$	$h=4$
Total average				
$t_h(1, 1)$	18.5	15.3	13.2	14.3
$t_h(2, 1)$	20.6	17.5	15.4	16.3
$t_h(1, 2)$	21.3	18.6	15.4	16.5
$t_h(2,2)$	23.4	20.7	17.6	18.5
t_h^{REF}	20.9	18.1	15.4	16.4

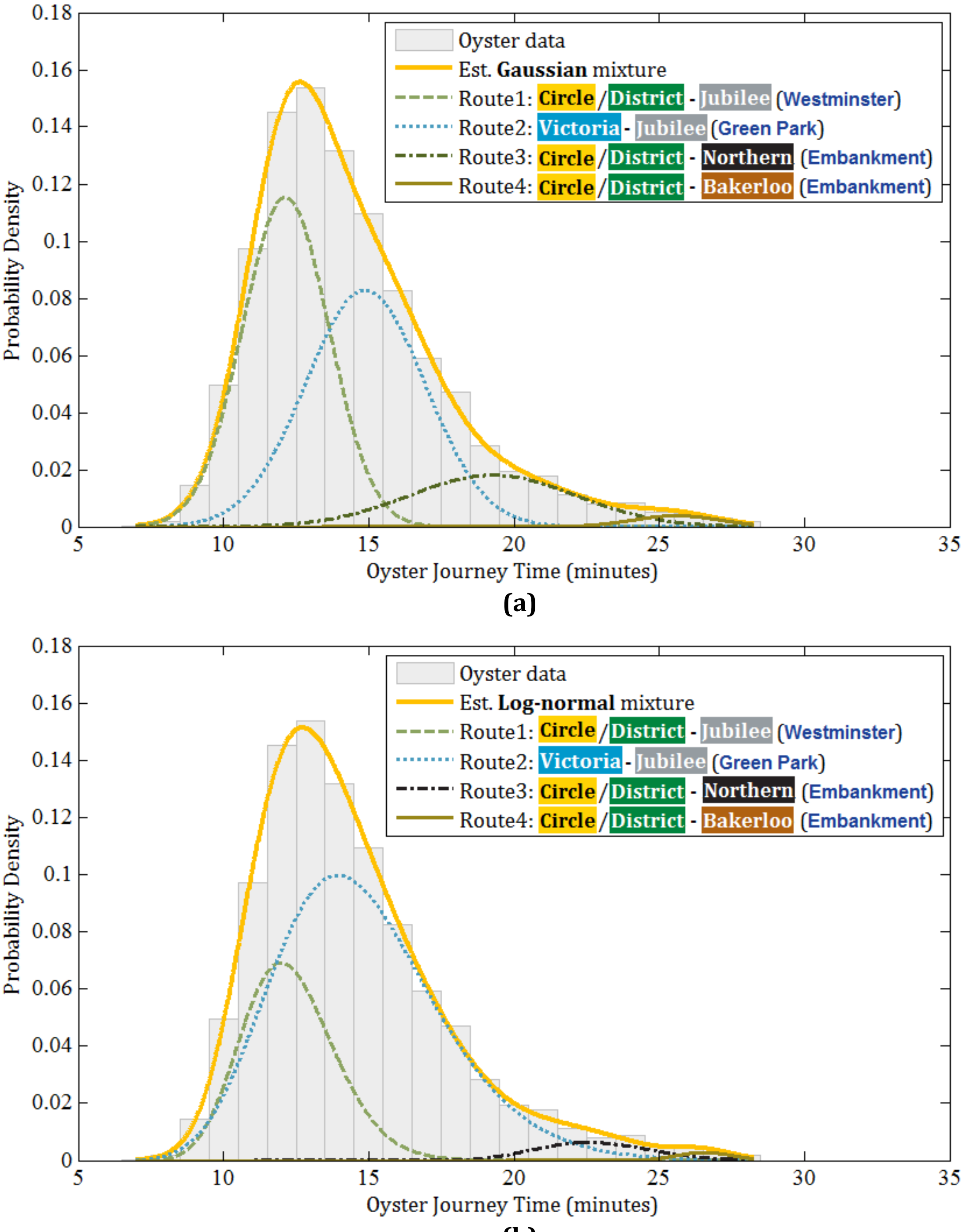
Table C.9 Matching the estimated mixture components with the real-world routes for Victoria – Waterloo

		r **matches** h			
Component-label r		$r=1$	$r=2$	$r=3$	$r=4$
Journey time (minutes)					
$\hat{\mu}_r$	GM	12.1	14.9	19.3	25.6
	LNM	12.3	14.8	22.8	26.6
t_h^{REF} $(\hat{\sigma}_h^{\mathrm{SEM}})$		15.4 (0.9)	16.4 (0.9)	18.1 (1.1)	20.9 (1.0)
CI for h	95% CL	[12.6, 18.3]	[13.7,19.1]	[14.5,21.6]	[17.7, 24.2]
Traffic distribution (%)					
$\hat{\omega}_r$	GM	43.0	42.6	13.0	1.4
	LNM	27.3	68.9	3.1	0.7
ω_h^{ROD} (n_h^{ROD})	AM Peak	48.2 (186)	36.5 (141)	15.3 (59)	
	A whole day	40.9 (410)	20.5 (206)	38.6 (387)	
Route-label h		$h=1$	$h=2$	$h=3$	$h=4$
		Circle/District – Jubilee Westminster	Victoria – Jubilee Green Park	Circle/District – Northern Embankment	Circle/District – Bakerloo Embankment

Table C.10 Goodness-of-fit test result for Victoria – Waterloo

The calculation of *gof* was repeated 1,000 times for each model.

	GM	**LNM**
Rate of obtaining lower *gof* (%)	44.6	55.4
Average *gof*	0.0916	0.0911



(a)

(b)

Figure C.8 Estimated mixture distributions, and weighted components thereof, of *OJT* for **Victoria – Waterloo** $(n = 7,935)$:

(a) estimated GM model; and
(b) estimated LNM model.

Appendix D
Updated posterior probabilities for *Case-2* – *Case-7*

D.1 Cases of two alternative routes

D.1.1 *Case-2*: Euston – St. James's Park

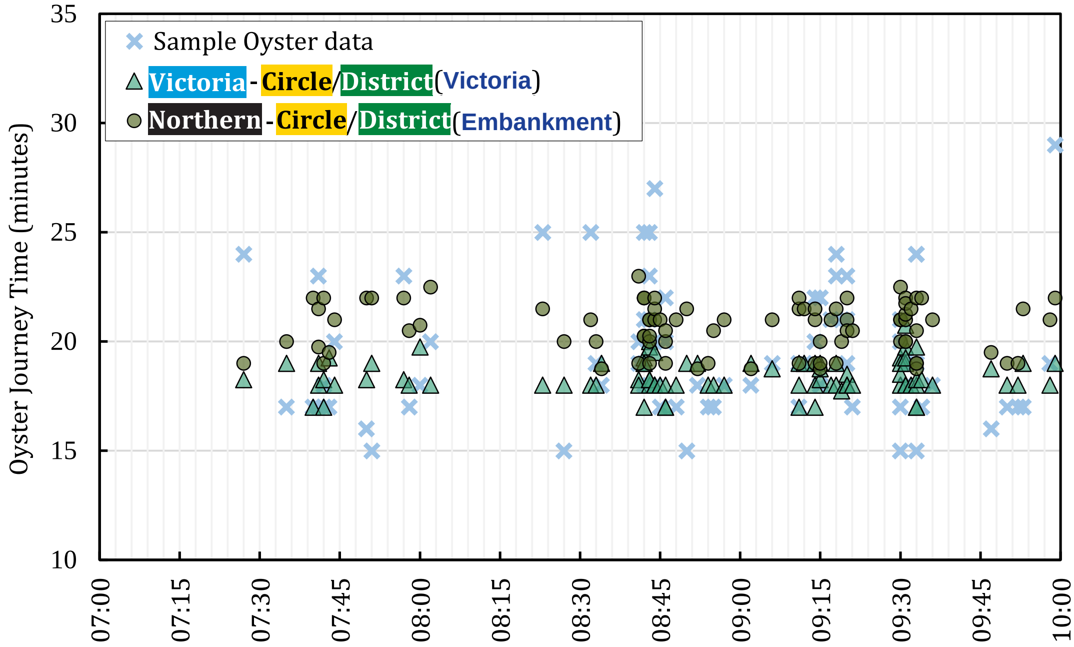
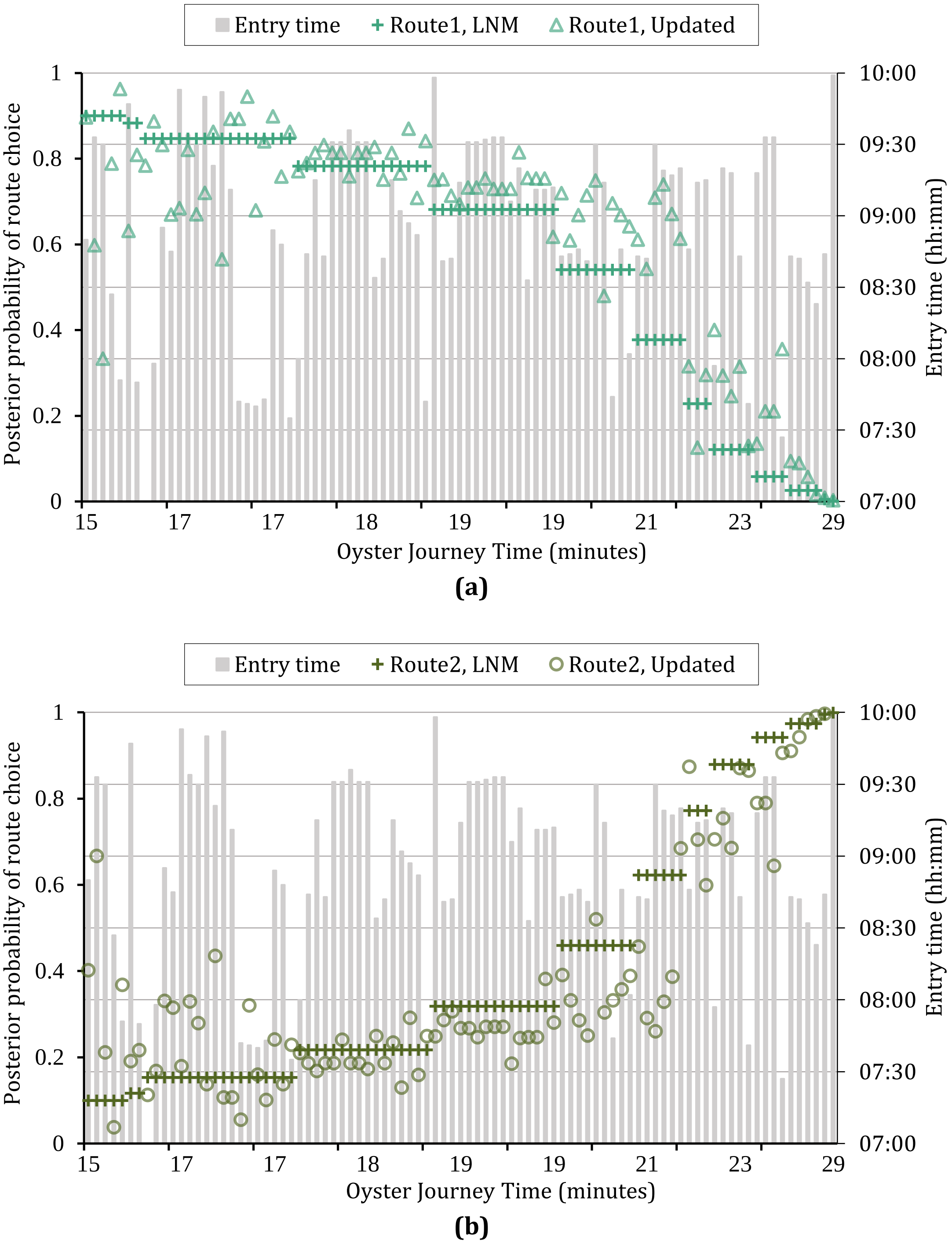


Figure D.1 Comparisons between OJT_q^{OBS} and δ_{qr}^{EXP} $\forall q, r$, given $\Delta_{5\%}$ for Euston – St. James's Park.

Table D.1 Proportion of passenger traffic for each alternative route on Euston – St. James's Park

In this case, ω_r^{INF} is calculated on the basis of LNM model estimates.

	Sample size	**Proportion of passenger-traffic** (%)	
		Victoria – Circle/District (Victoria) $r=1$	Northern – Circle/District (Embankment) $r=2$
ω_r^{ROD}	437	42.8	57.2
$\hat{\omega}_r$	22,379	82.8	17.2
ω_r^{INF}	22,379	82.7	17.3
ω_r^{INF}	89	76.7	23.3
ω_r^{UPD}	89	78.3	21.7



(a)

(b)

Figure D.2 Comparisons between π_{qr}^{MIX} (based on LNM) and π_{qr}^{UMM} for **Euston – St. James's Park**:

(a) Route1: **Victoria** – **Circle**/**District** (via **Victoria**); and
(b) Route2: **Northern** – **Circle**/**District** (via **Embankment**).

The interval between the tick-marks on the horizontal axis spans 10 bars each relating to an individual/journey record in the Oyster data.

D.1.2 *Case-3*: Victoria – Liverpool Street

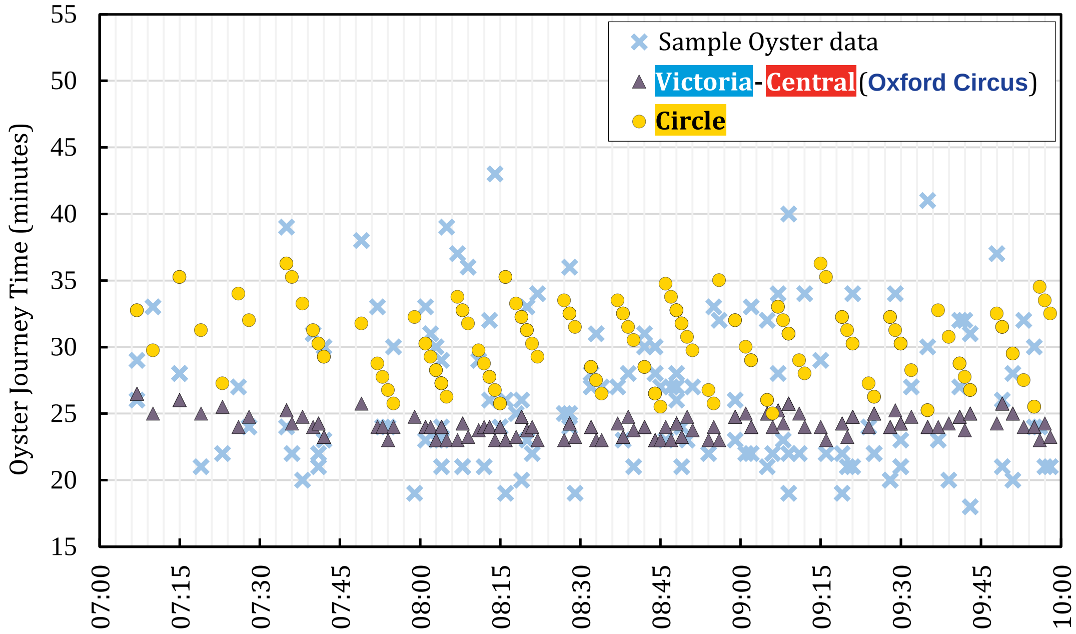
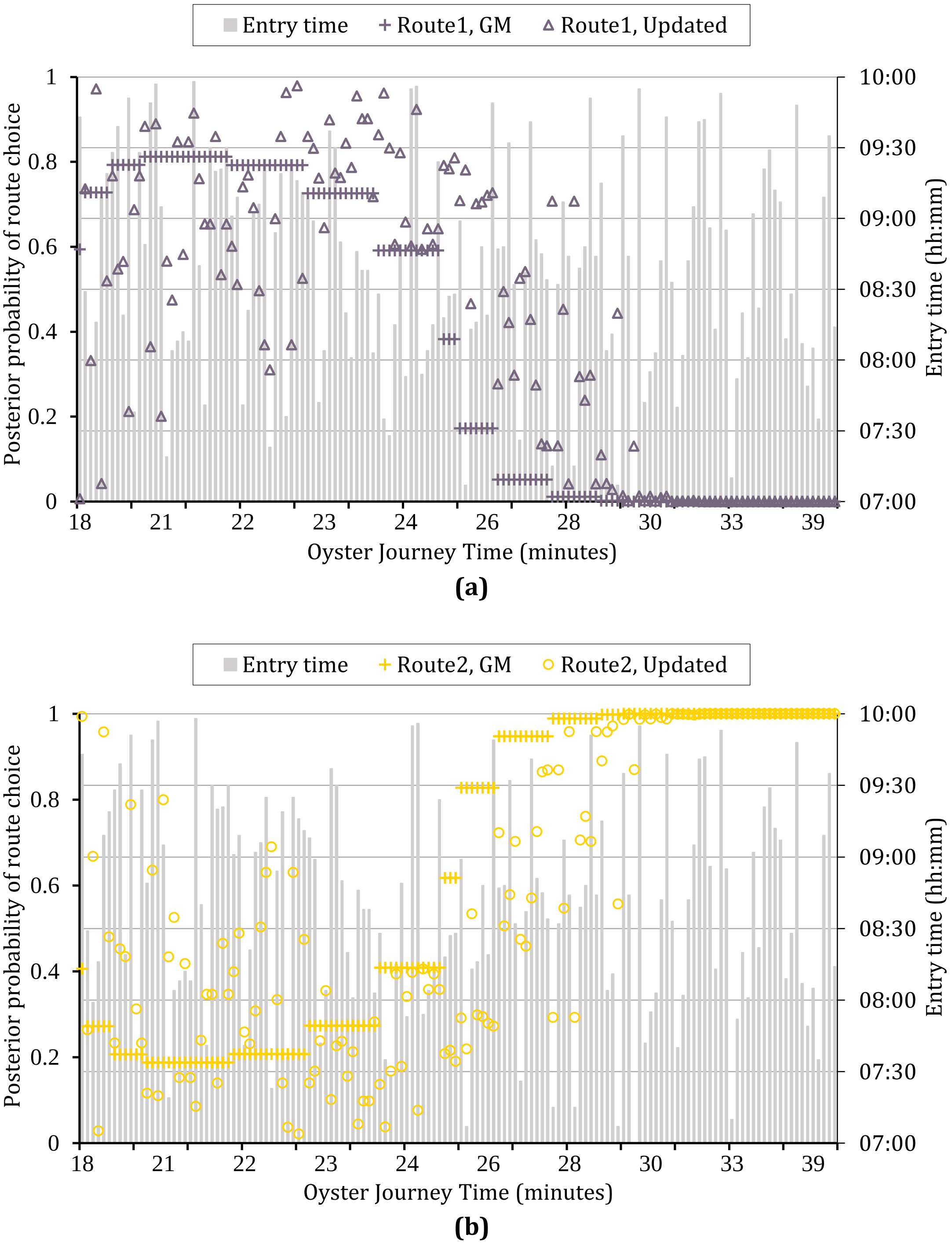


Figure D.3 Comparisons between OJT_q^{OBS} and δ_{qr}^{EXP} $\forall q, r$, given $\Delta_{5\%}$ for Victoria – Liverpool Street.

Table D.2 Proportion of passenger traffic for each alternative route on Victoria – Liverpool Street

In this case, ω_r^{INF} is calculated on the basis of LNM model estimates.

	Sample size	**Proportion of passenger-traffic** (%)	
		Victoria – Central (Oxford Circus) $r=1$	Circle – $r=2$
ω_r^{ROD}	557	48.1	51.9
$\hat{\omega}_r$	36,262	35.5	64.5
ω_r^{INF}	36,262	35.2	64.8
ω_r^{INF}	140	37.7	62.3
ω_r^{UPD}	140	42.6	57.4



(a)

(b)

Figure D.4 Comparisons between π_{qr}^{MIX} (based on LNM) and π_{qr}^{UMM} for **Euston – St. James's Park**:

(a) Route1: **Victoria** – **Central** (via **Oxford Circus**); and
(b) Route2: **Circle** (*direct service*).

The interval between the tick-marks on the horizontal axis spans 10 bars each relating to an individual/journey record in the Oyster data.

D.2 Cases of three alternative routes

D.2.1 *Case-4*: Angel – Waterloo

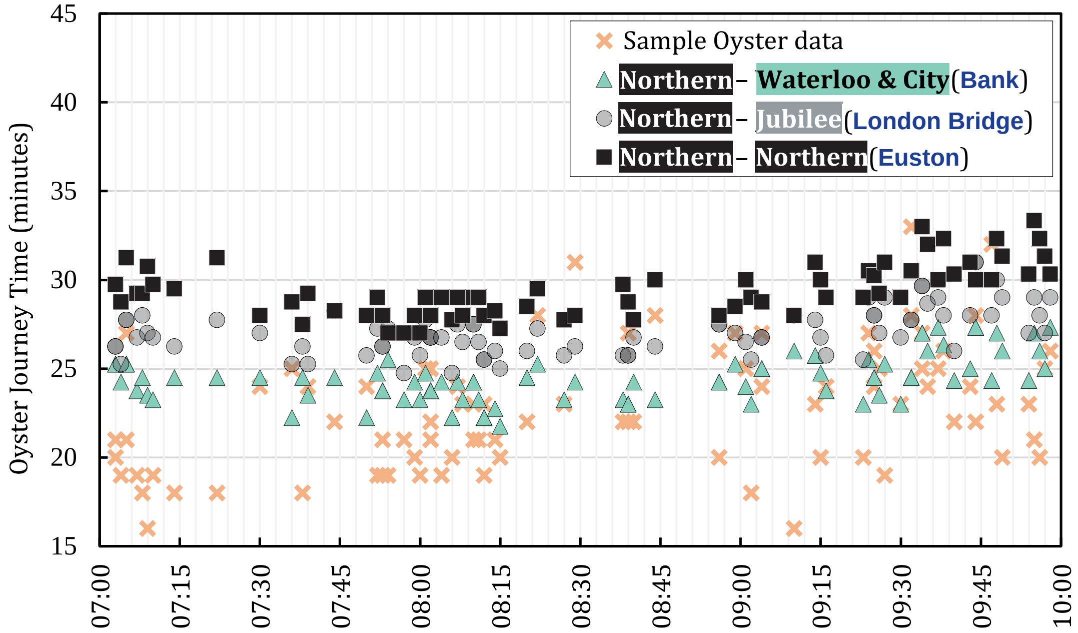
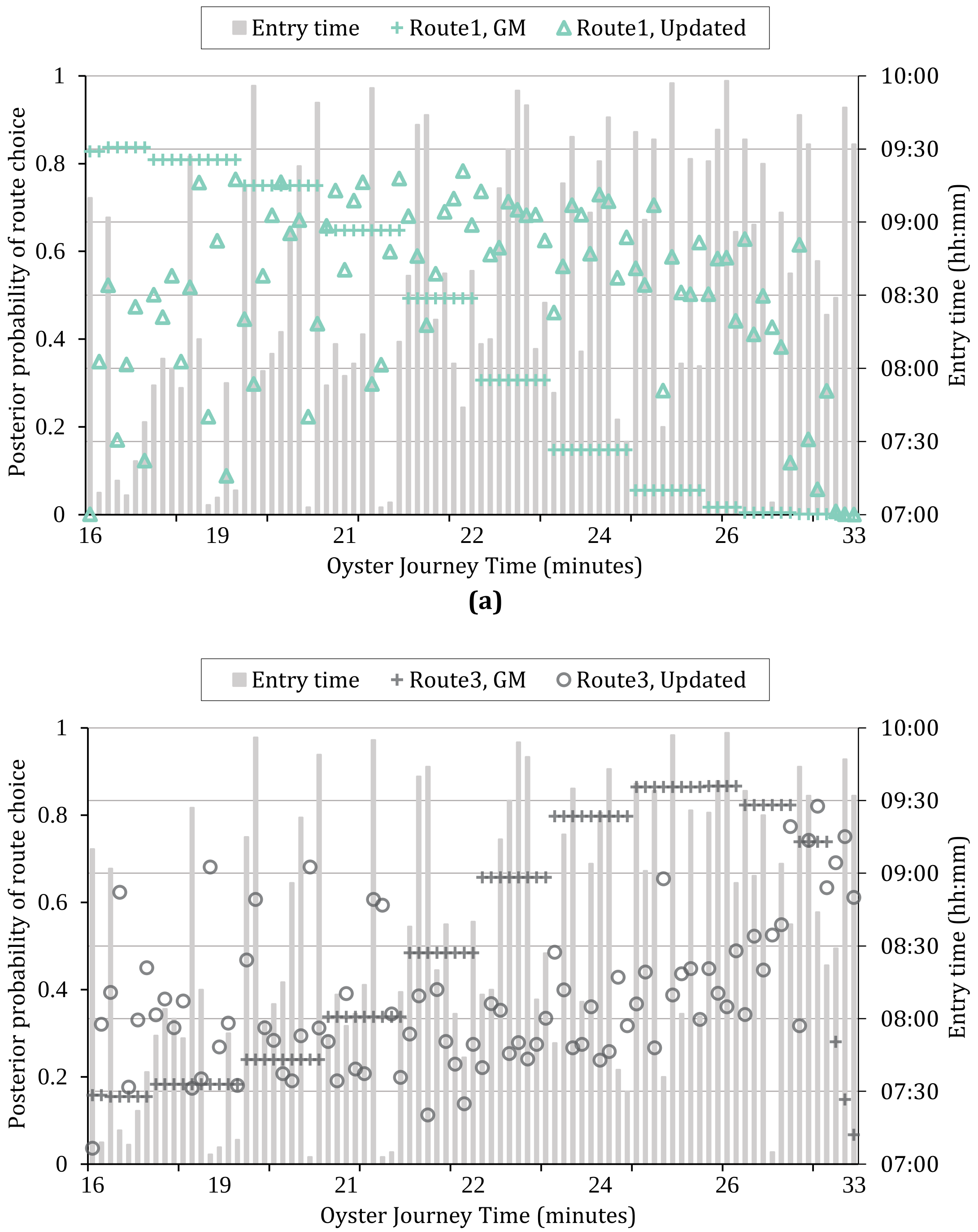


Figure D.5 Comparisons between OJT_q^{OBS} and δ_{qr}^{EXP} $\forall q, r$, given $\Delta_{5\%}$ for **Angel – Waterloo**.

Table D.3 Proportion of passenger traffic for each alternative route on **Angel – Waterloo**

In this case, ω_r^{INF} is calculated on the basis of GM model estimates.

	Sample size	**Proportion of passenger-traffic** (%)		
		Northern – **Waterloo & City** (**Bank**) $r = 1$	**Northern** – **Jubilee** (**London Bridge**) $r = 2$	**Northern** – **Northern** (**Euston**) $r = 3$
ω_r^{ROD}	77	42.9	44.1	13.0
$\hat{\omega}_r$	14,419	39.8	49.6	10.6
ω_r^{INF}	14,419	39.4	49.9	10.7
ω_r^{INF}	85	40.9	50.8	8.3
ω_r^{UPD}	85	50.3	38.0	11.8



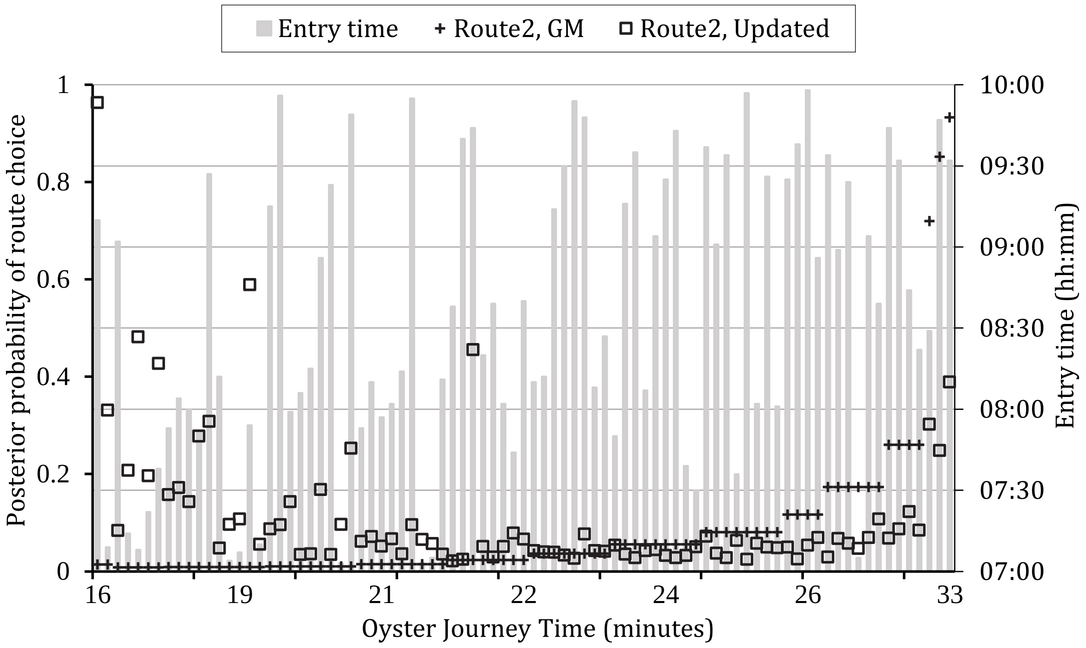
(a)

(b)

Figure D.6 Comparisons between π_{qr}^{MIX} (based on GM) and π_{qr}^{UMM} for **Angel – Waterloo**:

(a) Route1: **Northern** – **Waterloo & City** (via **Bank**);
(b) Route2: **Northern** – **Jubilee** (via **London Bridge**); and
(c) Route3: **Northern** – **Northern** (via **Euston**) (*see next page*).

The interval between the tick-marks on the horizontal axis spans 10 bars each relating to an individual/journey record in the Oyster data.



(c)

Figure D.6 (*Continued.*)

D.2.2 *Case-5*: Liverpool Street – Green Park

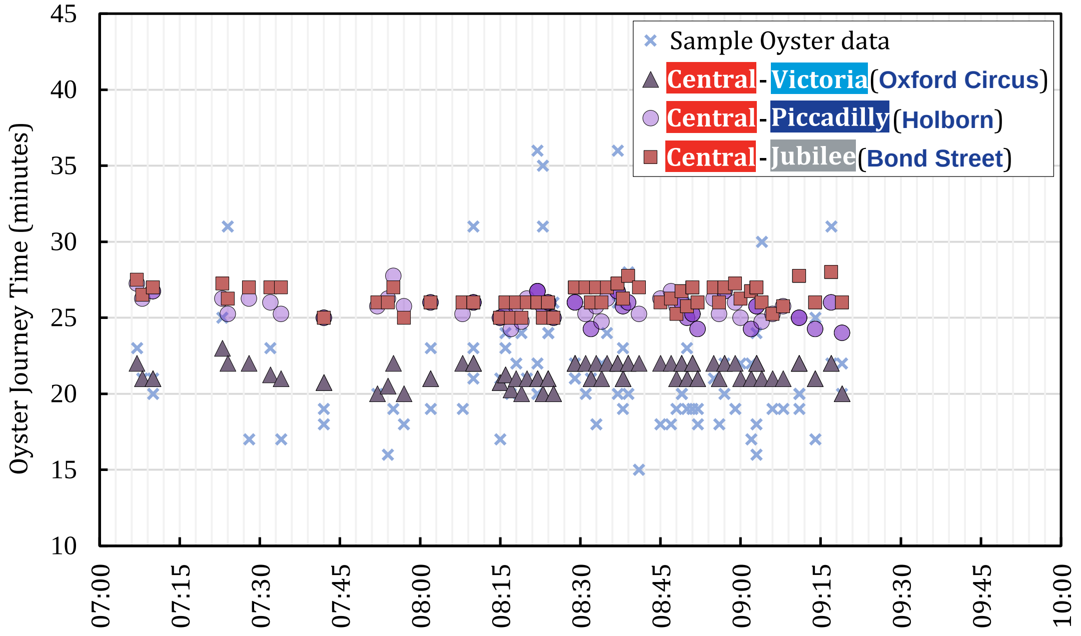
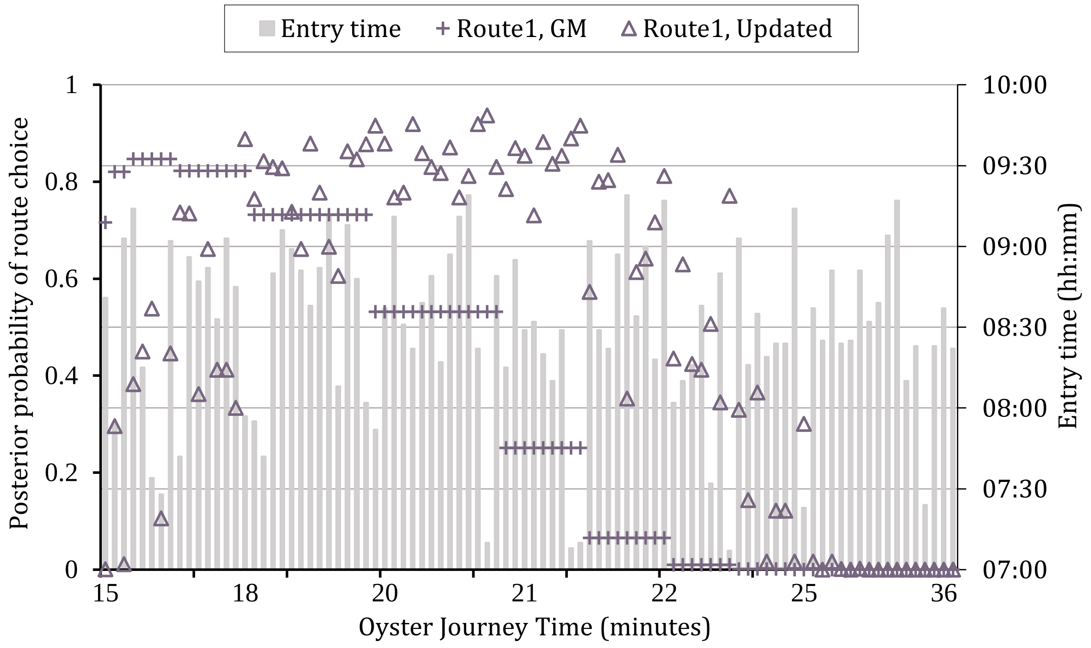


Figure D.7 Comparisons between OJT_q^{OBS} and δ_{qr}^{EXP} $\forall q, r$, given $\Delta_{5\%}$ for **Liverpool Street – Green Park**.

Table D.4 Proportion of passenger traffic for each alternative route on **Liverpool Street – Green Park**

In this case, ω_r^{INF} is calculated on the basis of GM model estimates.

	Sample size	**Proportion of passenger-traffic** (%)		
		Central – Victoria (**Oxford Circus**) $r=1$	Central – Piccadilly (**Holborn**) $r=2$	Central – Jubilee (**Bond Street**) $r=3$
ω_r^{ROD}	196	71.3	17.9	10.2
$\hat{\omega}_r$	17,102	35.9	47.7	16.4
ω_r^{INF}	17,102	35.4	48.3	16.3
ω_r^{INF}	92	35.9	46.3	17.8
ω_r^{UPD}	92	51.8	24.6	23.7

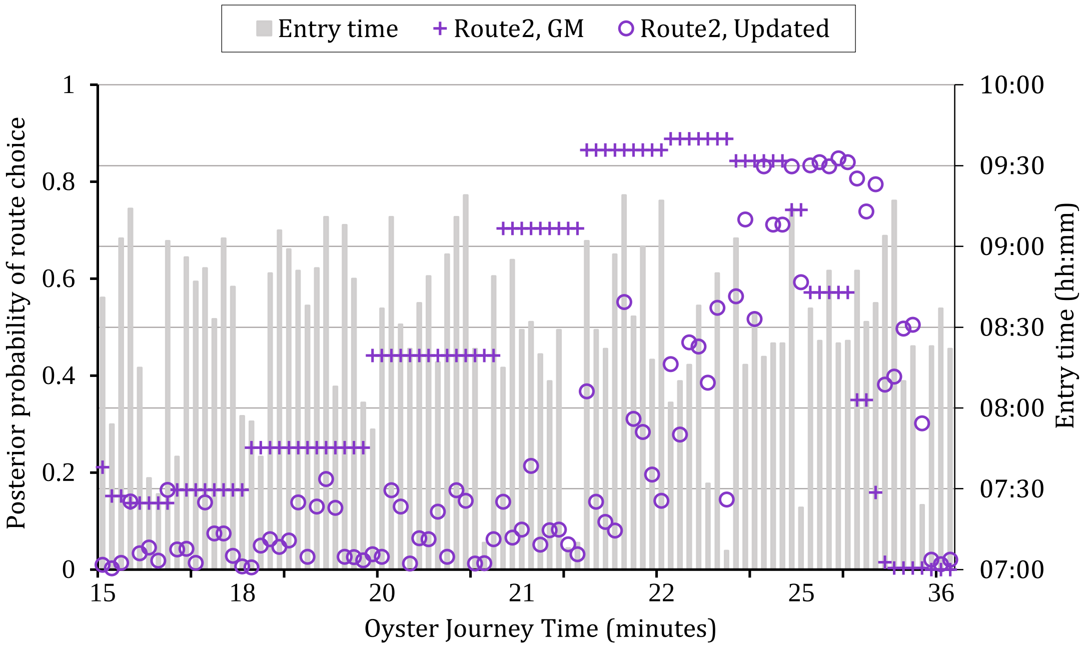


(a)

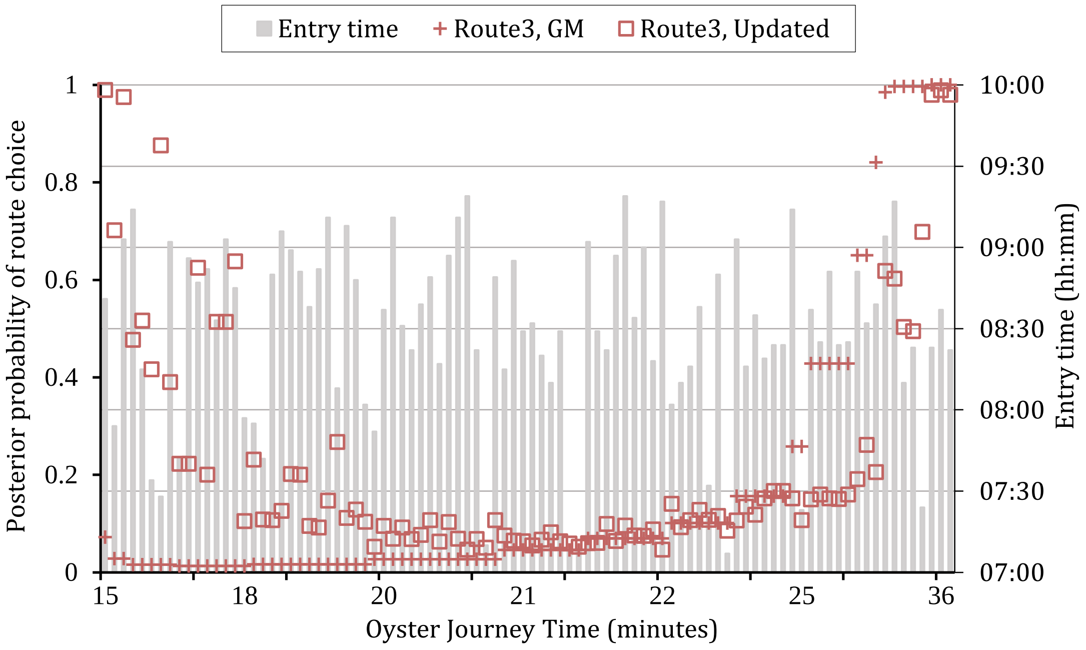
Figure D.8 Comparisons between π_{qr}^{MIX} (based on GM) and π_{qr}^{UMM} for **Liverpool Street – Green Park**:

(a) Route1: Central – Victoria (via **Oxford Circus**);
(b) Route2: Central – Piccadilly (via **Holborn**) (*see next page*); and
(c) Route3: Central – Jubilee (via **Bond Street**) (*see next page*).

The interval between the tick-marks on the horizontal axis spans 10 bars each relating to an individual/journey record in the Oyster data.



(b)



(c)

Figure D.8 (*Continued.*)

D.3 Cases of four alternative routes

D.3.1 *Case-6*: Euston – South Kensington

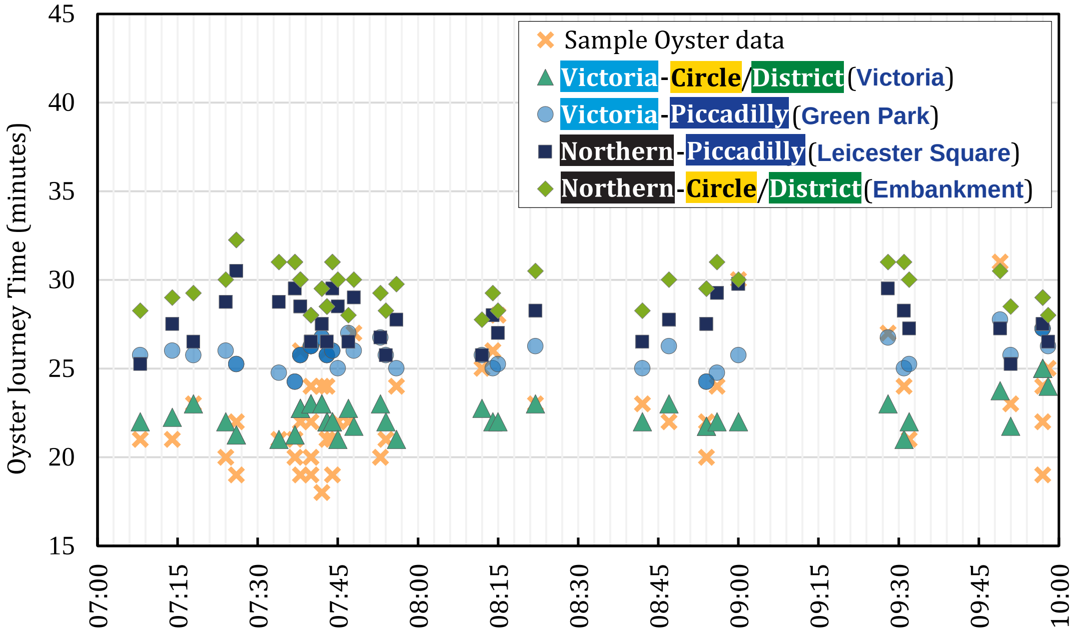
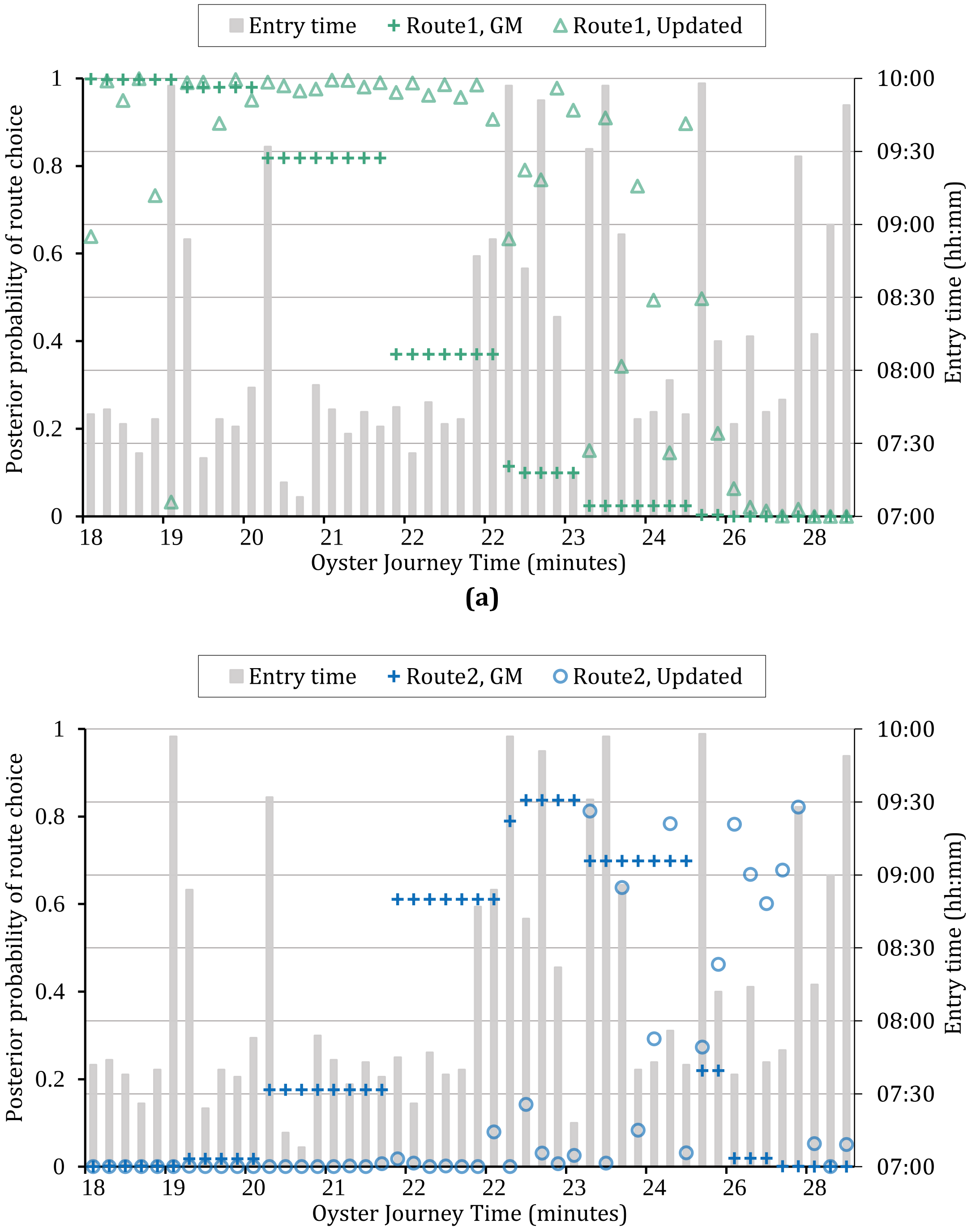


Figure D.9 Comparisons between OJT_q^{OBS} and δ_{qr}^{EXP} $\forall q, r$, given $\Delta_{5\%}$ for **Euston – South Kensington**.

Table D.5 Proportion of passenger traffic for each alternative route on **Euston – South Kensington**

In this case, ω_r^{INF} is calculated on the basis of GM model estimates.

	Sample size	**Proportion of passenger-traffic** (%)			
		Victoria – Circle/District (Victoria) $r=1$	Victoria – Piccadilly (Green Park) $r=2$	Northern – Piccadilly (Leicester Sq.) $r=3$	Northern – Circle/District (Embankment) $r=4$
ω_r^{ROD}	209	57.4	21.05	21.05	0.5
$\hat{\omega}_r$	8,116	40.9	26.6	19.8	12.7
ω_r^{INF}	8,116	40.8	26.4	20.4	12.4
ω_r^{INF}	48	43.3	31.6	17.5	7.6
ω_r^{UPD}	48	67.4	15.3	7.6	9.7



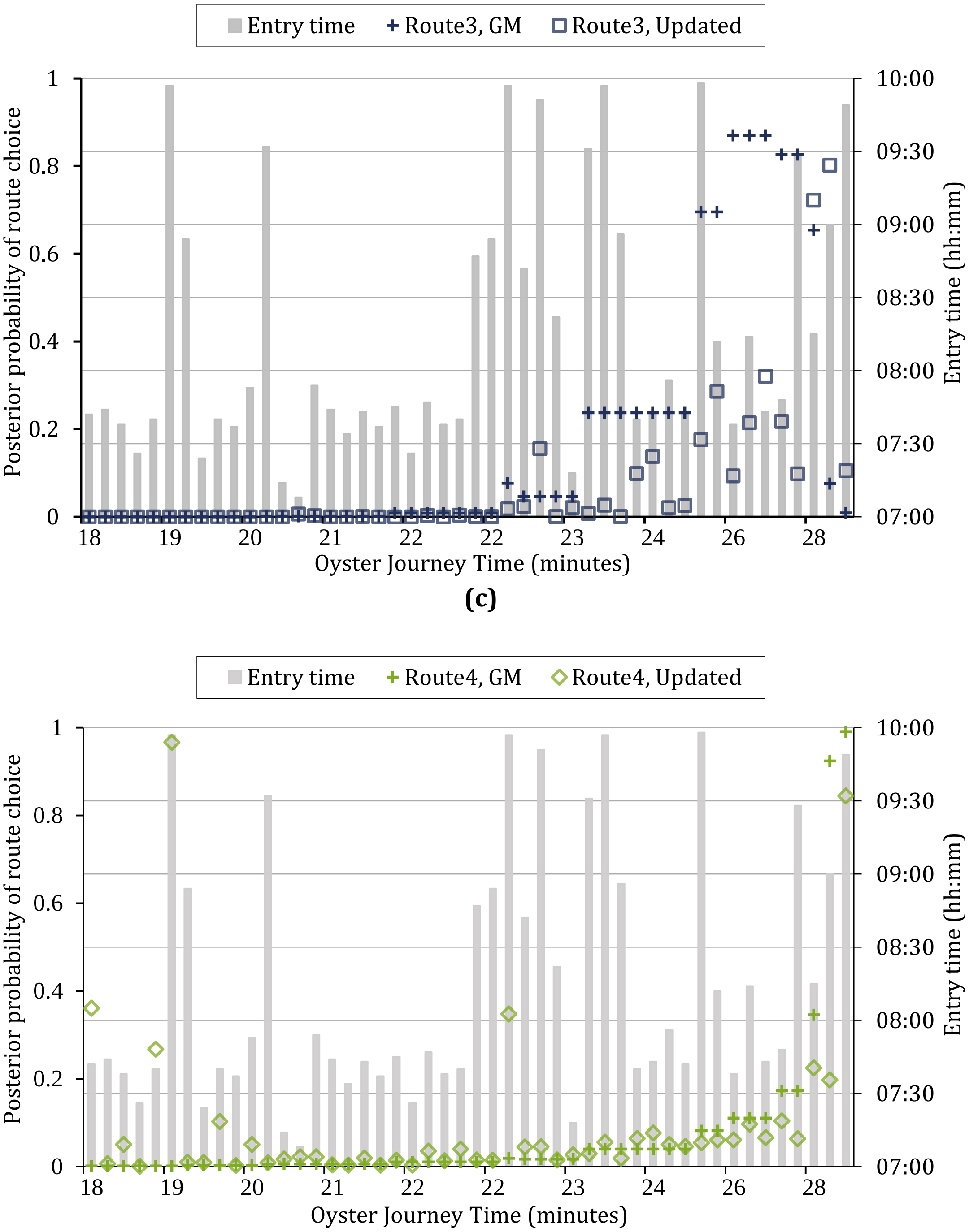
(a)

(b)

Figure D.10 Comparisons between π_{qr}^{MIX} (based on GM) and π_{qr}^{UMM} for **Euston – South Kensington**:

(a) Route1: **Victoria** – **Circle**/**District** (via **Victoria**);
(b) Route2: **Victoria** – **Piccadilly** (via **Green Park**);
(c) Route3: **Northern** – **Piccadilly** (via **Leicester Square**) (*see next page*); and
(d) Route4: **Northern** – **Circle**/**District** (via **Embankment**) (*see next page*).

The interval between the tick-marks on the horizontal axis spans 5 bars each relating to an individual/journey record in the Oyster data.



(c)

(d)

Figure D.10 (*Continued.*)

D.3.2 *Case-7*: Victoria – Waterloo

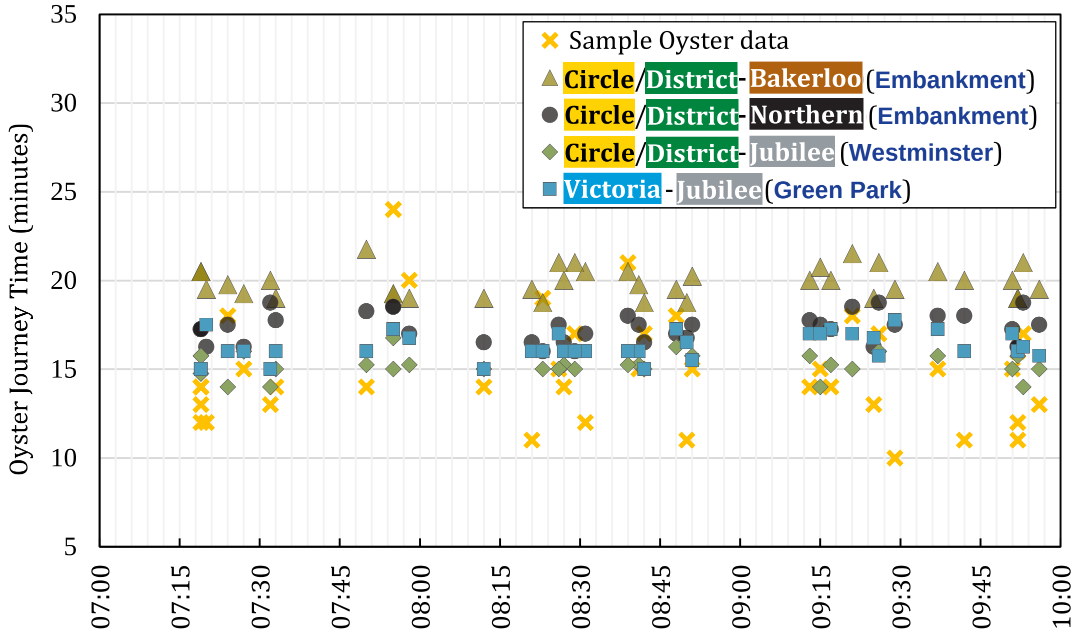
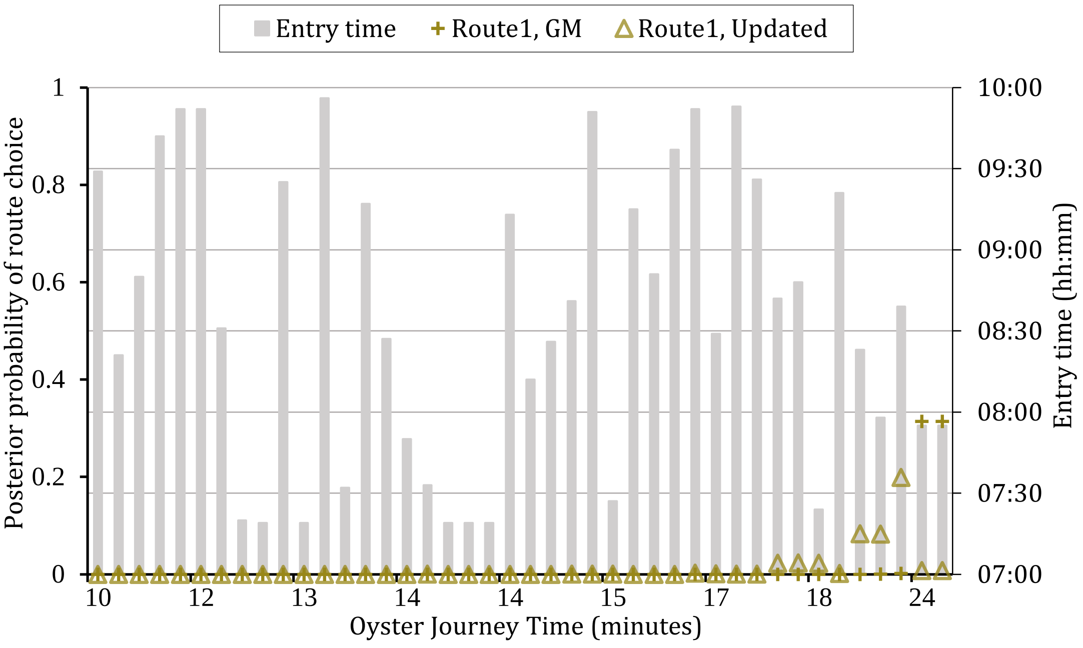


Figure D.11 Comparisons between OJT_q^{OBS} and δ_{qr}^{EXP} $\forall q, r$, given $\Delta_{5\%}$ for **Victoria – Waterloo**.

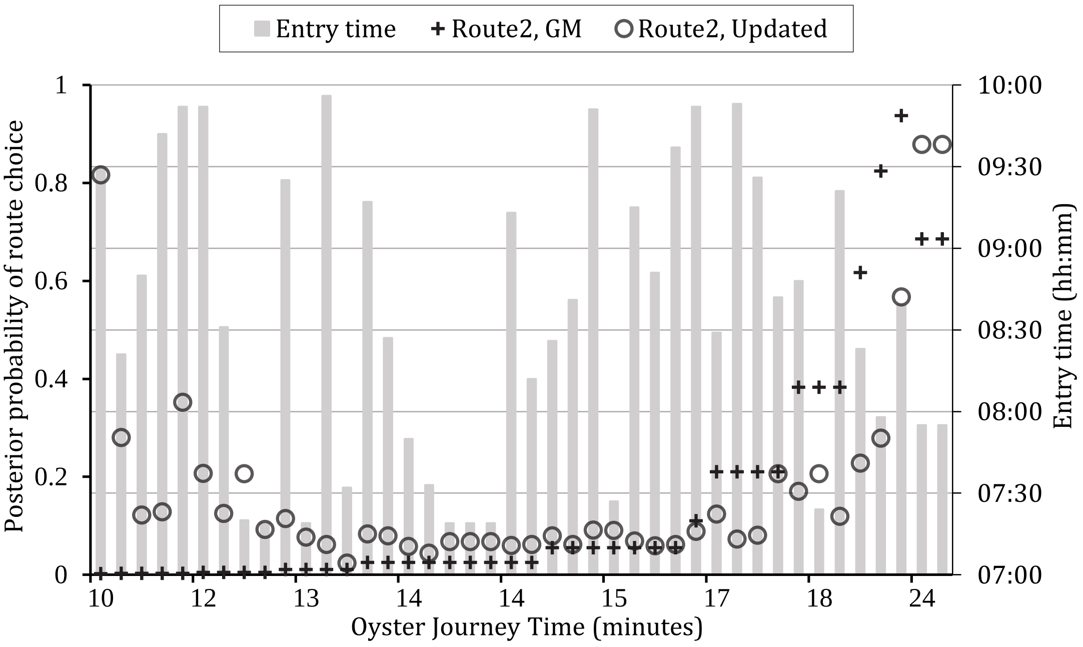
Table D.6 Proportion of passenger traffic for each alternative route on **Victoria – Waterloo**

In this case, ω_r^{INF} is calculated on the basis of GM model estimates.

	Sample size	**Proportion of passenger-traffic** (%)			
		Circle/District – Bakerloo (**Embankment**) $r=1$	**Circle/District – Northern** (**Embankment**) $r=2$	**Circle/District – Jubilee** (**Westminster**) $r=3$	**Victoria – Jubilee** (**Green Park**) $r=4$
ω_r^{ROD}	386	15.3		48.2	36.5
$\hat{\omega}_r$	7,935	1.4	13.0	43.0	42.6
ω_r^{INF}	7,935	1.4	13.5	42.6	42.6
ω_r^{INF}	42	1.6	15.5	35.7	47.2
ω_r^{UPD}	42	1.1	18.0	40.7	40.2



(a)

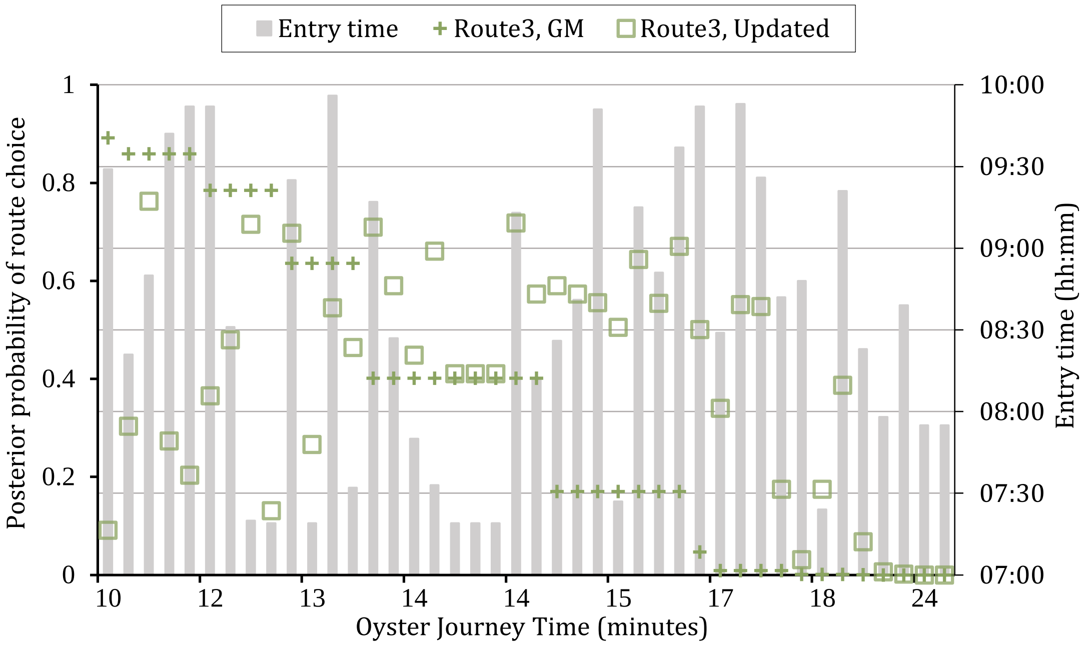


(b)

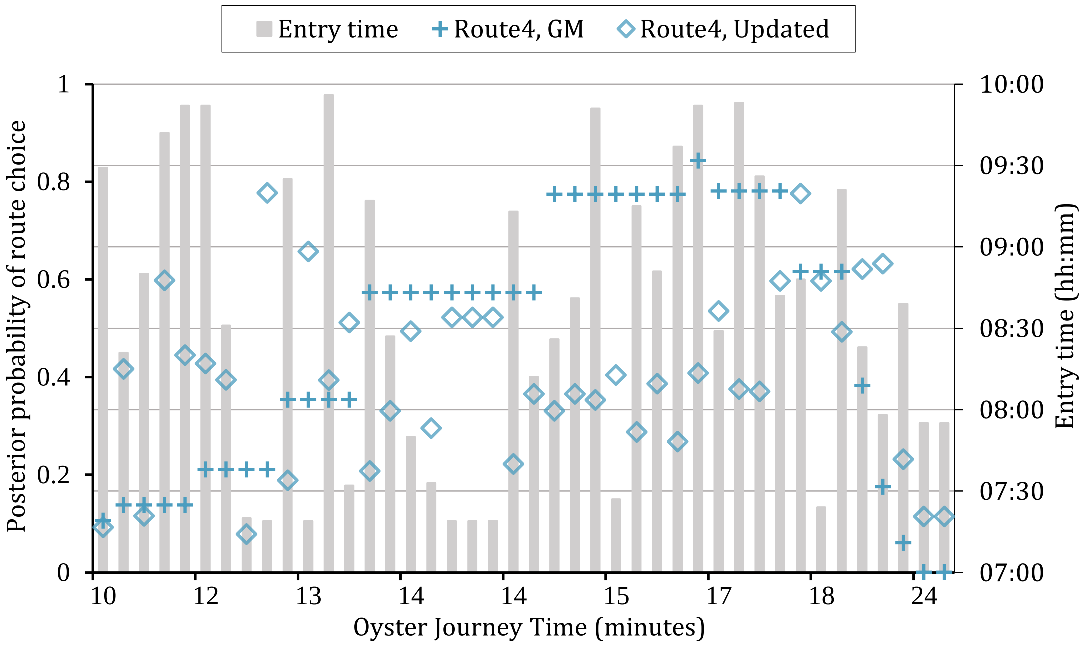
Figure D.12 Comparisons between π_{qr}^{MIX} (based on GM) and π_{qr}^{UMM} for **Victoria – Waterloo**:

(a) Route1: **Circle**/**District** – **Bakerloo** (via **Embankment**);
(b) Route2: **Circle**/**District** – **Northern** (via **Green Park**);
(c) Route3: **Circle**/**District** – **Jubilee** (via **Westminster**) (*see next page*); and
(d) Route4: **Victoria** – **Jubilee** (via **Green Park**) (*see next page*).

The interval between the tick-marks on the horizontal axis spans 5 bars each relating to an individual/journey record in the Oyster data.



(c)



(d)

Figure D.12 (*Continued.*)

Appendix E
Notation

E.1 Symbols based on the English alphabet

(Listed in alphabetical order)

$a(\cdot)$	function used for K-means clustering, which labels an journey time observation as belonging to a certain cluster
C_q	set of all elementary events within the sigma-field on the set of all possible route choices of passenger q
$C^{(\delta)}$	set of all elementary events of passengers' possible route choices, given the observations of every individual's journey time δ
$C_q^{(\delta)}$	set of all elementary events of possible route choices of passenger q, given his/her journey time δ_q^{OBS} (as an observed value of journey time variable δ)
$c_r(\cdot)$	PDF of journey time distribution of route r
$choice_{qr}$	elementary event that passenger q chose route r to make a single journey
$choice_{qr}^{(\delta)}$	elementary event that passenger q chose route r to make a single journey between a given O-D pair and spent a journey time of δ_q^{OBS} (as an observed value of journey time variable δ)
d	destination station of a give O-D pair
$f_{qr}(\cdot)$	PDF of journey time distribution of passenger q making a single journey by route r between a given pair of O-D stations
gof	indicator of **g**oodness-**o**f-**f**it between observed and simulated journey time data
h	label of travel route, referred to as 'route-label'
i, j	individual traveller/passenger (only used briefly so as to distinguish different individuals; $i \neq j$)
K	total number of clusters used in K-means clustering
l'	transit line for the first leg of a single journey

(*Continued*)

(*Continued*)

l''	transit line for the second leg of a single journey, referred to as ‘connecting line’
$l(\cdot)$	indicator function that indicates whether a transit line is l' or l''
$\ell(\cdot)$	likelihood function
$m(\cdot)$	PDF of a mixture distribution of journey time for a given O-D pair
N_Q	total number of passengers
N_R	total number of travel routes
$\mathbb{N}_{\geq 1}$	set of the natural numbers that are greater than or equal to 1
$\mathbb{N}_{\geq 2}$	set of the natural numbers that are greater than or equal to 2
$\mathcal{N}(\mu_r, \sigma_r)$	normal distribution given mean μ_r and standard deviation σ_r
n	sample size of a given data set
$n_r^{\texttt{INF}}$	number of passengers who chose route r, based on *effective* **inf**erence ($\texttt{INF}$) from a mixture model
$n_r^{\texttt{INF}_0}$	number of passengers who chose route r, based on *naive* **inf**erence ($\texttt{INF}_0$) from a mixture model
OJT	Oyster journey time
$OJT^{\texttt{OBS}}$	**obs**erved ($\texttt{OBS}$) (or, **obs**ervations of) Oyster journey time
o	origin station of a give O-D pair
$\Pr(\cdot)$	probability measure
Q	statistical population of passengers between a given O-D pair
Q_r	subpopulation of all passengers who chose route r
q	individual traveller/passenger
R	set of all alternative routes connecting a given pair of O-D stations, referred to as ‘route-choice set’
R_q	personal route-choice set of passenger q travelling between a given pair of O-D stations
r	travel route between a given pair of O-D stations; also component-label for mixture models and (cluster-label for) $K\text{-means}$ clustering

(*Continued*)

(*Continued*)

$r^{(q)}$ categorical variable of component-label, indicating the route choice of passenger q

s interchange station between a given pair of O-D stations

$T_{l',o}^{\mathrm{ARR}}$ time of passengers' **arr**ival ($\mathtt{ARR}$) on a platform for line l' at origin station o

$T_{l'',d}^{\mathrm{ARR}}$ time of passengers' **arr**ival ($\mathtt{ARR}$) on a platform for line l'' at destination station d

$T_{l',s}^{\mathrm{ARR}}$ time of passengers' **arr**ival ($\mathtt{ARR}$) on a platform for line l' at interchange station s

$T_{l',o}^{\mathrm{DEP}}$ time of passengers' **dep**arture ($\mathtt{DEP}$) from a platform for line l' at origin station o

$T_{l'',s}^{\mathrm{DEP}}$ time of passengers' **dep**arture ($\mathtt{DEP}$) from a platform for line l'' at interchange station s

$T_{l',o}^{\mathrm{dep}}$ time of **dep**arture ($\mathtt{dep}$) of a l'-train from its platform at origin station o

$T_{l'',s}^{\mathrm{dep}}$ time of **dep**arture ($\mathtt{dep}$) of l''-train from its platform at interchange station s

T^{ENT} time-stamp at which passengers pass through a ticket gate to enter an origin station o, referred to as '**ent**ry time' ($\mathtt{ENT}$)

T_q^{ENT} **ent**ry ($\mathtt{ENT}$) time of passenger q

T^{EXT} time-stamp at which passengers pass through a ticket gate to exit from an destination station d, referred to as '**exit** time' ($\mathtt{EXT}$)

T_q^{EXT} **exit** ($\mathtt{EXT}$) time of passenger q

T_{qr}^{EXT} **exit** ($\mathtt{EXT}$) time of passenger q, given that he/she chooses route r to make a journey

$\mathbf{t}_{qr}$ vector that contains all travel time variables for passenger q choosing route r

$t_h(\phi, \psi)$ journey time of any passenger travelling by route h, given that he/she boards the ϕ-th arriving train at origin station (and, if h involves interchange, the ψ-th arriving train at an interchange station)

(*Continued*)

(*Continued*)

$t_{l',o}^{\texttt{ACC}}$ **acc**ess (ACC) walking time from a gateline to a platform for line l' at origin station o

$t_{qr}^{\texttt{AEI}}$ passenger q's total walking time by using route r, including his/her **a**ccess and **e**gress, and the walk for **i**nterchange (AEI)

$t_{l'',d}^{\texttt{EGR}}$ **egr**ess (EGR) time from a platform for line l'' to a gateline at a destination station d

$t_{l',[o,s]}^{\texttt{OBT}}$ **o**n-**b**oard **t**ravel (OBT) time in a train of line l' running from origin station o to interchange station s

$t_{l'',[s,d]}^{\texttt{OBT}}$ **o**n-**b**oard **t**ravel (OBT) time in a train of line l'' running from interchange station s to destination station d

$t_{qr}^{\texttt{OBT}}$ passenger q's total **o**n-**b**oard **t**ravel (OBT) time by using r

$t_{h}^{\texttt{REF}}$ expected average journey time of travelling by route h, serving as a **ref**erence (REF) value for interpreting estimates from a mixture model

$t_{l',[o,s]}^{\texttt{run}}$ **run**ning (run) time of a train of line l', from origin station o to interchange station s

$t_{l'',[s,d]}^{\texttt{run}}$ **run**ning (run) time of a train of line l'', from interchange station s to destination station d

$t_{[l',l''],s}^{\texttt{TIC}}$ walking time to **t**ransfer from a platform for line l' to another for line l'' at **i**nter**c**hange (TIC) station s

$t_{qr}^{\texttt{TIC}}$ passenger q's walking time to **t**ransfer between platforms at an **i**nter**c**hange (TIC) station on route r

$t_{l',o}^{\texttt{WFD}}$ **w**aiting time to board a train of line l' **f**or **d**eparture (WFD) from the l'- platform at origin station o

$t_{qr}^{\texttt{WFD}}$ passenger q's **w**aiting time to board a train **f**or **d**eparture (WFD) from an origin station by using route r

$t_{l'',s}^{\texttt{WIC}}$ **w**aiting time to board a train of line l'' for departure from the l''-platform at **i**nter**c**hange (WIC) station s

$t_{qr}^{\texttt{WIC}}$ passenger q's **w**aiting time to board a train for departure from an **i**nter**c**hange (WIC) station on route r

$t_{qr}^{\texttt{WLK}}$ passenger q's total **wal**k**k**ing (WLK) time of both his/her access at an origin station and egress at a destination station by using route r

(*Continued*)

(*Continued.*)

$t_{qr}^{\texttt{WTT}}$ passenger q's total **w**ai**t**ing **t**ime on route r, including his/her waiting times at both the origin and interchange stations

$t_{*}(\cdot)$ Student's t-value with certain degrees of freedom and a given probability level $*$

U_{qr} utility that passenger q perceives he/she may gain from choosing route r to make a journey

$\mathcal{U}(0,1)$ standard uniform distribution

u underground station (representing station of origin o, destination d or interchange s)

$u(\cdot)$ function that indicates whether a station is an interchange or the destination of a given pair of O-D stations

V_{qr} deterministic (or observable) portion of utility U_{qr}

v_h (or v_r) indicator (or dummy variable) that equals one if route h (or r) is a direct service, and zero if it is an indirect service

$\mathbf{x}_{uh}$ vector that contains reciprocals of distances for each type of pathways at station u on route h;

$x_{uh}^{\texttt{DNS}}$ total run of **s**taircases used for going **d**ow**n** ($\texttt{DNS}$) to lower levels at station u on route h

$x_{uh}^{\texttt{ESC}}$ total run of **esc**alators/lifts ($\texttt{ESC}$) at station u on route h

$x_{uh}^{\texttt{PSG}}$ total length of level/ramp **p**a**s**sa**g**eways ($\texttt{PSG}$) at station u on route h

$x_{uh}^{\texttt{UPS}}$ total run of **s**taircases used for going to **up**per ($\texttt{UPS}$) levels at station u on route h

$\mathbf{y}$ vector that contains passengers' walking/moving speeds on each type of pathways

$y^{\texttt{DNS}}$ walking speed of going **d**ow**ns**tairs ($\texttt{DNS}$)

$y^{\texttt{ESC}}$ **esc**alators/lifts ($\texttt{ESC}$) speed

$y^{\texttt{PSG}}$ walking speed on level/ramp **p**a**s**sa**g**eways ($\texttt{PSG}$)

$y^{\texttt{UPS}}$ walking speed of going **ups**tairs ($\texttt{UPS}$)

E.2 Symbols based on the Greek alphabet

(Listed in alphabetical order)

α_{qr} binary indicator that equals one if passenger q actually chose route r, and zero otherwise

$\boldsymbol{\beta}$ vector that contains all coefficients, each being associated with a travel time variable

$\beta^{\mathtt{AEI}}$ coefficient of passenger q's total walking time by using route r, including his/her **a**ccess and **e**gress, and the walk for **i**nterchange ($\mathtt{AEI}$) between a given pair of O-D

$\beta^{\mathtt{I/C}}$ coefficient of a dummy variable for interchange/non-interchange

$\beta^{\mathtt{OBT}}$ coefficient of passengers' total **o**n-**b**oard **t**ravel ($\mathtt{OBT}$) time between a given pair of O-D

$\beta^{\mathtt{TIC}}$ coefficient of passengers' walking time to **t**ransfer between platforms at **i**nter**c**hange ($\mathtt{TIC}$) stations

$\beta^{\mathtt{WFD}}$ coefficient of passengers' **w**aiting time to board a train **f**or **d**eparture ($\mathtt{WFD}$) from an origin station

$\beta^{\mathtt{WIC}}$ coefficient of passengers' **w**aiting time to board a train for departure from an **i**nter**c**hange ($\mathtt{WIC}$) station

$\beta^{\mathtt{WLK}}$ coefficient of passengers' total **w**a**lk**ing ($\mathtt{WLK}$) time of both access (at an origin station) and egress (at a destination station)

$\beta^{\mathtt{WTT}}$ coefficient of passenger q's total **w**ai**t**ing **t**ime at both the origin and interchange stations on route r

$\hat{\gamma}_{qr}$ estimate of the location parameter for journey time distribution of passenger q making a single journey by route r

$\Delta^{\mathtt{SIM}}$ **sim**ulated ($\mathtt{SIM}$) data set of passengers' journey times, which is generated from a mixture model (being estimated)

Δ set of all journey time observations for a given O-D pair

$\Delta_{5\%}$ set of individual journey time observations, containing a sample of 5% Oyster card data (from 6th February (Sunday) to 5th March (Saturday) in 2011)

$\Delta_{N_R=2}$ set of posterior probabilities of passengers' route choice for selected O-D pairs, each of which involves two alternative routes ($N_R = 2$)

(*Continued*)

(*Continued*)

$\Delta_{N_R \le 3}$	set of posterior probabilities of passengers' route choice for selected O-D pairs, any one of which involves no more than three alternative routes ($N_R \le 3$)
$\Delta_{N_R \le 4}$	set of posterior probabilities of passengers' route choices for selected O-D pairs, any one of which involves no more than four alternative routes ($N_R \le 4$)
$\Delta^{\texttt{DES}}$	set of **des**ired ($\texttt{DES}$) data, which includes both of passengers' route choices and their journey times
$\Delta_r^{\texttt{INF}}$	set of journey time data of passengers who chose route r, based on *effective* **inf**erence ($\texttt{INF}$) from a mixture model
$\Delta_r^{\texttt{INF}_0}$	set of journey time data of passengers who chose route r, based on *naive* **inf**erence ($\texttt{INF}_0$) from a mixture model
$\Delta_r^{\texttt{KMS}}$	set of journey time observations, which is produced by K-means ($\texttt{KMS}$) clustering and labelled r
$\boldsymbol{\delta}_q$	elementary event that passenger q spent a journey time of $\delta_q^{\texttt{OBS}}$ travelling between a given pair of O-D stations
$\boldsymbol{\delta}_{qr}$	elementary event that the expected journey time of passenger q is $\delta_{qr}^{\texttt{EXP}}$, given that he/she chooses route r and his/her entry time is $T_q^{\texttt{ENT}}$
δ	journey time of travelling between a given pair of O-D stations
δ_q	journey time of passenger q travelling between a given pair of O-D stations
δ_{qr}	journey time of passenger q making a single journey by route r
δ_r	journey time of route r between a given pair of O-D stations
$\delta_q^{\texttt{OBS}}$	journey time **obs**ervation ($\texttt{OBS}$) of passenger q
$\delta_{qr}^{\texttt{EXP}}$	**exp**ected ($\texttt{EXP}$) journey time of passenger q using route r, given an observation of his/her entry time $T_q^{\texttt{ENT}}$
$\delta_{\mathscr{q}}^{\texttt{SIM}}$	**sim**ulated journey time, with subscript $\mathscr{q}$ being its index; $\mathscr{q} \in \mathbb{N}_{\ge 1}$
ε_{qr}	error term in utility U_{qr}
$\zeta(\cdot)$	assignment function used for *naive* inference of each passenger's route choice, based on a mixture model
$\eta_r^{\texttt{KMS}}$	median (or centroid-value) of set $\Delta_r^{\texttt{KMS}}$

(*Continued*)

(*Continued*)

$\boldsymbol{\Theta}$	vector that contains all parameters of a mixture model
$\hat{\boldsymbol{\Theta}}$	estimate of vector $\boldsymbol{\Theta}$
$\boldsymbol{\theta}_r$	vector of the distribution parameter(s) of $c_r(\delta)$, with $\hat{\boldsymbol{\theta}}_r$ being its estimate
$\hat{\boldsymbol{\theta}}_r$	estimate of parameter vector $\hat{\boldsymbol{\theta}}_r$
$\boldsymbol{\vartheta}_{qr}$	vector of a parameter (or parameters) for probability distribution of passenger q making a single journey by route r
$\hat{\boldsymbol{\vartheta}}_{qr}$	estimate of parameter vector $\boldsymbol{\vartheta}_{qr}$
$\kappa(\cdot)$	objective function to be minimised for *K*-means clustering
$\boldsymbol{\Lambda}$	vector of standard uniform variables used for the *effective* inference given data set of observed journey times, with each being associated with one of the observations
Λ_q	random variable for passenger q, which follows the standard uniform variable; $\Lambda_q \sim \boldsymbol{\mathcal{U}}(0,1)$
λ_q	generated (real-valued) number of standard uniform variable Λ_q
$\boldsymbol{\mu}$	vector that contains all subpopulation means μ_r
μ_r	mean of journey time distribution of route r (also referred to as subpopulation mean of Q_r)
$\hat{\mu}_r$	estimate of subpopulation mean μ_r
$\xi(\cdot)$	assignment function used for *effective* inference of each passenger's route choice, based on a mixture model
$\boldsymbol{\Pi}_{\Delta}^{\texttt{MIX}}$	matrix (of size $n \times N_R$) that enumerates all posterior probabilities of passengers' route choices, estimated from a **mix**ture (`MIX`) model on data set Δ
$\boldsymbol{\Pi}_{\Delta_{5\%}}^{\texttt{UMM}}$	matrix (of size $n \times N_R$) that enumerates all **u**pdated posterior probabilities of passengers' route choices, based on estimates from a **m**ixture **m**odel (`UMM`) on data set $\Delta_{5\%}$
$\pi(\cdot)$	posterior probability (density) function for passengers' route choices given their journey times
$\pi_{qr}^{\texttt{MIX}}$	posterior probability that passenger q chose route r (given his/her journey time $\delta_q^{\texttt{OBS}}$), estimated from a **mix**ture (`MIX`) model

(*Continued*)

(*Continued*)

$\pi_{qr}^{\mathtt{UMM}}$	**u**pdated posterior probability of passenger q choosing route r, based on the estimate from a **m**ixture **m**odel ($\mathtt{UMM}$)
$\boldsymbol{\sigma}$	vector that contains all subpopulation standard deviations σ_r
$\hat{\sigma}_h$	estimate of a sample standard deviation of journey time of route h (given that each of $t_h(\phi,\psi)$ $\forall\phi,\psi$ is treated an observation)
$\hat{\sigma}_h^{\mathtt{SEM}}$	estimate of a standard error of the mean journey time of route h (given that each of $t_h(\phi,\psi)$ $\forall\phi,\psi$ is treated as a sample mean)
σ_r	standard deviation of journey time distribution of route r (also referred to as subpopulation standard deviation of Q_r), with $\hat{\sigma}_r$ being its estimate
$\hat{\sigma}_r$	estimate of subpopulation standard deviation σ_r
$\sigma_r^{\mathtt{KMS}}$	standard deviation of set $\Delta_r^{\mathtt{KMS}}$
$\hat{\varsigma}_r$	estimate of the scale parameter for journey time distribution of passenger q making a single journey by route r
$\tau_{hu}^{\mathtt{WLK}}$	expected **walk**ing ($\mathtt{WLK}$) time at station u along route h
Φ_q	set of all possible route choices of passenger q
$\Phi^{(\delta)}$	set of all possible route choices of passengers travelling between a given pair of O-D stations, given their actual journey times
$\Phi_q^{(\delta)}$	set of all possible route choices of passenger q, given his/her actual journey time $\delta_q^{\mathtt{OBS}}$
ϕ	number of attempts that passengers make to successfully board a train at origin station o
ψ	number of attempts that passengers make to successfully board a train at interchange station s
$\boldsymbol{\omega}$	vector that contains all mixture weights of a mixture model
$\hat{\boldsymbol{\omega}}$	estimate of vector $\boldsymbol{\omega}$
ω_r	mixture weight of journey time distribution of route r
$\hat{\omega}_r$	estimate of mixture weight ω_r
$\omega_r^{\mathtt{INF}_0}$	proportion of passengers using route r, based on *naive* **inf**erence ($\mathtt{INF}_0$) from a mixture model of journey time

(*Continued*)

(*Continued.*)

$\omega_r^{\mathtt{INF}}$ proportion of passengers using route r, based on *effective* **inf**erence (`INF`) from a mixture model of journey time

$\omega_r^{\mathtt{KMS}}$ proportion of sub-dataset $\Delta_r^{\mathtt{KMS}}$ in data set Δ

$\omega_r^{\mathtt{ROD}}$ percentage of respondents who chose route r, according to the **R**olling **O**rigin and **D**estination (`ROD`) Survey data

$\omega_r^{\mathtt{UPD}}$ proportion of passenger using route r, based on *effective* inference from **upd**ated (`UPD`) route-choice probabilities