Imperial College London
Department of Civil and Environmental Engineering

Decomposing journey time variance on urban rail transit systems

Ramandeep Singh

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Civil and Environmental Engineering and
the Diploma of Imperial College, July 2019

Declaration of originality

I confirm that this thesis is my own original work. Any material from the published or unpublished work of others is appropriately referenced.

Copyright declaration

Abstract

In this thesis, automated fare collection (AFC) data are used to analyse and quantify transit journey time service quality on the London Underground metro system. The thesis comprises of three main research areas. The first part focuses on characterising passenger journey time variance through the generation of empirical probability distributions of journey times under regular and incident-affected operating conditions. The distributions are parametrically defined, and practical passenger-oriented performance metrics are proposed based on the moments of the distributions. The second area of research involves decomposing total passenger journey times recorded by the AFC data into sub-components that distinguish between the walking and in-vehicle phases of a passenger journey. To achieve this, a Bayesian assignment algorithm is proposed to allocate individual passengers to individual trains. Total journey times are then decomposed into the constituent components of access, on-train, and egress times. In the third area of research, the degree to which different service supply and demand factors influence journey times is analysed. Semiparametric regression methods are applied to quantify the effect of physical station and route characteristics, operational service supply factors, and passenger demand levels for each journey time component. To quantify the effect of individual passenger characteristics on journey times, passenger-level heterogeneity within each journey time component is analysed. As an extension to the access time model, the influence of train headways on passenger wait times at the origin station is also derived.

The main outputs of the thesis are the quantification of journey time performance, and the identification of the key service supply and demand factors that impact journey times. The results can be directly applied by operators to guide where potential interventions should be made in order to improve the reliability of journey times for urban rail transit networks.

Acknowledgements

The research presented in this thesis has been carried out under the supervision of Professor Daniel J. Graham, Research Director of the Transport Strategy Centre (TSC), and Mr Richard Anderson, Managing Director of the TSC at Imperial College London. I am deeply grateful to both Professor Graham and Richard for their continued guidance, advice, and support throughout the duration of this research. I further express special thanks to Dr. George Goldberg, who performed vital data cleaning of the raw Oyster card files at the beginning of the project, and Dr. Daniel Hörcher, for valuable feedback on the research and on the write up of the thesis.

Transport for London (TfL) and the TSC at Imperial College jointly funded the research, and these sources of financial support are gratefully acknowledged. Particular thanks goes to Mr Andrew Hyman of TfL. I am thankful for the support from Alexandra Williams and Sarah Willis from the Department of Civil and Environmental Engineering. Special thanks to Fionnuala Ni Dhonnabhain for being a constant source of support and for providing a listening ear at any time of the day over the last few years.

Many thanks to my colleagues in the TSC, particularly my fellow PhD students and the post-docs: Anupriya, Ana, Chuyu, Csaba, Farah, Jose, Keita, Laila, Liang, Nan, Praj, Saeed, and Shane. I must thank the fantastic friends that I was fortunate to make during my time here, including the researchers within the Centre for Transport Studies and Sophie, Nils, Jean Paul, Filip, Aurelien, Laura, Luis, Alex, Stas, Acile, and Magnus. Final thanks go to my family: my grandmother, parents, Harpreet, and Darash.

Contents

1 Introduction **1**

1.1 Background . . . 1

1.2 Motivation and objectives . . . 4

1.3 Contributions . . . 5

1.4 Thesis outline . . . 8

2 Literature review **11**

2.1 Definition of journey time variance . . . 12

2.2 Definition of journey time components . . . 15

2.3 Journey time performance metrics . . . 16

2.3.1 Operations-based performance metrics . . . 17

2.3.2 Passenger-oriented performance metrics . . . 24

2.3.3 London Underground performance metrics . . . 32

2.4 Journey time distributions . . . 36

2.4.1 Underlying principles of journey time distributional form . . . 37

2.4.2 Unimodal distribution studies . . . 38

2.4.3 The effect of temporal and spatial aggregation . . . 42
2.4.4 Fitting and evaluation methods for unimodal distributions . . . 43
2.4.5 Multimodal distribution studies . . . 44
2.5 Factors that affect journey times . . . 46
2.5.1 Theoretical regression model framework . . . 47
2.5.2 Empirical studies on total journey times . . . 49
2.5.3 Train running times . . . 53
2.5.4 Train dwell times . . . 56
2.5.5 Platform wait times . . . 60
2.5.6 Summary of factors that influence urban rail journey times . . . 64
2.5.7 Evaluation of regression methods used in empirical work . . . 66
2.5.8 A note on machine learning methods . . . 67
2.6 Summary of literature review and research contributions . . . 69

3 Data 72

3.1 Definition of London Underground study area . . . 73
3.2 Passenger trip data . . . 73
3.2.1 Data summary . . . 73
3.2.2 Descriptive statistics . . . 76
3.2.3 Limitations . . . 78
3.3 Train movement data . . . 80
3.3.1 Data summary . . . 80

3.3.2 Descriptive statistics . . . 82
3.3.3 Limitations . . . 87
3.4 Incident data . . . 88
3.4.1 Data summary . . . 88
3.4.2 Descriptive statistics . . . 89
3.4.3 Limitations . . . 91
3.5 Additional network data . . . 93
3.5.1 Rolling Origin Destination survey data . . . 93
3.5.2 Physical network and station characteristics . . . 94

4 Characterising journey time distributions **96**
4.1 Introduction . . . 96
4.2 Study area and data . . . 98
4.3 Methods . . . 100
4.3.1 Operating scenarios . . . 101
4.3.2 Unimodal distribution fitting . . . 104
4.3.3 Evaluation criteria for unimodal distributions . . . 106
4.3.4 Multimodal distribution fitting . . . 109
4.3.5 Journey time performance metrics . . . 112
4.4 Results . . . 120
4.4.1 Distributional form . . . 120
4.4.2 Journey time performance metrics . . . 127

4.5 Conclusions 137

5 Passenger to train assignment 139

5.1 Introduction 139

5.2 Literature review 141

5.2.1 Prior to automated data availability 142

5.2.2 Automated data methods 142

5.2.3 Limitations of reviewed methods and thesis contributions 146

5.3 Study area and data 148

5.4 Methods 150

5.4.1 Proposed assignment algorithm 150

5.4.2 Comparison of relative performance of proposed assignment algorithm . . 159

5.5 Results 160

5.6 Conclusions 164

6 Decomposing journey times via semiparametric regression 165

6.1 Introduction 165

6.2 Study area and data 168

6.3 Methods 169

6.3.1 General regression model framework 170

6.3.2 Dependent variables 172

6.3.3 Model covariates 173

6.3.4 Supply side covariates . . . 174
6.3.5 Demand covariates . . . 177
6.3.6 Model evaluation . . . 179
6.4 Results . . . 183
6.4.1 Data preparation . . . 183
6.4.2 Model form . . . 184
6.4.3 Covariate significance . . . 188
6.4.4 Model residuals . . . 190
6.4.5 Results for model covariates . . . 191
6.4.6 Results for service supply covariates . . . 195
6.4.7 Results for demand covariates . . . 202
6.5 Applications . . . 206
6.5.1 Headways and waiting times . . . 206
6.5.2 Case study: Oxford Circus to Brixton . . . 213
6.6 Conclusions . . . 217

7 Individual passenger-level heterogeneity via semiparametric regression 219

7.1 Introduction . . . 219
7.2 Study area and data . . . 221
7.3 Methods . . . 224
7.3.1 General regression model framework . . . 224
7.3.2 Dependent variables . . . 227

7.3.3 Model covariates . . . 228

7.3.4 Passenger demand covariates . . . 229

7.3.5 Supply side covariates . . . 230

7.3.6 Model evaluation criteria . . . 233

7.4 Results . . . 234

7.4.1 Data preparation . . . 234

7.4.2 Model form . . . 235

7.4.3 Covariate significance . . . 238

7.4.4 Model residuals . . . 239

7.4.5 Network supply and demand covariates . . . 242

7.4.6 Individual passenger effects . . . 243

7.4.7 Station and line effects . . . 248

7.4.8 Case study: Oxford Circus to Brixton . . . 251

7.5 Conclusions . . . 253

8 Conclusions 255

8.1 Summary of thesis objectives . . . 255

8.2 Summary of thesis contributions . . . 256

8.3 Summary of main findings . . . 258

8.4 Potential applications . . . 262

8.5 Future work . . . 264

References **267**

Appendix A Regression studies on bus journey times **284**

Appendix B Map of London Underground **287**

Appendix C List of TfL station and line codes **289**

Appendix D Train assignment method comparison **291**

Appendix E REML optimisation algorithm **296**

Appendix F Chapter 6 supplementary material **300**

Appendix G Chapter 7 supplementary material **307**

List of Tables

2.1 Summary of operator-oriented journey time metrics 18

2.2 Summary of passenger-oriented journey time metrics 26

2.3 London Underground performance metrics 33

2.4 London Underground journey time metric components 35

2.5 London Underground incident cause and disruptions categories 37

2.6 Regression studies on rail journey times 51

2.7 Regression studies on train running times 55

2.8 Empirical studies on train dwell times 59

2.9 Factors that influence journey time variance 65

3.1 Study area - number of stations and unique OD pairs 73

3.2 Trip data set attributes 75

3.3 Passenger data categories 77

3.4 Summary statistics for passenger trips 77

3.5 Train data set attributes 81

3.6 Summary statistics for train movement data 83

3.7 Incident data set attributes 89

3.8 Summary statistics for incident events 90

3.9 Summary statistics for station and network characteristics 95

4.1 Number of passenger trips and incident events by line and direction 100

4.2 Number of journey time distributions by line 102

4.3 Incident band definition 103

4.4 Theoretical probability density functions 105

4.5 Aggregate data - number of empirical distributions per distributional form by line and direction 122

4.6 Aggregate data - number of empirical distributions per distributional form by time period 122

4.7 Incident data - number of empirical distributions per distributional form by line and direction 125

4.8 Incident data - number of empirical distributions per distributional form by time period 125

4.9 Aggregate data - journey time performance measures by line, direction, and time period 128

4.10 Incident data - journey time performance measures by line, direction, and time period 133

5.1 Train headways by time period 153

5.2 Summary of unambiguous trips vs ambiguous trips by line and direction 160

5.3 Egress times of unambiguous trips by time period 161

5.4 Summary statistics of egress times and headways of unambiguous trips in off-peak 161

6.1 Summary statistics of dependent variables 173

6.2 Summary of regression covariates by model 174

6.3 Summary statistics of dynamic service supply covariates 176

6.4 Summary statistics of static service supply covariates 177

6.5 Definition of passenger categories 179

6.6 Summary statistics of passenger demand covariates 179

6.7 Discarded model outliers 184

6.8 Model goodness-of-fit 187

6.9 Form and significance of model covariates 188

6.10 Covariate elasticities - dynamic covariates 193

6.11 Covariate elasticities and predicted values - static covariates 194

6.12 Passenger effects (log-mins) 204

6.13 Derivative of access times vs headway model goodness-of-fit statistics 209

6.14 Marginal platform wait times by time period 210

6.15 Oxford Circus to Brixton mean predicted values and elasticities 213

6.16 Oxford Circus to Brixton station and OD route-specific intercepts 214

7.1 Number of trips per passenger 222

7.2 Summary statistics of dependent variables 228

7.3 Summary of regression covariates by model 229

7.4 Number of passengers per time period 229

7.5 Summary statistics of demand covariates 230

7.6 Summary statistics of service supply covariates 232

7.7 Model form 1 goodness-of-fit statistics - final model form 238

7.8 Model form 2 goodness-of-fit statistics - fixed passenger effects 238

7.9 Model form 3 goodness-of-fit statistics - without passenger effects 239

7.10 Covariate significance - AM peak 240

7.11 Covariate significance - inter-peak 240

7.12 Covariate significance - PM peak 241

7.13 Mean predicted values and covariate elasticities - AM peak 242

7.14 Mean predicted values and covariate elasticities - inter-peak 243

7.15 Mean predicted values and covariate elasticities - PM peak 244

7.16 Individual passenger effects summary statistics 245

7.17 Station effects summary statistics 249

7.18 Line/direction effects summary statistics 250

7.19 Oxford Circus to Brixton mean predicted values and elasticities 251

A.1 Regression studies on bus journey times 284

A.1 Regression studies on bus journey times 285

A.1 Regression studies on bus journey times 286

C.1 TfL codes for lines and directions 289

C.2 TfL codes for stations 290

D.1 Finsbury Park to Oxford Circus data sample summary statistics (mins) 292

D.2 Summary statistics for egress times of unambiguous trips (mins) 293

D.3 Summary statistics for egress times of assigned ambiguous trips (mins) 294

F.1 Model residuals summary statistics (mins) 304

G.1 Model residuals summary statistics (mins) 307

List of Figures

1.1 Example distribution of passenger journey times, Brixton to Oxford Circus, London Underground 3

2.1 Example distribution of journey times, Brixton to Oxford Circus 13

2.2 Journey time components of a typical passenger trip on an urban rail system . . 15

2.3 Wait time metrics for long headway service (Source: Furth et al. (2006)) 23

2.4 Example of different forms of journey time distributions 39

3.1 Study area, London Underground (adapted from Transport for London (2014a)) 74

3.2 Passenger entry times 78

3.3 Train departure times 84

3.4 Train headways by time of day 85

3.5 Distribution of train headways 86

3.6 Total delay due to incident events 91

3.7 Time of incident event 92

4.1 Passenger trips and incident events by time of day 100

4.2 Example of theoretical probability distribution functions - Victoria to Oxford Circus, 8:45-9:00am 106

4.3 Example of multimodal probability distribution function - Brixton to Oxford Circus, 8:00-8:15am, 13 November 2013 112

4.4 Moments of a journey time distribution used in the calculation of performance metrics 116

4.5 Benchmarking format - normalised mean of journey time by time period 129

4.6 Benchmarking format - normalised reliability buffer time by time period 129

4.7 Benchmarking format - normalised standard deviation of journey time by time period 130

4.8 Benchmarking format - I/R ratio normalised mean of journey time by time period 134

4.9 Benchmarking format - I/R ratio normalised reliability buffer time by time period 134

4.10 Benchmarking format - I/R ratio normalised standard deviation of journey time by time period 135

5.1 Distribution of passenger trip entries and train departures by time of day 150

5.2 Journey time components using available train movement data 150

5.3 Example of train assignment - unambiguous and ambiguous trips 152

5.4 Distribution of train headways by time period 154

5.5 Summary of passenger to train assignment process 158

5.6 Distribution of egress times and headways for unambiguous trips in off-peak . . 162

6.1 Distribution of journey times by journey time component 169

6.2 Regression plot of headway 195

6.3 Regression plot of *cv* headway . . . 196

6.4 Regression plot of headway normalised by mean headway . . . 196

6.5 Regression plot of number of platform entries/exits interaction term . . . 199

6.6 Regression plot of train speed . . . 200

6.7 Regression plot of origin platform passenger volume indicator . . . 202

6.8 Regression plot of destination platform passenger volume indicator . . . 203

6.9 Regression plot of passenger line loading indicator . . . 203

6.10 Regression plot of number of passenger trips . . . 205

6.11 Derivative of access time with respect to headway . . . 211

7.1 Distribution of the number of trips per individual passenger . . . 222

7.2 Distribution of regular commuter trips by time of day . . . 223

7.3 Distribution of journey times by journey time component for regular commuters 224

7.4 Distribution of individual passenger effects by time period - access times . . . 245

7.5 Distribution of individual passenger effects by time period - egress times . . . 246

7.6 Distribution of individual passenger effects by time period - total journey times . 246

B.1 Map of London Underground . . . 288

D.1 Distribution of ambiguous egress times by assignment method . . . 294

F.1 Regression plot of dynamically available gates . . . 300

F.2 Regression plot of dynamically available escalators . . . 301

F.3 Regression plot of dynamically available lifts . . . 301

F.4 Regression plot of dynamically available stairs 302

F.5 Regression plot of platform length 302

F.6 Regression plot of station walk distance 303

F.7 Access time model residuals 304

F.8 On train time model residuals 305

F.9 Egress time model residuals 305

F.10 Total journey time model residuals 306

G.1 Access time model residuals 308

G.2 Egress time model residuals 308

G.3 Total journey time model residuals 309

Nomenclature

List of Symbols

α	Model constant in semiparametric regression models
$\bar{x}$	Sample mean of observations of variable x
$\bar{y}$	Mean value of dependent variable in regression models
β	Parameter coefficient estimates for covariates modelled parametrically in semiparametric regression models
ϵ_i	Random error term in semiparametric regression models
$\hat{f}_k(x)$	Kernel density estimator of the observations of variable x
λ	Smoothing parameter in semiparametric regression models
$\mathcal{A}$	Set of ambiguous unassigned passenger trips
$\mathcal{N}(0, \sigma^2)$	Definition that given variable of interest is normally distributed with mean 0 and variance σ^2
$\mathcal{U}$	Set of unambiguously assigned passenger trips
$\mathcal{U}_{op}$	Set of unambiguously assigned passenger trips in off peak times
μ	Population mean of given variable of interest

ν_r	Restricted maximum likelihood of semiparametric regression model
ρ_{qt}	Indicator of passenger volumes for quantity of interest q over time t
σ	Population standard deviation of given variable of interest
σ^2	Population variance of given variable of interest
θ	Set of model parameters
$\widehat{y_i}$	Predicted (or fitted) value of dependent variable in regression models
AIC	Akaike Information Criterion
AT_{dj}	Train arrival time at destination station
AT_{oj}	Train arrival time at origin station
BIC	Bayesian Information Criterion
c	Vector of estimated group-specific fixed effects in semiparametric regression models
cv	Coefficient of variation of given variable of interest
$D_{explained}$	Proportion of deviance explained by semiparametric regression model under consideration
DT_{dj}	Train departure time at destination station
DT_{oj}	Train departure time at origin station
e	Elasticity of journey time with respect to model covariate
$E[x]$	Expected value of variable x
$f(x)$	General probability density function of variable x

f_k	Smooth functions based on penalised thin plate regression splines in semiparametric regression models
FF	Free flow time
$g(\cdot)$	Link function (identity function) applied to dependent variable of journey times in semiparametric regression models
h	Observed train headway (mins)
H_0	Null hypothesis
H_1	Alternative hypothesis
h_k	Kernel density smoothing parameter
I	Set of incident-affected trips
$k(\cdot)$	Kernel density function (uniform distribution)
$L(\cdot)$	Likelihood function
$log(L)$	Log-likelihood
$M_{ld,t}$	Weighted value of journey time performance metric for given line-direction over given time
$m_{od,t}$	Unweighted value of journey time performance metric for given origin-destination pair over given time
n	Number of observations of given variable of interest
$N\%x$	Nth percentile value of variable x
p	Number of estimated model parameters
$P(X)$	Probability of event X

$P(X_1 \mid X_2)$	Conditional probability of event X_1 given event X_2
R	Set of regular trips
R^2	Coefficient of determination
$R^2_{adj.}$	Adjusted coefficient of determination
r_{ij}^{eg}	Randomly sampled egress time from the set of unambiguous trips in the off-peak period
RBT	Reliability buffer time
s	Sample standard deviation of observations for given variable of interest
t_i^{entry}	Passenger entry time at origin station
t_i^{exit}	Passenger exit time at destination station
t_{oi}^{walk}	Walk time from the ticket gates to the platform at the origin station
U	Mann-Whitney U test statistic
u	Vector of estimated group-specific random effects in semiparametric regression models
$Var[x]$	Expected variance of variable x
w	Passenger wait time at origin station platform (mins)
x	Nominal variable of interest, typically journey time
X_i	Set of parametric covariates in semiparametric regression models
X_i^*	Set of non-parametric covariates in semiparametric regression models
y_{ij}^{ac}	Passenger access time (mins)

y_{ij}^{eg}	Passenger egress time (mins)
y_i^{jt}	Total passenger journey time (mins)
y_{ij}^{ot}	Passenger on-train time (mins)
y_i	Observed value of dependent variable in regression models, typically observed journey time
y_i^c	Component of journey time for individual trip observation in semiparametric regression models
Z_i^f	Set of fixed categorical factors in semiparametric regression models
Z_i^r	Set of random categorical factors in semiparametric regression models

List of Subscripts and Superscripts

ac	Passenger access time
c	Component of passenger journey time
d	Destination station
eg	Passenger egress time
$entry$	Passenger entry timestamp
$exit$	Passenger exit timestamp
f	Fixed categorical factors
i	Individual observation of given variable of interest, typically journey time of single trip
j	Individual train itinerary

jt	Total passenger journey time
k	Denotes nominal increment in summations, typically increments of value 1
l	Line
ld	Line-direction
M	Weighted journey time performance metric
m	Unweighted journey time performance metric
o	Origin station
od	Origin-destination pair
op	Passenger trips in off-peak time period
ot	Passenger on-train time
r	Random categorical factors
t	Time interval

List of Abbreviations

AFC	Automated fare collection
AVL	Automated vehicle location
C-E	Central line eastbound
C-W	Central line westbound
CoMET	Community of Metros benchmarking group
CuPID	Contract Performance Information Database (Transport for London)

FHWA	Federal Highway Administration (United States)
GAMMs	Generalised Additive Mixed Models
HH:MM:SS	Hours:minutes:seconds
hr	Hours
J-E	Jubilee line eastbound
J-W	Jubilee line westbound
km	Kilometres
KPIs	Key performance indicators
Kurt.	Kurtosis
LCH	Lost Customer Hours (London Underground)
log-mins	Log-transformed minutes
LU	London Underground
m	Metres
Max.	Maximum
Min.	Minimum
mins	Minutes
MLE	Maximum likelihood estimation
NACHs	Nominally Accumulated Customer Hours (London Underground)
NetMIS	Network Management Information System (Transport for London)

NLC	National Location Codes (Transport for London)
No.	Number
OD	Origin-destination pair
OLS	Ordinary least squares
PIRLS	Penalised iteratively reweighted least squares
REML	Restricted maximum likelihood
RODs	Rolling Origin Destination survey (Transport for London)
sec	Seconds
Skew.	Skewness
TCQSM	Transit Capacity and Quality of Service Manual (United States)
TfL	Transport for London
TSC	Transport Strategy Centre
UK	United Kingdom
US	United States
V-N	Victoria line northbound
V-S	Victoria line southbound
W-E	Waterloo and City line eastbound
W-W	Waterloo and City line westbound

Chapter 1

Introduction

1.1 Background

With the introduction of automated fare collection (AFC) on mass transit systems, there now exists access to large volumes of data, with scope for more in-depth analysis of transit service performance than has been previously possible. Conventional methods of quantifying transit service quality rely on indicators that capture the performance of the network from an operational perspective. For urban metro systems, the indicators are predominately evaluated using data on train movements from the network signalling system. Service performance is then measured relative to predetermined target levels of service as per train schedules and/or other network operations plans (Anderson et al., 2013; Barron et al., 2013; Kittelson & Associates Inc et al., 2013; Furth et al., 2006; Kittelson & Associates Inc et al., 2002). The resulting indicators typically measure the average or aggregate performance of network operations, however, these indicators do not represent the passenger experience.

Using passenger AFC data as an input, there has been an increasing focus on quantifying journey time performance from the passenger perspective from researchers and operators alike. Measures of the range and variance of journey times on a route have been put forward as providing a better representation of the journey time performance as experienced by passengers (Kittelson & Associates Inc et al., 2013; Taylor, 2013; Lomax et al., 2003; Kittelson & Associates

Inc et al., 2002; Abkowitz et al., 1978).

Literature on travel behaviour demonstrates that the variance of journey times influences the route and mode choice decisions of passengers. In transport demand modelling, the cost of travel for a passenger comprises of monetary costs, costs associated with travel time, and costs associated with factors that affect perceptions of travel (e.g. crowding). The summation of all costs incurred is termed the generalised cost of travel (Button, 2010). If journey times on a route have a high degree of variance, and therefore a low degree of predictability, the passenger experience of unreliable service results in a higher generalised cost of travel compared to if the service was more reliable (Noland and Polak, 2002; Bates et al., 2001; Noland and Small, 1995). This can result in a lower willingness to pay for the service, and can potentially lead the passenger to select an alternative route or mode with a lower generalised cost of travel, or in the extreme case, opt to forgo travel altogether.

Following on from this, if all other travel conditions are constant and equal, a passenger is more likely to select the route with a higher degree of journey time reliability. Indeed, recent empirical research on elasticities of passenger demand for rail transit demonstrate that improvements in journey time reliability are associated with higher levels of ridership. For example, studies of rail journeys in the Netherlands and in Britain observe that a 10% improvement (decrease) in journey time reliability is associated with increases (decreases) in passenger patronage in the range of approximately 0.8-3.6% (van Loon et al., 2011; Preston et al., 2009). Therefore, improving the understanding of journey time variance as experienced by passengers is evident; better quantification of journey time variance and the factors that influence it can lead to the development of better policies to improve journey time performance, and can ultimately lead to higher levels of ridership and revenue.

The variance of journey times on a route can be represented by constructing probability density functions of observed journey times of passengers travelling over a given time period. An example of a distribution of journey times is shown in Figure 1.1. The form and moments of the distribution provide a direct representation of the travel conditions that passengers experience, and this can be used to infer service quality from the passenger perspective.

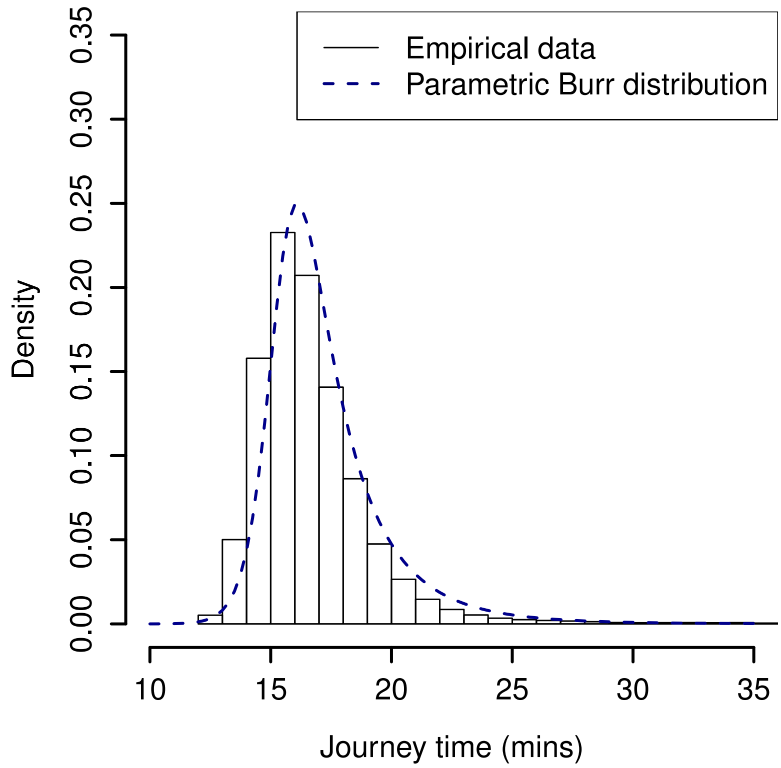

Figure 1.1: Example distribution of passenger journey times, Brixton to Oxford Circus, London Underground

Large volumes of automated data on passenger journeys from the London Underground metro system are used to both develop and verify the journey time analysis methods presented in this thesis. The London Underground is one of the largest and oldest urban metro systems in the world (Transport for London, 2019b). First established in 1863 as the single-line Metropolitan Railway serving six stations, the London Underground system in its current form comprises of a total of 270 functioning stations over twelve lines (Transport for London, 2019a,b). The most recent travel statistics report that on average, 2.8 million trips are made each day on the London Underground (including the Docklands Light Rail system), which corresponds to an approximate 11% transport mode share for daily trips undertaken in the Greater London area (Transport for London, 2018b).

The Oyster smart card system for automated fare collection on the London Underground was first introduced in 2003 (Transport for London, 2019b; Verma, 2010). The Oyster system records passenger entry and exit timestamps and locations, enabling the calculation of passenger journey times between the origin and destination stations. In the thesis, Oyster data covering the time period from October to December 2013 are used. Although statistics are not available for 2013, in 2007, it is estimated that approximately 70% of all trips on the London Underground were made via Oyster card, increasing to over 90% of trips being made by Oyster or other contactless payment modes by the year 2018 (Transport for London, 2018a; Chan, 2007). The automated data records therefore represent a majority of trips on the system, and are a valuable source of data capturing travel patterns of passengers on the London Underground.

1.2 Motivation and objectives

As mentioned in the previous section, transit operators have identified the need to understand and quantify passenger perspectives of journey time reliability. Understanding the passenger experience enables operators to improve the quality of service delivered, and as a result, attract and retain ridership. Through this motivation, the overarching aim of the thesis is to quantify the degree to which journey times vary, and identify the underlying drivers of journey time variance from the passenger perspective. The objectives of the research are split into three broad areas, under which more specific sub-objectives are defined, as follows:

1. Define and quantify the distributional properties of journey times:

   i. Functionally define how journey times are distributed between different origin-destination pairs under different operating conditions.

   ii. Quantify to what extent journey times vary, and which parts of the network perform best and worst.

2. Decompose total journey times into sub-components:

   i Assign individual passengers to individual trains.

   ii Once assigned, undertake empirical decomposition of journey times to distinguish between the walking and in-vehicle phases of a passenger journey.

3. Identify and quantify the factors that influence total journey times and the sub-components of journey time:

   i. Identify the key network service supply and demand factors that influence journey time variance.

   ii. Determine whether individual passenger-specific characteristics influence journey times.

   iii. Quantify the extent to which the aforementioned factors affect total journey times, as well as the different components of journey time.

The thesis objectives are addressed through the analysis of large volumes of passenger trip data, using the London Underground metro system as a representative case study. Combined with additional data sets including train movement data, incident data, and data on network infrastructure and operations, a detailed analysis of passenger journey times is able to be undertaken. First, statistical methods are applied to characterise and quantify degree of journey time variance. Second, the passenger trip and train movement data are merged and a probabilistic assignment algorithm is applied to allocate passengers to trains. This enables the total journey times to be decomposed into their constituent sub-components. Semiparametric regression methods are then applied to identify and quantify the factors that influence each component of passenger journey time. To further illustrate how operators can interpret and apply the results, detailed case studies of representative lines and origin-destination pairs are presented.

1.3 Contributions

The research presented in this thesis contributes to the literature on the analysis of journey time variance for urban metro systems. The specific contributions are summarised in the following sections.

The data used in the thesis are a combination of large volumes of passenger trip data, train movement data, incident data, and additional data on physical network characteristics and operations. The combination of data sets enables a deeper level of analysis of journey times compared to previous work, and as a result, new empirical results on the variance of passenger journey times are presented.

The first part of the analysis involves defining and quantifying the distribution of passenger journey times. The metrics to quantify performance of the London Underground are currently more operator-oriented. To provide improved measurement of journey time performance from the passenger perspective, empirical journey time distributions are generated under different operating conditions. The distributions are parametrically defined to functionally characterise the variance of journey times. From the moments of the distributions, three new metrics to capture journey time performance from the passenger perspective are proposed.

In the second part of the analysis, individual passenger trips are assigned to individual trains. Most of the existing assignment methods reviewed in the literature either require the application of idealised assumptions of passenger travel behaviour and/or require access to vast amounts of manually collected data on walking speeds and station layouts. The probabilistic assignment method proposed in this thesis is based on inputs from automated data sources only, and is therefore more efficient in terms of computational time and cost for operators implement in practice. In existing assignment methods, a potential bias has been identified whereby the assignment is based on the travel characteristics of the fastest passengers using the system. The assignment algorithm presented in this thesis builds on existing methods with the following contributions: first, the algorithm aims to correct for the potential bias in existing methods through the use of egress times of passengers travelling in off-peak periods; second, a new method of assigning probabilities to each feasible train itinerary based on Bayes' Theorem is presented; and third, statistical tests are undertaken to more rigorously verify that the assignment is non-biased.

In the third part of the analysis, semiparametric regression is used to quantify the effect of different network and passenger characteristics on journey times. In the literature, the majority of regression-based studies of journey times adopt linear regression methods. The imposition of a linear relationship results in a simplified model that may omit key characteristics of the relationship between the dependent and independent variables, which can in turn, produce misleading model results. Semiparametric regression methods on the other hand, are able to model flexible relationships between the variables based on the data points themselves, generating parameter estimates with a higher degree of fidelity to the data compared to the conventional linear models. To current knowledge, the application of semiparametric regression methods in the analysis of journey time variance on urban metro systems has not been previously undertaken.

From the assignment of passengers to trains, total journey times are split into three constituent components, namely: access time, on-train time, and egress time. Each component is used as the response variable in the regression models. The decomposition and subsequent analysis of journey times in this manner has not been previously undertaken, and so this is another contribution of the thesis. The decomposition enables elasticities of journey time with respect to key service supply and passenger demand characteristics to be derived, enabling a direct comparison of the relative degree to which the different factors impact journey times.

Two levels of analysis are undertaken using semiparametric regression methods. The focus of the first set of regression models is to quantify the effects of network service supply and passenger demand factors on journey times, including characteristics of stations, origin-destination routes, lines, and generic passenger attributes. The focus of the second set of regression models is to quantify the effects of individual passenger-level heterogeneity on journey times. The analysis of the influence of individual passenger-specific characteristics on transit journey times via a semiparametric mixed model framework has not been previously undertaken.

The regression model of the access time component in the first set of regression models is further extended to derive estimates of marginal passenger platform wait times at the origin station with respect to train headways. To current knowledge, this is the first application of advanced regression techniques to estimate the relationship between wait times and headways while conditioning for other key service supply and passenger demand factors.

1.4 Thesis outline

The remainder of the thesis comprises seven chapters. The first two chapters establish the context for the research contributions presented in the subsequent analysis chapters, beginning with a review of the literature and description of the data sets. The first part of the analysis focuses on total journey times as recorded in the passenger trip data set, and involves defining and quantifying the properties of the journey time distributions. The total journey times are then decomposed into sub-components, via the assignment of individual passenger trips to individual trains. The output of this process forms the basis of the semiparametric regression analyses in the two subsequent chapters. First, the effects of different network supply and demand characteristics on each component of journey time are derived. The models are then extended to analyse the impact of individual passenger travel behaviours within each component of journey time. A detailed outline of the thesis by chapter is summarised in the following sections.

The thesis begins with a review of the literature in Chapter 2 relating to the research areas addressed in Chapters 4, 6, and 7 the thesis. The research relating to the passenger to train assignment methods is reviewed separately in Chapter 5, as the exposition of the methods is improved by reviewing the literature directly prior to the analysis. The literature review chapter is organised into four parts, beginning with a general definition of the concepts of a journey time distribution and journey time variance. This is followed by a discussion of the metrics used to quantify journey time performance from an operator perspective, the passenger perspective, and a summary of the metrics currently adopted by TfL for the London Underground. The third part includes a review of studies on the parametric definition of journey time distributions. Finally, a review of empirical work on the factors affecting rail transit journey times is presented.

The basis of the thesis is the analysis of large volumes of empirical data in the order of millions of observations from the London Underground metro system. As such, it is important to clearly define the properties and limitations of the data, and this is presented in Chapter 3. The data sets included in the analysis, namely passenger trip data, train movement data, incident data, and additional data on network operations and infrastructure, are summarised and the main short-comings of the data in reference to the analyses are discussed in detail.

The next four chapters present the main results of the analyses of journey times. Chapter 4 focuses on the descriptive characterisation of journey times under different operating conditions. Empirical distributions of journey time are constructed at an origin-destination route level on different lines, at varying times of the day. The distributions are generated separately for regular operating conditions and incident-affected conditions to determine the effect of incident events on journey times. Under the different operating conditions and temporal and spatial boundaries, the empirical distributions are parametrically defined, and a series of practical performance metrics based on the moments of the distributions are proposed.

In Chapter 5, the allocation of individual passengers to individual trains is undertaken. The chapter begins with a review of the existing assignment methods, followed by the presentation of the assignment method developed. The passenger trip data are merged with the train movement data, and a probabilistic algorithm based on Bayes' Theorem is proposed to assign passengers to trains. The assignment enables total passenger journey times from entry at the origin station to exit at the destination station to be decomposed into the sub-components of access time, on-train time, and egress time. The sub-components of journey time are then used as inputs for the regression analyses presented in the subsequent chapters.

Assigning the components of journey time as response variables, a semiparametric regression analysis is performed in Chapter 6. The covariates in the models represent different service supply and passenger demand characteristics of the network. Elasticities of journey time are derived for each factor to determine the underlying drivers of variance within each component of journey time. As an extension of the regression results for access times at origin stations, functions of marginal passenger platform wait times with respect to train headways are derived.

In Chapter 7, an extension of the semiparametric regression models from Chapter 6 is presented. The focus of the analysis is to capture and evaluate the effect of individual passenger-level heterogeneity on journey times. Individual passengers are tracked via their unique Oyster card numbers, and a data set of repeat observations for each passenger travelling on the same origin-destination route is extracted. The network supply and demand covariates from the models in Chapter 6 are aggregated and treated as conditioning variables, and individual passenger-specific effects are quantified for each component of journey time at different times of the day.

In the final chapter of the thesis, Chapter 8, a summary of the conclusions from the analysis chapters is presented, along with recommendations for potential future research.

Chapter 2

Literature review

In the following chapter, a review of the literature relating to journey time variance on urban rail systems is presented. The chapter begins with an introduction to the concept of journey time variance, and definition of the components of a typical rail journey. Beyond this, the literature review comprises of three main parts corresponding to the research areas of Chapters 4, 6, and 7 of the thesis. The literature review of the passenger to train assignment methods is included in Chapter 5.

The first part of the literature review presents a discussion of the metrics used to evaluate journey time performance on metro systems, split into three sub-sections: metrics used by metro operators, passenger-oriented performance metrics, and a summary of the metrics used by Transport for London to monitor the performance of journey times on the London Underground. The second part of the literature review covers research that focuses on defining the form of journey time distributions, beginning with a discussion of the fundamental concepts of a journey time distribution, followed by a review of the empirical work in the field. The third part reviews empirical research related to identifying and quantifying the factors that contribute to journey time variance. The chapter concludes with a summary of how the thesis contributes to the existing literature.

2.1 Definition of journey time variance

If journey times of multiple individuals are measured between a given origin and destination, the journey times are not likely to be exactly identical due to differences in service supply, passenger demand, and individual specific travel behaviours. The variance of journey times can be illustrated by generating a journey time distribution. A journey time distribution is simply the probability density function which represents the relative likelihood of observing each journey time. An example is illustrated in Figure 2.1 for weekday trips between Brixton station and Oxford Circus station, northbound on the London Underground Victoria Line.

In Figure 2.1, the probability density function of observed journey times is estimated using the widely adopted kernel density estimation method. The method is non-parametric, i.e. the form of the probability density function is generated from the data points rather than being specified parametrically a priori. The method applies a kernel density function and a smoothing parameter to capture the trade-off between matching the data points and the smoothness of the function (Scott, 2015). The kernel density function applied here is the default normal distribution, though it should be noted that other distributions such as the uniform or triangular distribution are also used in other applications. It should be noted that the majority of the density plots included for illustrative purposes in the thesis are generated via kernel density estimation using the normal distribution, unless otherwise stated.

There are three notable features of a journey time distribution: 1) the minimum journey time, 2) the typical journey time at which most journeys are undertaken, and 3) the spread of journey times. The minimum journey time is the quickest observed travel time between the origin and destination, and represents the fastest passenger travelling in free flow conditions. The US Highway Capacity Manual defines free flow conditions as the travel conditions that arise when vehicle (or passenger) density is zero. The only restrictions under free flow conditions are those related to the geometric and operational environment and personal travel preferences (Transportation Research Board Committee on Highway Capacity and Quality of Service, 2010). The mean or median journey times represent the typical travel conditions on a route, reflecting the typical levels of passenger demand and operational service. The spread of journey times,

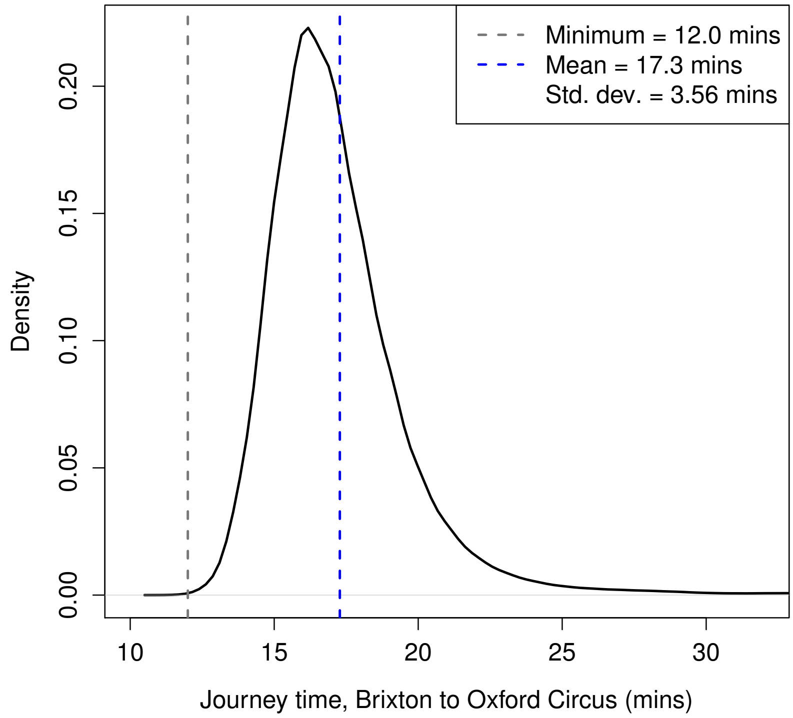

Figure 2.1: Example distribution of journey times, Brixton to Oxford Circus

traditionally measured as the variance or standard deviation of journey times, represents the degree of consistency in travel conditions. A longer spread of journey times reflects more variable travel conditions, whereas a shorter spread of journey times indicates a higher degree of consistency.

In the literature, the terms journey time "reliability" and "variability" are widely used and can generate confusion. The variability of journey times refers the total range of journey times observed on a route. For example, if journey times on route A are more spread across a longer range compared to journey times on route B, it can be said that the variability of journey times on route A is greater than the variability of journey times on route B. In their study on travel conditions on highway networks, Wong and Sussman (1973) propose three sub-categories of journey time variance:

1. Regular condition-dependent variations: predictable variations, such as those arising from peak hour congestion, or seasonal changes.

2. Irregular condition-dependent variations: unpredictable variations from disruptions and incidents that cause delay that propagates through a network.

3. Random variations: unpredictable residual random variance in travel times.

Journey time variability therefore refers to the total variation in journey times from both predictable and unpredictable sources. The definition of journey time reliability relies on the distinction between these two sources of variance. In a detailed review of the literature on transit journey time reliability, Taylor (2013) states that reliability "relates to the ability of a system to perform to expectations under a given set of conditions", highlighting definitions of reliability from transit authorities in the United Kingdom (UK) and United States (US) as follows:

- UK Department of Transport (2007): the degree of reliable service relates to "the variations in journey time that travellers cannot predict".

- US Department of Transportation Federal Highway Administration (2006): reliability is "the consistency or dependability in travel times, as measured from day-to-day and/or across different times of the day".

From these definitions, it can be concluded that journey time reliability is a measure of the degree to which unpredictable variations in journey time occur. For example, if an unexpected incident event occurs during a journey, then journey times on the route may become longer, however, the degree to which the journey times increase is unable to be predicted. Therefore, the service is considered to be unreliable. On the other hand, if journey times are observed to be consistently longer during congested peak periods on a given route, the journey time of the route can be considered as being "reliably late", as the degree of the regular fluctuation in journey times is predictable.

In this research, quantifying both the predictable and unpredictable sources of journey time variance is of interest. The concepts of journey time variability and reliability are therefore both applicable, and are subsequently referenced throughout the thesis.

2.2 Definition of journey time components

To be consistent with the definition used by Transport for London (TfL), in the thesis, the term "trip" refers to a one-way movement from an origin point to a destination point (Transport for London, 2018b). Each passenger trip on an urban rail transit system comprises of distinct components, and in the following literature review, frequent references are made to each component of journey time. For clarity, the components of a passenger trip are defined as illustrated in Figure 2.2.

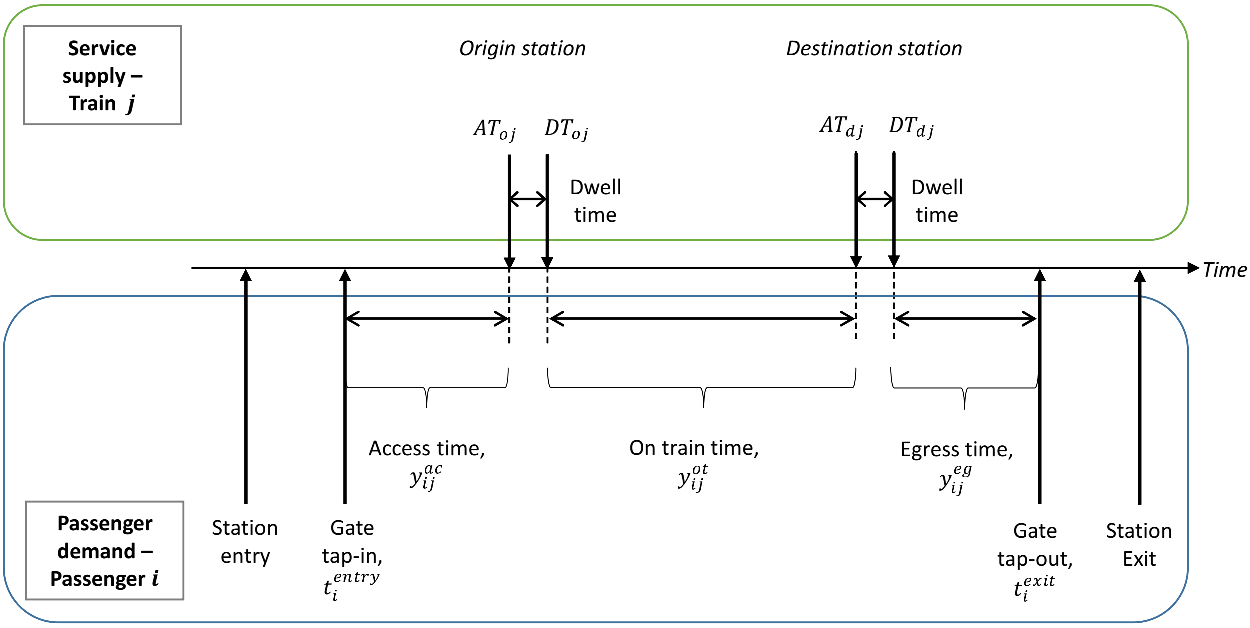

Figure 2.2: Journey time components of a typical passenger trip on an urban rail system

In the thesis, the term "journey time" is defined as the total journey time of a passenger from tap-in at the origin station to tap-out at the destination station on an urban rail system. As shown in Figure 2.2, the total passenger journey time comprises of the following components:

1. Access time - the time taken from passenger tap-in at the origin station until the passenger boards the train. This component includes the time taken for the passenger to walk to the platform as well as wait time at the platform prior to train boarding.

2. Dwell time - represents the time that a train waits at each station platform to allow passengers to alight and board.

3. On-train time (also referred to as the train run time) - the running time of the train as it travels from the origin to the destination station.

4. Egress time - the time taken from when the passenger alights the train at the destination to when the passenger taps-out at the exit.

It should be noted that there is another component of journey time termed "interchange time", which captures the time taken for the passenger to travel from one platform to another within a station in cases where the passenger is required to make a transfer from one line to another line. In this research, the analysis is undertaken on single-line journeys only, and so detailed discussions of this component of journey time are generally excluded in the literature review.

2.3 Journey time performance metrics

Quantifying the degree of variability in journey times is a key input in monitoring the level of service quality of a transit system. There are two main stakeholders of a transit service: the operator and the passenger. In order to achieve a reliable and therefore high quality of service, the needs of both the operator and passenger must be considered.

From a business operations perspective, the fundamental objective for the operator is to minimise costs and maximise social benefits and/or revenue. To minimise costs, the operator develops a timetable for the service that minimises the total travel time and/or the total energy usage of the system. It is therefore important to monitor how well the actual train operations adhere to the optimised schedules or timetables. As a result, the metrics that operators use to monitor service quality are those that are based on measuring the degree to which actual operations adhere to the service plans. The input data used to calculate the operator-oriented metrics have traditionally been more readily available. For rail systems, this includes data on train movements from signalling systems and data on the scheduled services from the operations plans. Arising from data availability and the operator imperative to monitor costs, the operator-oriented metrics have been the dominant form of metric to measure transit service quality.

If journey times on a route have a high degree of variability and a low degree of predictability, research has shown that the passenger experience of unreliable service can lead a passenger to rank the route as having a higher generalised cost of travel compared to if the service was more reliable (Bates et al., 2001; Noland and Polak, 2002). Consequently, the passenger could likely choose a less costly option of travel for future trips, and this could lead to a loss of ridership for the operator. It is therefore in the operator's interest to consider journey time performance as experienced by the passenger. With the more widespread availability of automated fare collection data in recent times, both academic researchers and operators have begun to develop metrics to measure the performance of journey times from a passenger perspective.

In the following sections, a summary of the most commonly cited and used journey time performance metrics are presented, beginning with a review of the operator-oriented metrics, followed by the passenger-oriented metrics, and finally a summary of the metrics used by the London Underground is presented.

2.3.1 Operations-based performance metrics

The US Federal Transit Administration Transit Capacity and Quality of Service Manual (TC-QSM) is one of the most widely cited manuals for the specification of performance metrics used by transport operators (Kittelson & Associates Inc et al., 2013). The manual identifies three categories of metrics related to journey time performance used by rail operators: schedule adherence, regularity of headways, and platform wait times. The metrics under each of these categories are discussed in the following sections. The mathematical expressions of the metrics are summarised in Table 2.1.

Schedule adherence

The schedule adherence metrics are more applicable to low frequency services, to provide a measure of the degree to which actual train running times deviate from the scheduled running

Table 2.1: Summary of operator-oriented journey time metrics

Metric	Definition	Source
Schedule adherence metrics		
On time performance	% "on time" trains Where "on time" trains are those within scheduled AT or $DT+$ (2 to 5 mins)	Kittelson & Associates Inc et al. (2013); Anderson et al. (2013)
Headway regularity metrics		
Headway adherence	$cv_H = \frac{s_H}{H}$	Kittelson & Associates Inc et al. (2013)
Service regularity metric	% "regularity" Where "regular" headways are those within scheduled $H \pm 50\%H$	Kittelson & Associates Inc et al. (2002)
Headway ratio	$\frac{actual\ H}{scheduled\ H} \times 100\%$	
Headway deviation	$actual\ H - scheduled\ H$	
Headway regularity index	$1 - \frac{2\sum(H_r - \bar{(}H)r)}{n^2\bar{H}}$	Henderson et al. (1991)
Platform wait time metrics		
Short headways - expected wait time	$E[w] = 0.5E[H](1 + cv_H^2)$	Furth and Muller (2006)
Short headways - excess wait time	$0.5E[H_{actual}](1 + cv_{H\ actual}^2) - 0.5E[H_{schedule}](1 + cv_{H\ schedule}^2)$	
Potential wait time	$95\%DT - E[DT]$	
Long headways - excess wait time	$actual\ DT - 2\%DT$	
Long headways - excess budgeted time	$95\%DT - 2\%DT$	
Delay metrics		
Number of delays	Number of trains operating $\geq AT$ or $DT+$ (2 to 20 mins)	Anderson et al. (2013)
Service operated between consecutive incidents	Million car km between incidents, where trains operate $\geq AT$ or $DT+$ 5 mins	
Passengers affected ratio	Number of passengers on trains operating $\geq AT$ or $DT+$ 5 mins	

Note: AT is train arrival time, DT is train departure time, H is train headway, cv is the coefficient of variation, s is the sample standard deviation, r is the ranking of headways smallest to largest, and w is wait time.

time. This type of indicator is important for services with longer headways greater than 10 minutes, where services are operated according to a published timetable. For these services, passengers typically consult the timetables to plan their arrival times in order to minimise platform wait times. The earliest metrics of schedule adherence were simplistic, including measures of the average difference between actual and scheduled arrival times, and the standard deviation of mean arrival times, as cited by Abkowitz et al. (1978) in their seminal report on Transit Service Reliability produced for the US Department of Transportation.

The most common metric currently used in practice is the on-time performance metric. The metric, also cited in Abkowitz et al. (1978), is defined as the percentage of train trips at a given point in the system that arrive or depart within a given “on-time” threshold relative to the scheduled arrival/departure times. The metric is typically evaluated at the busiest points in the system or at terminal stops. Higher values of the on-time performance metric correspond to a greater degree of service reliability. The US Transit Capacity and Quality of Service Manual states that if on-time performance is less than 70%, the service is deemed highly unreliable (Kittelson & Associates Inc et al., 2013). The definition of “on-time” can vary between operators. In their assessment of service quality metrics used by operators, Anderson et al. (2013) review 31 different metros around the world and report that on-time trips are calculated at thresholds 2, 3 and 5 minutes later than the scheduled arrival/departure times. The US Transit Capacity and Quality of Service Manual currently suggests an allowable window of train departures from 1 minute early up to 5 minutes late relative to the scheduled service (Kittelson & Associates Inc et al., 2013).

Headway regularity

For higher frequency services with headways shorter than 10 minutes, passengers typically arrive randomly without consulting timetable information and services are operated in accordance to specified headways. A higher degree of regularity or evenness between headways minimises the aggregate wait time across all passengers and leads to more balanced vehicle utilisation rates. As such, measures of headway regularity are adopted to assess the reliability of high frequency services (Kimpel et al., 2000; Kittelson & Associates Inc et al., 2013). The US Transit Capacity and Quality of Service Manual recommends the headway adherence metric, evaluated as the coefficient of variation of observed headways, as the most preferred metric to measure the regularity or degree of bunching of train arrivals (Kittelson & Associates Inc et al., 2013). The metric is calculated as the standard deviation of headways divided by the mean, and the TCQSM states that values greater than 0.75 indicate that most vehicles are bunched together, leading to highly unreliable service.

Aside from the headway adherence metric, other measures of headway regularity are also used by operators, including the service regularity metric, the headway ratio, the headway deviation, and the headway regularity index, as documented in the accompanying guidebook to the TCQSM by Kittelson & Associates Inc et al. (2002). The service regularity metric is defined as the percentage of headways that operate within a given acceptable threshold of deviation from the scheduled headways. The acceptable threshold varies between operators and can be defined as either a percentage of the scheduled headway or a number of minutes; for example, the New York City metro uses a threshold of ±50% from the scheduled headway for high frequency services (Kittelson & Associates Inc et al., 2002). The headway ratio is defined as the ratio between the actual and scheduled headway multiplied by 100 to give a percentage value of the degree of schedule adherence, and the headway deviation is simply defined as the difference between actual and scheduled headways in minutes (Kittelson & Associates Inc et al., 2002). The headway regularity index, initially proposed by Henderson et al. (1991), measures the degree of regularity using Gini's ratio on a normalised scale of 0 to 1. In the computation of Gini's ratio, the y-axis represents cumulative headways while the x-axis indexes cumulative individual services. The axes are computed as proportions of total headway minutes and total services respectively, to enable comparisons between routes with different operating headways. Higher values of the headway regularity index indicate higher degrees of regularity.

Platform wait times

Waiting at the platform for a train to arrive is perceived by passengers to be one of the most inconvenient components of a rail transit journey. In typical rail transit appraisal applications, the value of platform wait time is weighted at least twice as much as the value of in-vehicle time (Wardman, 2004). Transport for London applies a higher weighting of 2.5 for wait time in uncrowded conditions, compared to a base weighting of 1 allocated for in-vehicle time in uncrowded conditions (Transport for London, 2013). As wait time is reported to cause at least double the amount of disutility to passengers as when they are travelling on empty trains, it is considered an important component of service quality. A series of metrics are used to quantify wait times under different conditions.

As mentioned in the previous sections, researchers observe that the pattern of passenger arrivals varies depending on the frequency of train services. Two types of arrival behaviour are noted in the literature: 1) for high frequency services, passengers are observed to arrive randomly; 2) for low frequency services, arrivals are non-random, as passengers tend to consult schedule information to plan their arrivals in order to minimise their wait times. A 10 minute headway is the boundary cited in the US Transit Capacity and Quality of Service Manual for distinguishing between high frequency ($\leq$ 10 mins) and low frequency ($\geq$ 10 mins) services (Kittelson & Associates Inc et al., 2013). Recent research, however, suggests that with the increased access to real-time information, passenger arrival behaviour can switch from random to non-random arrivals at headways as low as 5 minutes (further discussion on the literature on platform wait times is presented in section 2.5).

The most commonly adopted metric by operators is the excess wait time metric, which is defined as the difference between the actual wait time and the scheduled wait time. The value represents the deviation of wait time due to operational irregularities, and therefore quantifies operator performance while also capturing the effect on passengers. The excess wait time metric is defined as one of the key metrics to assess network performance by the London Underground, and is also considered as a "best practice" metric in a survey of 31 international metros on transit service quality (Anderson et al., 2013). The excess wait time metric is evaluated according to the passenger arrival conditions specific to high and low frequency services. In the next sections, the wait time metrics for the two modes of passenger arrival are presented.

For high frequency services under ideal conditions, the expected passenger wait time is defined as half of the scheduled headway. This definition of wait time is conditional on the following assumptions: passengers arrive perfectly randomly, the service is operated according to schedule, and passengers board the first available train (Kittelson & Associates Inc et al., 2013). To account for the fact that headways may vary in operation, the half headway value is adjusted by the coefficient of variation of headways to give an expected passenger wait time as per equation 2.1 (Furth and Muller, 2006).

$$E[w] = 0.5E[h](1 + cv_h^2) \tag{2.1}$$

where $E[w]$ is the expected or average wait time in minutes, $E[h]$ is the average headway in minutes, and cv_h is the coefficient of variation of headways.

Following on from equation 2.1, the excess wait time is then taken as the difference between the actual observed headways and the scheduled headways as given in Table 2.1. Furth and Muller (2006) also introduce the concept of budgeted and potential wait times, which are adopted as key metrics in the US Transit Capacity and Quality of Service Manual. The budgeted wait time is defined as the 95th percentile departure time, which is taken to provide an appropriate level of confidence of on-time arrival at the destination, as the likelihood of arriving late is 1 out of 20 trips (or 5%). The potential wait time is then defined as the difference between the budgeted wait time and the expected or average departure time.

For low frequency services under ideal conditions, passengers arrive at the platform just prior to train departure and board the latest available service that will ensure on-time arrival at the destination (Kittelson & Associates Inc et al., 2013). In practice, passengers factor in a buffer to allow for irregularities in service to improve their likelihood of arriving on time at their destination. The biggest inconvenience for passengers arises in situations where the train departs earlier than scheduled, and the passenger must then wait one full headway until the next service. For this reason, as a rule of thumb, it is assumed that all passengers plan their arrival to coincide with a 2nd percentile train departure time to minimise the likelihood of the train departing early; this is termed as the target arrival time (Furth et al., 2006). In addition to this, passengers also account for the likelihood of the train departing late and include an allowance for a 95th percentile train departure (Furth et al., 2006). As for short headways, this is referred to as the budgeted wait time. The definition of potential wait time for long headways is also equivalent to that for short headways; it is the difference between the budgeted departure time and the expected departure time.

In terms of evaluating the excess wait time for long headways, Furth and Muller (2006) make an allowance for an additional component of wait time termed the synchronisation time. This is included to allow for the possibility that passengers make a conservative decision to arrive earlier than the 2nd percentile target arrival time, to further account for any uncertainties in arriving at the platform before their preferred service departs. For ease of understanding, each component of wait time is visually represented in Figure 2.3, sourced from Furth et al. (2006). From Figure 2.3, two types of excess wait time metrics for long headway services are shown: 1) excess platform wait time, which is the difference between the actual departure time and the target 2nd percentile departure time and 2) the excess budgeted time, which is the difference between the budgeted 95th percentile departure time and the target 2nd percentile departure time.

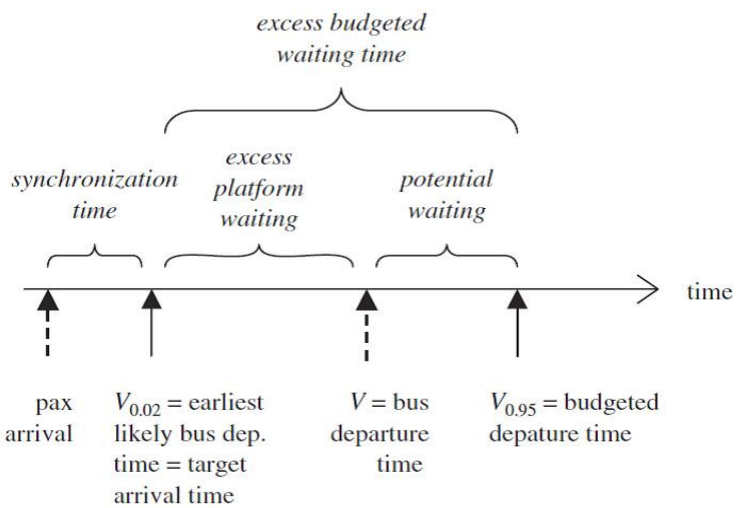

Figure 2.3: Wait time metrics for long headway service (Source: Furth et al. (2006))

Measures of delay

Along with measuring adherence to operating schedules and headways, it is also important for operators to track the occurrence of incident events and their resulting impacts on service. As documented in the US Transit Capacity and Service Quality Manual and a review of delay metrics used by 22 metros across the world by Barron et al. (2013), most operators simply report on the frequency of incidents only. The typical metrics reported are the total number of incidents, and the amount of service operated between consecutive incident events. Different operators define different thresholds for the measurement of incident events. For example, Hong Kong MTR measure the number of incidents that cause delays $\geq$ 2, 5, 8, and 20 minutes (Wood, 2015). In their review of 31 metro operators around the world, Anderson et al. (2013) note that 53% of metros report on the number of incidents that cause delays $\geq$ 2 and 5 minutes. In an effort to unify the measurement of delay across metros for comparative purposes, the number of million car km between incidents that cause delays $\geq$ 5 minutes is reported by the international CoMET and Nova metro benchmarking groups administered by the Imperial College London Transport Strategy Centre.

Individual operators also use additional metrics to quantify the impact of delays on passengers. Hong Kong MTR use the passengers affected ratio, which is defined as the number of passengers whose trips are affected due to delays $\geq$ 5 minutes (Anderson et al., 2013; Wood, 2015). For the London Underground, Transport for London report on the effect of delays on total journey times via the excess journey time metric, and also report on the number of hours of delay incurred by all affected passengers through the Lost Customer Hours metric. These are further discussed in section 2.3.3.

2.3.2 Passenger-oriented performance metrics

The operator-oriented metrics do not necessarily reflect the travel conditions as passengers experience, and in some cases, can provide a misleading picture of service quality from the passenger perspective. The limitations of the operator metrics in terms of representing the passenger experience are as follows:

- The operator metrics are formulated to give an indication of the service measured against a prescribed set of operating standards. If the level of service standards change, the operator metrics can indicate an improvement or reduction in operating service quality, however the base passenger experience remains the same (Abkowitz et al., 1978; Uniman, 2009).

- The data used in the computation of the operator metrics only represent service performance. To capture the passenger perception of reliability, the total journey time of the passenger must be considered, including walk times within stations and waiting times at the platform as well as the in-vehicle times. The operator-based metrics represent data from the in-vehicle portion of a passenger's journey only, and so the complete passenger travel conditions cannot be quantified.

- The operator metrics are designed to reflect the typical operating conditions, and are mostly based on average or aggregated values. For example, most metrics are structured to measure the percentage of trips that operate within predefined standards of good or bad service. This provides an aggregate indication of the percentage of good and bad service operations, however, the total range of operational conditions including extreme values are not reported (Abkowitz et al., 1978; Wilson et al., 1992; Furth et al., 2006; Uniman, 2009).

The passenger-oriented reliability metrics aim to correct for the limitations of the operator metrics. The most important consideration in developing passenger-oriented metrics is to capture the full range of travel conditions between a given origin and destination. With the introduction of automated fare collection data, this is able to be achieved. Disaggregate data on individual passenger journey times are available, and from this data, a distribution of journey times along a route can be generated. The moments of the journey time distributions represent the range and variability of journey times on a route, and this reflects the passenger experience. The most commonly used distributional moments for the development of passenger-oriented metrics include the mean, median, standard deviation, and other measures of the range and spread of journey times. In their review of passenger-oriented journey time metrics, Lomax et al. (2003) classified the metrics into four broad categories: statistical range metrics, buffer time metrics, tardy trips metrics, and probabilistic metrics. Here an updated summary of the most-cited metrics is presented under the same typology of categories. The mathematical forms of the metrics are summarised in Table 2.2.

Table 2.2: Summary of passenger-oriented journey time metrics

Metric	Definition	Source
Statistical range metrics		
Travel time window	$\bar{x} \pm s$	US Transportation Research Board, in NCHRP (1997)
Percent variation	$\frac{s}{\bar{x}} \times 100\%$	California DoT, in Lomax et al. (2003)
Variability Index	$\frac{(Upper\ 95\%x - lower\ 95\%x)_{peak}}{(Upper\ 95\%x - lower\ 95\%x)_{off\ peak}}$	Albert (2000), in Lomax et al. (2003)
Skew-width method	Skew $= \frac{90\%x - 50\%x}{50\%x - 10\%x}$ Width $= \frac{90\%x - 50\%x}{50\%x}$	Van Lint and Van Zuylen (2005)
Buffer time metrics		
Buffer time	$95\%x - \bar{x}$	US FHWA, in Cambridge Systematics Inc. and Texas Transport Institute (2005)
Buffer index	$\frac{95\%x - \bar{x}}{\bar{x}}$	
Reliability factor	$(95\%x - 50\%x)_{od}$	Chan (2007)
Reliability buffer time	$RBT = (95\%x - 50\%x)_{od,t}$	Uniman (2009)
Excess RBT	$(RBT_{Overall} - RBT_{Recurrent})_{od,t}$	
Normalised RBT	$\frac{RBT_{od}}{50\%x_{od}}$	Schil (2012)
Individual RBT	$median(RBT)_{Individual,od,t,day}$	Wood (2015)
Platform-to-platform RBT	$F_{od}^{-1}(0.95) - F_{od}^{-1}(0.5)$ where $F_{od}(x)$ is the empirical platform to platform travel time cumulative density function.	
Tardy trip metrics		
Florida reliability statistic	$100\% - \%$"Unreliable" trips Where "Unreliable" trips have journey times $\geq$ $\bar{x}+$ (5, 10, 15,20% of $\bar{x}$)	Florida DoT, in Lomax et al. (2003)
On time arrival	$100\% - \%$"Unreliable" trips Where "Unreliable" trips have journey times $\geq$ $110\%\bar{x}$ during peak	California DoT, in Lomax et al. (2003)
Misery index	$\frac{(\bar{x}\ 20\%\ longest\ trips) - (\bar{x}\ all\ trips)}{(\bar{x}\ all\ trips)}$	Albert (2000), in Lomax et al. (2003)
Probabilistic metrics		
Probabilistic threshold	$P(\alpha) = P(JT \geq \alpha 50\%x)$ Where α is 1.2	Dutch MoT, in Van Lint et al. (2008)

Note: x is journey time, $\bar{x}$ is mean journey time, s is the sample standard deviation, od denotes origin-destination pair, and t denotes time.

Statistical range metrics

Some of the earliest indicators of journey time reliability, defined initially for road applications, were based on the standard deviation of journey times. The US Transportation Research Board introduced the concept of a travel time window, expressed as the average journey time ± one standard deviation of journey times. This indicator defines a window of approximately 68% of the most observed journey times for a given time period, and can be interpreted as the range of acceptable or reliable journey times (NCHRP, 1997). Extensions to the measure include expanding the range of the window through the addition of an appropriate multiplier to the standard deviation (e.g. a multiplier of 2 would capture approximately 95% of journey times), or adopting the inter-quartile range (75th - 25th percentile journey times) as the range of acceptable journey times (Lomax et al., 2003).

Another early measure of journey time variation was proposed by the California Department of Transportation. The percent variation metric evaluates a coefficient of variation measure for journey times. The metric is defined as the standard deviation of journey times divided by the average journey time for the time period of interest (Lomax et al., 2003). As the standard deviation is normalised by the mean, the metric can be used to compare the journey time variability across routes of different lengths.

In more recent work, using distributional properties other than the standard deviation have been proposed. The variability index is a measure of the spread of peak journey times in relation to off-peak journey times (Lomax et al., 2003). It is computed by calculating the difference between the upper and lower 95% of journey times for peak periods and off-peak periods, and then taking the ratio of the peak spread to the off-peak spread (refer to Table 2.2). The indicator provides a measure of how variable the peak is compared to the off-peak for a given route.

Van Lint and Van Zuylen (2005) extend the concept of statistical range measures to incorporate a measure of the skew of journey time distributions. The distribution of journey times tends towards a more symmetrical shape on more congested routes, while right-skewed distributions indicate relatively lower levels of congestion (see section 2.4 for further discussion). The skew metric is defined as the ratio of the upper 40th percentile (90%-50%) and lower 40th percentile (50%-10%) journey times for a given route over a given time period. Higher values of the skew metric indicate a more left-skewed or congested route. Van Lint and Van Zuylen (2005) also propose a width metric to indicate the degree of spread or reliability of journey times. The width metric is defined as the difference between the 90th percentile journey time and 10th percentile journey time, normalised by the median (50th percentile) for a given route and time period. A higher value of the width metric indicates a greater degree of spread in journey times, which can be interpreted as journey times being less reliable.

Buffer time metrics

In 2005, the US Federal Highway Administration (FHWA) published a guide on reliability metrics for road travel based on work undertaken by Lomax et. al. at the Texas AM Transport Institute from 2000 onward. In this body of work, the concept of expressing reliability in terms of a "buffer time" is introduced. The proposed buffer time metric is defined as the difference between the 95th percentile journey time and the mean journey time for a route (see Table 2.2). The metric can be interpreted as the amount of extra (buffer) time allowed on a typical journey to ensure on-time arrival at a 95% confidence level. A further metric, the buffer index, is the buffer time normalised by the average travel time on a route. The normalised metric is proposed to enable comparisons between different origin-destination pairs.

For the buffer time metric, the upper 95th percentile value is chosen as it is seen to strike an appropriate balance between the amount of extra buffer time to be budgeted and the likelihood of arriving late at the destination. At 95%, the likelihood of arriving later than the budgeted time is 1 in 20, which can be interpreted as once a month for a regular weekday commuter. If the threshold is set lower, the buffer time that passengers need to budget is less, however, there is a higher degree of likelihood of a late arrival. For example, there is a 1 in 10 likelihood for a late arrival at a 90% upper limit. Conversely, a higher threshold reduces the risk of a late arrival, however, the buffer time is longer and this leads to higher travel time costs arising from the inconvenience of regular early arrival at the destination.

Following on from the initial work on the buffer metric for roads applications, the transportation department at the Massachusetts Institute of Technology (MIT) have more recently produced a number of studies focusing on the concept of buffer time for metro systems using data from the London Underground system. The following sections provide a brief description.

One of the initial studies was undertaken by Chan (2007). Chan characterises the reliability of metro services using the Bakerloo, Central, Jubilee, Piccadilly, and Victoria lines as test examples and proposes the reliability factor metric, which forms the basis for further work undertaken by MIT. The reliability factor is defined as the difference between the 95th percentile journey time (indicator of threshold of certainty for reliable service) and the median journey time (indicator of typical service). The reliability factor is a variation of the buffer time metric introduced by the US FHWA, whereby the median is selected as an indicator of typical service rather than the mean. The median value is less sensitive to outliers, and is therefore interpreted as better representing the typical travel conditions. In comparative terms, a smaller value of the reliability factor reflects a more compact journey time distribution, indicating a more reliable service. The reliability factor as presented by Chan (2007) can be aggregated at a line level by applying weightings for passenger flow.

Uniman (2009) makes use of the reliability factor metric proposed by Chan (2007), renaming it the reliability buffer time (RBT) metric. Uniman extends the work by Chan by separating the journey time distributions to distinguish between recurrent and incident-related performance, using data from the Jubilee, District/Circle, and Northern-Victoria lines during the AM peak period. Linear step-wise regression is performed on the 95th percentile journey times aggregated at a day level for each OD pair, and the outlying values are defined as incident-related. The remaining days are classified as recurrent performance days. From this, a new metric is proposed termed the excess reliability buffer time (ERBT). The ERBT is defined as the difference between the overall RBT and the RBT of recurrent performance, and quantifies the effect of incidents on journey time performance.

Building on the work of Chan (2007) and Uniman (2009), Schil (2012) analyses journey time performance on the Jubilee, Piccadilly and Jubilee-Victoria/Central/Bakerloo lines during the PM peak. Schil (2012) makes use of the RBT metric proposed by Uniman (2009), however the analysis focuses on identifying “good” days within a set time period (20 days) and using this as a benchmark to measure unreliable service. The threshold between “good” and “bad” days is defined as the 50th percentile RBT. The 25th percentile corresponds to the median of the "good" service range, and so the day that falls on the 25th percentile of the RBT distribution is nominated as the "good" service day. The use of a rolling average median calculated over the previous 20 day period is also proposed to explore the use of a quasi-dynamic definition of "good" days. The percentage of passengers that travel 5 minutes (or any other set threshold) above the "good" day journey time is defined as a key metric. Schil (2012) also proposes the use of a normalised reliability buffer time metric (NRBT), where the RBT at a line level is normalised by its median journey time to enable comparison of the RBT across lines.

Wood (2015) highlights that the RBT captures journey time variance arising from both variance in the operating service characteristics and the variance in travel behaviours across passengers, such as walking speeds and boarding positions, amongst other factors. Wood argues that individual passengers would tend to exhibit more consistency in their travel behaviour on a given route over time, and so the aggregate RBT overestimates the buffer time from an individual passenger perspective. To provide a measure of a typical individual's buffer time, Wood defines an individual-based reliability buffer time (IRBT). It is calculated by first obtaining the RBT of each individual with ≥ 20 trips on a given route, and then taking the median of these values to represent the IRBT of a typical regular passenger. To assess trip reliability completely excluding individual-specific variance, Wood proposes an additional measure termed the platform-to-platform reliability buffer time (PPRBT). The platform-to-platform time is defined as the time taken from passenger arrival at the origin station platform to passenger arrival at the destination station platform, and consists of the wait time at the origin platform (and at interchange platforms as applicable) along with the in-vehicle travel time. The PPRBT is then computed as the difference between the 95th and 50th percentile platform-to-platform times. Both the IRBT and PPRBT can be aggregated at line level by applying weightings of passenger flow per OD pair.

Tardy trip metrics

This series of metrics focuses on defining thresholds for what constitutes late or unreliable journey times, and using these to provide a measure of the degree of lateness. For the aforementioned buffer time metrics, a limit of the 95th percentile journey time is chosen as the upper bound of reliable journey times, implying that 1 in 20 journey times may be deemed late. The Florida reliability metric and the on-time arrival metric provide alternative ways to define the threshold of lateness, and the misery index metric provides a way to index the degree of lateness. The metrics are briefly detailed in the following section.

The Florida reliability metric involves trialling four different thresholds for the upper limit of journey times i.e. the mean journey time with the addition of 5%, 10%, 15%, or 20% of the mean journey time. The percentage of journey times above the calculated upper limits are then determined from the probability distribution of journey times on a given route for a given time period. An appropriate threshold is chosen, and the upper percentile for reliable journey times is calculated as 100% (all trips) subtracted by the percentage of unreliable trips.

The on-time arrival metric, developed by the California Department of Transportation, is similar to the Florida Reliability Method in that it defines an upper threshold for reliable journey times, but focuses on peak period travel times. The metric defines unreliable trips as trips with journey times greater than or equal to the mean journey time during the peak plus 10% of the mean journey time during the peak. The upper limit for reliable journey times is then defined as 100% (all trips) subtracted by the percentage of unreliable trips.

The misery index provides a measure of how unreliable or late the longest trips are on a given route and time period. The difference between the average travel times for the longest 20% of trips and the average of all trips is evaluated, and then this value is normalised by the average travel time for the route and time period of interest (Lomax et al., 2003). Higher values of the misery index represent a long tail in the distribution of journey times, and therefore indicate that a higher degree of lateness or unreliability is experienced.

Probabilistic metrics

The Dutch Ministry of Transport measures travel time reliability through the use of a probability-based indicator. This type of indicator defines reliable travel times by setting a threshold of the probability that a trip can be made within a specified upper journey time limit. The upper journey time limit is defined as 20% above the median journey time, and it is specified that 95% of all trips on long routes for a given time period should occur within a 20% margin above the median journey time to be considered reliable (Van Lint et al., 2008). Table 2.2 gives the mathematical expression of the indicator.

2.3.3 London Underground performance metrics

Transport for London periodically reports on a number of metrics to assess the performance of the London Underground system. The metrics are categorised into four “Targeted” measures, and ten “Other” measures. The London Underground is also part of the international benchmarking group CoMET (Community of Metros), a group independently administered by the Imperial College Transport Strategy Centre to undertake benchmarking on 16 of the largest metros around the world. Annual benchmarking is undertaken on seven key performance indicators (KPIs). The metrics are summarised in Table 2.3.

Out of the measures listed, there are no measures that explicitly assess the reliability of journey times. From the TfL metrics, the excess journey time metric (measure 4 of the targeted measures) and the total journey time metric (measure 1 of the other measures) are aggregate indicators of journey time performance. The excess journey time metric can be interpreted as an indirect measure of reliability. The lost customer hours metric (measure 10 of the other measures) captures the total amount of system delay. From the CoMET KPIs, measure 5 addresses reliability from a technical systems performance perspective, rather than the reliability of passenger journey times. The following sections describe in more detail the measures of journey time performance and measures of incidents and delays.

Table 2.3: London Underground performance metrics

London Underground metrics: Targeted measures	London Underground metrics: Other measures	CoMET benchmarking metrics
1. Customer satisfaction (score) 2. Passenger journeys (millions) 3. Percentage of scheduled kms (%) 4. Excess journey time (minutes)	1. Total journey time (Minutes) 2. Percentage of scheduled kms - peak and off-peak (%) 3. Percentage of timetabled kms - weekdays and weekends (%) 4. Station closures (number) 5. Escalator availability (%) 6. Lift availability (%) 7. Rolling stock mean distance between failures (km) 8. Signal and points related delays > 2 mins (number) 9. Customer related delays > 2 mins (number) 10. Lost customer hours (hours, '000)	1. Operating cost recovery ratio - Total revenue per total operating cost 2. Operating cost per car km 3. Maintenance cost per car km 4. Labour productivity - Passenger journeys per total staff and contractor hours 5. Reliability - Car km between incidents that cause > 5 mins delay 6. Environment - Grams of CO_2 per million passenger km 7. Safety - Staff hours lost to accidents

Source: Transport for London (2019a, 2015)

Measures of journey time

To evaluate the performance of journey times experienced by customers, the London Underground reports on an in-house developed metric, the journey time metric (JTM). The JTM comprises of five components which represent different stages of a typical passenger journey: (i) access, egress and interchange time (AEI), (ii) ticket purchase time (TPT), (iii) platform wait time (PWT), (iv) on-train time (OTT), and (v) closures time. Table 2.4 provides details on each component.

Value of time (VOT) weighting multipliers are applied to the different activities (e.g. walking, queuing, waiting) for each component of the JTM. "Scheduled" and "actual" values are calculated for each of the five components of the JTM. The measures of performance are then calculated as per equations 2.2 and 2.3. The measures can be aggregated at line and network level by applying passenger demand weightings by line section and time period.

$$Total\ journey\ time_i = (AEI + TPT + PWT + OTT + Closures)_i \tag{2.2}$$

Where i indexes either "scheduled" or "actual" journey times

$$Excess\ journey\ time = Actual\ journey\ time - Scheduled\ journey\ time \tag{2.3}$$

Source: Transport for London (1999) and Transport for London (2016)

The excess journey time (EJT) metric is used as a proxy to measure the reliability of the network, by comparing the actual journey times experienced by customers against ideal scheduled journey times. Lower EJT is associated with reduced levels of "lateness" and interpreted as being more reliable. In their most recent CoMET benchmarking report, TfL also propose the use of the excess wait time metric, calculated from the weighted difference between actual and scheduled platform wait times (Transport for London, 2015). This is put forward as a more useful measure of reliability to compare the performance of different lines on the London Underground.

Measures of incidents and delays

The lost customer hours (LCH) metric is a measure of the amount of delay experienced by passengers as a result of service disruptions from incident events. The LCH metric is defined as the sum of the total extra journey time experienced across all affected passengers due to service disruptions of 2 minutes or more. The metric provides an indication of the severity of the incident in reference to the number of passengers affected. For example, an incident of 3 minutes during an inner zone trip at a weekday peak time would incur more LCH compared to an incident of an equivalent 3 minute duration during an outer zone trip at a weekday off-peak time, as more passengers are travelling in the first scenario compared to the second.

Table 2.4: London Underground journey time metric components

Component	Measurement	Scheduled vs actual
Access, Egress and Interchange (AEI)	Access time - entry into station to midpoint of platform Interchange time (if applicable) - midpoint to midpoint of interchange platforms Egress time - midpoint of platform to station exit	Scheduled times reflect free-flow walking conditions, actual times include estimation of passenger congestion
Ticket Purchase Time (TPT)	Queuing time - time taken for last customer in queue to reach ticket machine Transaction time - time to make ticket purchase	Scheduled time is measured at 90% of average TPT times observed over the past 12 month period
Platform Wait Time (PWT)	Time from arrival at midpoint of platform to wheel-start of train, calculated as: $\frac{\sum H^2}{2\sum H}$ Platform crowding and "left behinds" are also included based on train service frequencies and passenger demand	Scheduled wait time is based on the set working train timetables
On-Train Time (OTT)	Time from wheel-start of train to door opening at alighting platform. Crowding and delays are included through appropriate VOT multipliers	Scheduled train time is based on set working train timetables
Closures	Additional time due to incidents and/or engineering work	Scheduled closures defined as those that have a duration of 2 weeks or more and customers have been notified at least two weeks prior to closure

Note: H is train headway.

To estimate the number of LCH from an incident event, TfL uses the NACHs (Nominally Accumulated Customer Hours) calculator tool. The NACHs tool references the following input data sets: a database of historical incident data grouped by incident cause, the London Underground Train Service Model, which is a simulation model of all timetabled train movements on the network, and historical data on passenger volumes throughout the network aggregated at 15 minute intervals. The historical data sets and train network simulation model are updated and calibrated approximately every 10 years, with the latest known update undertaken in 2011 (Freemark, 2013).

To use the NACHS tool, descriptive information about the incident event must be specified including: the category of incident cause and resulting disruption type, as well as the time of day, location, and the amount of initial delay (for example, for incidents that affect train operations, a train controller will know the exact amount of time a train is delayed in operation). With these inputs, the tool first estimates the total amount of delay caused by the incident event. This is calculated as the specified initial delay plus an estimate of the time taken for the system to recover back to normal operating conditions. Then an estimate for the number of lost customer hours is derived (Transport for London, 2014b). Table 2.5 lists the defined categories for incident cause and disruption type as documented in Transport for London (2014b) and Transport for London (2019a).

The CoMET benchmarking group makes use of the incident data to report on the technical reliability of the London Underground, using the distance (in million car km) between incidents that cause delays of 5 minutes or more as a key performance indicator. The other measures that TfL report on i.e. the number of station closures, escalator and lift availability, mean distance between rolling stock failures, and the number of signal/points and customer related delays, are also used to indicate the magnitude of disruptions on the network from a technical perspective.

2.4 Journey time distributions

Some of the earliest research on journey time distributions was initially motivated by the need to improve journey time estimates in travel behaviour models (Bates et al., 2001; Noland and Polak, 2002). Since then, the functional definition of journey time distributions has become an area of research in itself. As previously discussed, the shape and moments of the distribution, including the mean, mode, range, and variance represent the service conditions on a route as experienced by the passenger. Public transit operators are increasingly recognising the need to measure service quality from a passenger perspective and with access to vast volumes of passenger trip data, it has become easier to generate and functionally define these distributions.

In the following sections, an introduction to the general principles of the distributional form

Table 2.5: London Underground incident cause and disruptions categories

Incident cause	Disruption type
• Customers and public • External factors - external parties e.g. external contractors • Fleet - defects in rolling stock services • Safety and security - emergency/security alerts • Signals • Staff • Station infrastructure - including lift/escalators • Track and civils - track, structures, drainage	• Train delays • Signal failures • Train cancellations/train withdrawals • Depot late start-up • Full and partial line suspensions • Loss of route • Speed restrictions • Train degradation • Partial line degradation • Full and partial station closures • Ticket hall closures • Routeway closures • Escalator failures • Lift and passenger conveyor closures • Platform closures

of journey times is first presented, followed by a review of empirical studies which detail the specific functional forms that best represent journey times for different travel conditions. The majority of the empirical work is on road-based car and bus travel, and so the literature review comprises mostly of these transport modes. A limited number of studies on urban rail systems, however, have been undertaken, and are presented alongside the work on car and bus travel.

2.4.1 Underlying principles of journey time distributional form

Travel times on a route between a given origin and destination possess the following base properties: a) travel times are non-negative, and b) travel times are bound by a lower limit, which is defined as the minimum travel time on a route in free-flowing conditions without delays. The shape of the journey time distribution on a given route is dependent on the service supply and demand characteristics specific to the route. Holding the service factors constant, for a given route at a given time period, distinct distributional properties can be observed depending on the level of passenger demand (Taylor, 1999).

In the absence of congestion, travel times tend to be arranged in a right skewed distributional form. This form arises when the majority of trips are made around a shorter median travel time closer to the free flow time and longer travel times may be observed as a result of individual travel characteristics and/or other operational performance factors. In congested conditions, passenger arrivals tend to be more random, and so travel times are observed to be more symmetrically distributed around a relatively higher mean value. The normal distribution is typically reported as best-representing these congested conditions.

The temporal and spatial resolution over which the journey time distribution is generated is important. Over shorter time intervals and shorter distances, travel conditions are more consistent and the resulting distributions are typically unimodal, i.e. a singular peak of journey times is observed. Over longer temporal and/or spatial intervals, mixed levels of demand can be captured and this can give rise to multiple peaks in the journey time distribution. These distributions are termed multimodal distributions. For illustrative purposes, an example of different journey time distributional forms is given in Figure 2.4, including a right-skewed unimodal log normal distribution, a unimodal normal distribution, and a multimodal distribution.

In the following sections, empirical studies on journey time distributions are reviewed beginning with a review of studies that focus on unimodal distributions, followed by a review of studies that specifically investigate the effect of different spatial and temporal levels of aggregation, and studies that focus on multimodal distributions.

2.4.2 Unimodal distribution studies

Road based travel

To date, empirical research on the definition of journey time distributions has predominately focused on road travel - both private car travel, and public bus services. Wardrop (1952) and Berry (1952) are two of the earliest researchers to note that road travel times follow a right-

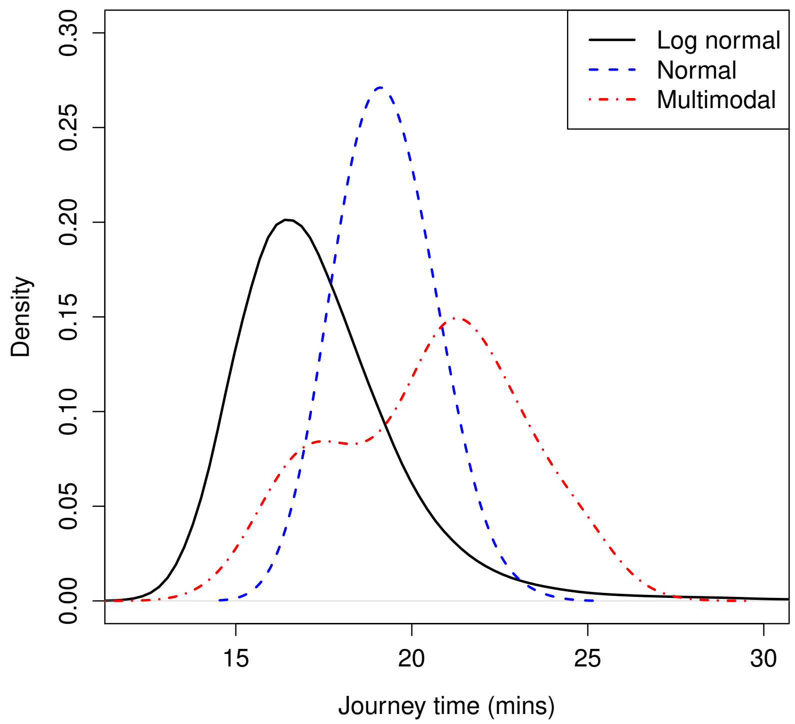

Figure 2.4: Example of different forms of journey time distributions

skewed distribution without specifying functional definitions. Herman and Lam (1974) perform one of the earliest empirical studies, collecting data on car-based work trips in Detroit, US, over a period of 20 months. The resulting journey time distribution is reported as being right-skewed, with the gamma and log normal distributions providing the best fit. Richardson and Taylor (1978) study road based commuter travel in Melbourne, Australia and also identify a right-skewed distribution of travel time, with a log normal distribution providing the best fit. Polus (1979) studies travel time data from 14 arterial road routes in Chicago, US. The distribution of journey times are also found to be right-skewed, following a gamma distribution.

In more recent work, authors have highlighted limitations of the log normal and gamma distributions. Although the distributions are right-skewed and are able to capture typical non-congested conditions, the distributions are not well suited for routes where a larger proportion of longer outlying journey times are observed. As such, more flexible, multi-parameter distributions have been put forward to better capture long-tailed distributions, including the four-parameter stable distribution, the three-parameter Burr distribution, and the two-parameter Generalised

Pareto distribution.

Fosgereau and Fukuda (2012) observe journey times of road users in Copenhagen, Denmark and conclude that the four-parameter stable distribution appropriately represents the distribution of travel times standardised by the mean and interquartile range. The authors do note, however, that the stable distribution overestimates the degree of skewness in the tail of the distributions.

Taylor and Susilawati (2012) and Susilawati et al. (2013) apply the Burr distribution in their studies of journey times in the Adelaide Central Business District in South Australia. Taylor and Susilawati (2012) assess journey times on one of the main routes through the city, and conclude that a Burr distribution provides a good representation of journey times on both the entire route and individual links that comprise the route. Susilawati et al. (2013) assess journey times on a larger portion of the road network in Adelaide, gathering data on two routes comprising of 16 and 22 links respectively. Normal, log normal, Burr, Generalised Pareto, Weibull, and gamma distributions are fitted to each link. The Generalised Pareto and Burr distributions are found to provide a good fit to the journey time data, however, the Burr distribution is seen to be superior as it is statistically significant for a higher proportion of the links. A number of links exhibit a bimodal distribution, and this is discussed further in the section describing multimodal journey time distributions.

Kieu et al. (2014) assess journey times of buses in Brisbane, Australia, trialling 23 different right-skewed distributions, including the Burr and Generalised Pareto distributions. High degrees of skewness and long tails are not observed in the data set, and so the log normal distribution is found to provide the best fit to the data. In some isolated peak periods, journey times are found to be normally distributed. Guessous et al. (2014) study the distribution of road travel times on motorways in Paris, France under different levels of road congestion. It is found that at lower levels of congestion, the distributions are right-skewed, and a Burr distribution provides the best fit. At the highest levels of congestion, the distributions are more symmetrically distributed and there is evidence of multiple peaks. As a result, a mixture distribution comprising of three normal distributions outperforms a unimodal Burr distribution.

Rail transport

While the basic concepts of journey time distributional form remain valid, travel on urban metro systems is inherently different from road transport. While in-vehicle time is the sole consideration for road travel, travel by metro consists of the following components: 1) access time from the entry point to the platform including platform wait time, 2) on-train time, 3) interchange time (if required), and 4) egress time from the train to the exit. Depending on demand levels, service conditions, and individual passenger behaviour, there can be high levels of variance in travel times within each component of the journey. The literature on the definition on rail journey time distributions is limited; two early studies, and three more recent studies are reviewed.

Taylor (1982) presents a comparison of journey time distributions on the Paris metro against travel by bus. Data are collected on morning peak weekday trips over three weeks. The bus journey times follow a normal distribution, while the metro journey times follow a log normal distribution. Overall, the metro journey times are found to be less variable than the bus journey times. Dandy and McBean (1984) perform an empirical study on multimodal travel between Arlington and Cambridge, Massachusetts in the US, assessing three options: 1) car only, 2) bus and subway (metro), and 3) bus and bicycle. Probability distributions are fitted to each component of journey time (in-vehicle times, wait times, transfer times), with normal, log normal, and gamma distributions trialled. In-vehicle travel times by bus and private car are found to follow a log normal distribution. Log normal or normal distributions are deemed appropriate for in-vehicle travel times by metro. Transfer and wait times for bus and metro adhere to a modified uniform distribution, with a tendency toward zero wait time to reflect the condition that passengers tend to arrive and board concurrently as buses or trains arrive.

Duran-Hormazabal and Tirachini (2016) characterise the distribution of journey times for car, bus and metro travel modes in Santiago, Chile. The majority of distributions are found to be right-skewed, with a log normal distribution representing 80% of car distributions, and a log-logistic distribution providing the best fit for rail (46% of all distributions) and bus (45% of distributions) modes. Li et al. (2017) develop a statistical model based on limited dependent variable theory to characterise train running times for the metro in Shenzen, China. The study concludes that a log normal distribution is more suitable than a normal distribution in representing the distribution of running times. Random outlying values in the tail of the distributions, however, are not able to be adequately captured by either distribution. Buchel and Corman (2018) perform a distribution fitting analysis of journey times for trams in Zurich, Switzerland, trialling the Weibull, normal, log normal, and log-logistic distributions. The authors conclude that the log normal distribution provides the best fit to total journey times, as well as the individual components of dwell time and inter-stop running times considered separately.

2.4.3 The effect of temporal and spatial aggregation

The level of temporal and spatial aggregation in generating a journey time distribution influences its form. Several studies have been undertaken to specifically assess the effect of generating distributions at different time intervals (Mazloumi et al., 2010; Ma et al., 2016; Li et al., 2006), and the effect of analysing different lengths along the same route (Rahman et al., 2018; Ma et al., 2016).

Li et al. (2006) study distributions of road travel in Melbourne, Australia at temporal intervals of 5 minutes, 1 hour, and all day. As the level of temporal resolution decreases, the journey time distributions tend towards a more symmetrical form reflecting localised congested conditions, and the normal distribution provides the best fit. Mazloumi et al. (2010) study the effect of using different temporal time intervals to generate the distributions of bus travel times in Melbourne, Australia. The authors find that at short intervals of 5 minutes, the journey time distributions are normally distributed. At intervals of 15 minutes, 1 hour, 2 hours and up to 6.5 hours, the distributions tend to become more right-skewed for off-peak periods, however, remain more symmetrically distributed during the peak. At a time interval of 12 hours, the distribution of journey times in both the peak and off-peak periods tend toward a normal distribution once again.

Ma et al. (2016) gather data on bus travel times in Brisbane, Australia for two routes: one route operates on a dedicated busway service, and the other travels on the general road network. Travel time distributions are generated at 5, 15, 30, and 60 minutes, and at aggregate peak and off-peak periods ranging from 2 hours to 6 hours. It is found that at the shortest interval of 5 minutes, the distributions tend toward a more normal shape, while at longer intervals, the distributions show greater degrees of skewness. At the longest interval of 6 hours capturing the midday inter-peak period, the distribution tends towards a more symmetric normal shape. Ma et al. (2016) also investigate the effect of different levels of spatial aggregation, and find that at an individual link level, travel times tend to exhibit multimodal peaks.

Rahman et al. (2018) test five distributions for bus travel times in Calgary, Canada: normal, log normal, gamma, log-logistic, and logistic. The effect of different levels of spatial aggregation on the shape of travel time distributions is investigated. The authors find that travel times on routes less than 7-8 km are right-skewed, and best represented by the log normal distribution. When travel times are aggregated over routes longer than approximately 7 km, the distributions converge to a more symmetrical form and are best represented by the normal distribution. It is postulated that over very long distances, the journey time distribution is more likely to be influenced by random variation, and so according to the central limit theorem, the distribution would follow a normal form.

2.4.4 Fitting and evaluation methods for unimodal distributions

For most of the unimodal empirical studies reviewed, the maximum likelihood estimation method is applied to fit different theoretical functions to the empirical distribution of journey times. Maximum likelihood estimation is a commonly used fitting technique which selects the theoretical function that maximises the log-likelihood when fitted to the empirical data (Hastie et al., 2009).

The fit of the theoretical function to the empirical data is assessed by goodness-of-fit significance tests and indicators. The most common test used in the empirical unimodal studies is the Kolmogorov-Smirnov test (Polus, 1979; Dandy and McBean, 1984; Taylor, 1982; Taylor and

Susilawati, 2012; Susilawati et al., 2013; Kieu et al., 2014; Rahman et al., 2018). Ma et al. (2016) use the similar Anderson-Darling test. Taylor (1982) and Duran-Hormazabal and Tirachini (2016) use Pearson's chi-squared significance test. All three goodness-of-fit tests are based on evaluating the distance between the theoretical functions and empirical data, under the null hypothesis that the fitted theoretical distribution function is uniformly close to the empirical distribution function (Lehmann and Romano, 2005). The three tests, however, lend focus to different aspects of the distributions. The Kolmogorov-Smirnov test is more powerful in terms of detecting deviations from the median range of the data, whereas the Anderson-Darling test gives relatively more emphasis to the tail of the distributions (Lehmann and Romano, 2005; Delignette-Muller and Dutang, 2015). The Chi-squared test treats all parts of the distribution in an equal way, however, this test requires data that are discrete. As a result, some information on the journey times may be lost when the data are grouped into discrete bins, and so it can be inferred that the Chi-squared test is not the most suitable for assessing the fit of journey time distributions (Lehmann and Romano, 2005).

A small sample of studies use log-likelihood based indicators to directly compare the goodness-of-fit of different theoretical distributions. Guessous et al. (2014) use the Akaike Information Criterion (AIC), while Kieu et al. (2014) and Buchel and Corman (2018) make use of the Bayesian Information Criterion (BIC). Both the AIC and BIC penalise models with high degrees of model complexity, however, the BIC penalises complexity to a greater degree than the AIC. Smaller values of the AIC and BIC indicate a better fitted model (Hastie et al., 2009).

2.4.5 Multimodal distribution studies

When journey times are observed for heterogeneous operating conditions ranging from free-flowing to congested conditions, the journey time distributions may exhibit more than one peak. In this case, estimation of a multimodal distribution function is more appropriate than a unimodal distribution. The majority of work on multimodal distributions has been undertaken for roads-based travel, a review of which is presented in the following sections.

A mixture model framework is typically used to define multimodal distributions. The general form is given in equation 2.4. The fitting process involves first confirming the presence of multiple peaks in the distribution, and then fitting a mixture model with an appropriate number and form of component distributions. In most studies, Hartigans' dip test is initially used to establish the presence of multimodal peaks; the null hypothesis of the test is that the distributions are unimodal. Once the presence of multimodal peaks is established, a mixture model is fitted using the iterative expectation maximisation fitting technique.

$$f(x|\lambda_m, \theta) = \sum_{k=1}^{K} \lambda_m k f_k(x|\theta_k) \tag{2.4}$$

where

x represents journey time,

λ_m is a vector of mixture coefficients,

θ is a vector of model parameters, and

$f_k(x|\theta_k)$ is the normal probability distribution function for the kth component distribution. For bimodal distributions k=2, and for three peaks $k = 3$.

Susilawati et al. (2013) observe travel times on 22 road links in the city of Adelaide, South Australia. Five links indicate statistically significant evidence of bimodality. The authors suggest three conditions that could contribute to the bimodality of travel times: traffic signal performance, levels of congestion, and shortness of link lengths. High levels of congestion on shorter links could result in a greater proportion of traffic being left behind during sub-optimal signal cycles, which could cause substantially longer travel times compared to free flow conditions.

Jun (2010) observe average speeds in 15 minute intervals on the interstate highway in Virginia, US, over one week during the Thanksgiving holiday period to investigate how congestion levels alter before, during, and after the national Thanksgiving holiday. Two-component normal distributions are chosen to model low speed (congested) and high speed (non-congested) operating conditions. The mixture component of the low speed distribution is used to assess the levels of congestion for the time period of interest; a high mixture coefficient indicates higher proportions of low speed (congested) travel.

In their study of bus times, Ma et al. (2016) also observe mixed congested and non-congested conditions on shorter link lengths, which are modelled via multimodal distributions. A mixture model comprising of up to three normal component distributions is used to represent the multimodal form. When distributions are generated at a total route level, the distributions tend toward a unimodal form for the dedicated busway service, but maintain the multimodal form for the non-busway service. For services that run on the dedicated busway route, the authors postulate that drivers can adjust their speed to catch up with schedules. In contrast, the non-busway service is still subjected to influencing factors from signals and differing levels of road congestion on the general road network, therefore giving rise to more mixed travel conditions.

Barkley et al. (2012) study freeway travel times in San Diego, California to define how many distinct states of traffic are present and identifiable in travel time distributions at 5 minute intervals. To avoid over-fitting, a minimum threshold of 1.5 minutes between consecutive peaks is specified. For the peak periods, a three-component model reflecting three distinct peaks provides the best fit, while a two-component model is more appropriate for the off-peak periods. The authors also track the effect of incident events, and find that incidents are typically reflected in the longer second or third peaks of the distributions, rather than the first peak, which captures trips in more free-flowing conditions.

2.5 Factors that affect journey times

The preceding sections of the literature review have focused on research quantifying the degree to which journey times vary in a descriptive manner through the application of journey time performance metrics and the functional definition of journey time distributions. One of the key objectives of the thesis is to quantify the underlying supply and demand factors that influence the variance in journey times. To address this objective, statistical regression techniques are applied.

In the literature, studies that make use of regression methods to analyse transit journey times have been performed, however, the majority of this work is on bus transit. There is a limited body of work that analyses the journey time performance of rail transit systems. These studies, however, are typically used as inputs or as supporting information for other modelling processes such as route choice, or economic appraisal applications, and as a result, the regression models are not as detailed as the models for bus transit. The analysis of the underpinning factors that affect rail transit journey times is also undertaken using alternative empirical methods; much of this work concentrates on individual components of journey time assessed separately. Particular research areas include the analysis of train run times, train dwell times, and passenger wait times at platforms.

In the following sections, research undertaken to identify and quantify the factors that affect rail transit journey times is reviewed. The section begins with an introduction to the general regression model framework, followed by a review of the empirical results of regression analyses on bus and rail journey times. This is followed by an overview of the empirical work on the individual components of rail journey times, including train run times, train dwell times, and platform wait times. The section concludes with a tabular summary of the factors that influence journey times, and an evaluation of the regression methods used in the research to date.

2.5.1 Theoretical regression model framework

In section 2.1, the Wong and Sussman (1973) framework was presented, outlining the three components of journey time variance: regular condition-dependent variation, irregular condition-dependent variation, and random variation. These components of journey time variance can be related to the concepts of deterministic and stochastic variation that can be quantified through regression analysis. The general form of a regression model can be used to represent journey time as illustrated in equation 2.5.

$$y = \overbrace{\underbrace{X_i^T \beta_i}_{\substack{\text{regular condition-dependent}\\\text{variations}}} + \underbrace{Z_k^T u_k}_{\substack{\text{irregular condition-dependent}\\\text{variations}}}}^{\text{deterministic component}} + \overbrace{\underbrace{\epsilon}_{\text{random variations}}}^{\text{stochastic component}} \tag{2.5}$$

Where y represents journey time on a given route. $X_i^T \beta_i$ represents the predictable or regular condition-dependent variations, where X_i is a set of variables that capture sources of known variation in journey times arising from regular operations. $Z_k^T u_k$ is the irregular condition-dependent variation in journey times, and this term represents the effect of unpredictable events, such as incidents and weather events. If ex-post information can be obtained on these events, such as operator incident logs, and historical weather data, a set of explanatory variables Z_k can be included in the model to capture the effect of these irregular events. The remaining random variations represent the variations in travel that typically cannot be measured. These variables are therefore captured by the random error component of the model, ϵ.

The deterministic component of journey time variation, comprising of the set of variables X_i and Z_k, refers to the measurable factors that contribute to journey time variation. In one of the earliest works on the concept of journey time reliability, Abkowitz et al. (1978) propose that the causes of journey time variance as represented by the deterministic component of journey time variance can be categorised into two groups: either intrinsic or exogenous. The US Transit Capacity and Quality of Service Manual also adopts the same categorisation referring to internal versus external factors (Kittelson & Associates Inc et al., 2013). The intrinsic or internal factors are those that are within the control of the operator, and include network and service planning and operations. The intrinsic factors therefore represent those related to service supply. The exogenous or external factors refer to the effects that are not within the direct control of the operator, including the predictable temporal and seasonal variations in passenger volumes, as well as external unpredictable events that can cause variations in passenger volumes, such as weather, or incident events. The exogenous factors therefore represent both the regular factors that influence passenger demand as represented by the set of variables X_i in equation 2.5, as well as irregular sources of variance as captured by the variables Z_k in equation 2.5.

In order to maintain tractability with respect to the statistical regression framework, a distinction is made between the deterministic and stochastic sources of variance. The factors that affect journey time variance are categorised into the sub-categories of supply, demand, unpredictable events, and random variations, and are allocated as follows:

1. Deterministic regression components:

   i. Regular condition-dependent variables, X_i

      – Supply factors

      – Demand factors

   ii. Irregular condition-dependent variables, Z_k - Ex-post information on unpredictable events (eg. weather, incidents)

2. Stochastic regression component, ϵ - Unobserved random variations in travel.

   Where X_i, Z_k, and ϵ refer to the variables defined in equation 2.5.

2.5.2 Empirical studies on total journey times

Bus transit

Table A.1 in Appendix A summarises a selected number of the most widely cited regression analysis studies on bus journey times by location, regression model form, the response variable/s, and explanatory variables. As shown in Table A.1, the most commonly evaluated response is total bus journey time (also referred to as running time), which represents bus operating times between stops and the dwell time at stops. In some cases, the individual components of dwell time and in-vehicle time are set as the response (El-Geneidy and Vijayakumar, 2011; Bertini and El-Geneidy, 2004). A number of studies also model the difference between scheduled and actual journey times as the response (El-Geneidy et al., 2011; Diab and El-Geneidy, 2013; Strathman et al., 1999). To explicitly evaluate the impact of factors on journey time reliability, measures of journey time variance are set as the response. Measures include the coefficient of variation of journey times (El-Geneidy et al., 2011, 2006; Diab and El-Geneidy, 2013; Ma et al., 2015; Yetiskul and Senbil, 2012), the standard deviation of journey times (Yetiskul and Senbil, 2012), and buffer time measures (Ma et al., 2015; Mazloumi et al., 2008).

A similar set of covariates is used across all the models, these can be grouped under the supply/demand framework as discussed in section 2.5.1. On the supply side, the covariates can be allocated to two further sub-groups: 1) the static route characteristics, and 2) the dynamic operational characteristics. The static route characteristics represent the fixed components of the route which are pre-determined during the transit planning process, and include route length, route road type, number of stops, whether the bus travels with other traffic or whether there are reserved bus lanes, the presence of signalised intersections, and bus vehicle specifications such as whether there is smart card functionality, the age of the fleet, and other related factors. The operational factors are those that reflect the variable in-service conditions, and include the amount of delay at the start of the route, headway, the number of actual stops made, the number of lift operations made, and in some cases driver experience is also included.

The covariates capturing the demand side of the framework relate to passenger volumes and factors that influence passenger volumes. Passenger volumes are typically measured at three points: passenger boardings, alightings, on-vehicle passenger volumes. The time of day is shown to have an impact on passenger demand levels, and so this covariate is also modelled. In terms of unpredictable factors that can influence both the supply and demand side of bus transit, weather is the only external factor that is typically included in the models. Interestingly, in all the models reviewed, individual incident events are not modelled as independent variables. This could be a result of insufficient data records, or simply a lack of interest in modelling incidents explicitly. As a result, the effects of incidents are captured by either the other covariates (e.g. the effect of an incident could be reflected in the headway value between buses) or the regression random error term.

Rail transit

Table 2.6 summarises the regression studies on rail transit journey times including information on the location, regression model type, response variable/s, and explanatory variables.

Compared to the regression models for bus transit journey times, the models on rail transit are limited in terms of the number of studies, and less consistent in terms of the response and explanatory variables included in the models. This is a result of the different applications of the models. In most instances, the journey time regression models are used as inputs in subsequent modelling processes rather than as an end result. As a consequence of this, the response and explanatory variables are chosen to meet the specific requirements of the different studies.

Table 2.6: Regression studies on rail journey times

Location and source	Model	Response variables	Significant explanatory variables
Metro and bus, Paris, France (Taylor, 1982)	Linear reg. - OLS	Standardised cv: $\frac{s}{\bar{x}}\bar{x}^{0.5}$	Congestion index: $\frac{\bar{x}}{FF}$ where FF is free flow time
Multimodal - train plus bus, tram or metro, Amsterdam, Netherlands (Krygsman et al., 2004)	Linear reg. - OLS	Journey time	Gender, age, number of children in household, no. of interchanges, trip motive, mode of travel to and from stations, time to travel to origin station, urban form (CBD vs non-CBD)
Metro, London, UK (Uniman, 2009)	Linear reg. - OLS, Feasible Generalised Least Squares (FGLS), FLGS with first-order serial correlation	Reliability buffer time: $95\% - 50\%x$	Median travel time, dummy for whether route involves interchange, dummy for incident event
Metro, Singapore (Sun et al., 2012)	Linear reg. - OLS	Journey time	Access and egress distances, number of intermediate stops/dwell time, train speed

Note: x is journey time, $\bar{x}$ is mean journey time, s is sample standard deviation, and cv is the coefficient of variation.

In the earliest example, journey times on the Paris metro are analysed to investigate the impact of congestion on the variance of journey times. Journey time data are collected via stated preference surveys, and an univariate OLS regression is performed to show that the standardised coefficient of variation of journey times increases with higher levels of congestion (Taylor, 1982). With access to automated data sources, the more recent studies extend the univariate model form to assess the concurrent impact of different covariates using multivariate linear regression.

Total journey times are designated as the response variable in models by Krygsman et al. (2004) and Sun et al. (2012). Krygsman et al. (2004) use the model as an input into mode choice analysis for public transit in the Netherlands, while Sun et al. (2012) use the outputs of the journey time model to allocate passengers to trains. Uniman (2009) sets the response variable as the reliability buffer time, as part of the assessment of the reliability of journey times on the London Underground.

With the obvious mode-specific differences aside, the covariates modelled in the rail studies can be categorised in a similar way as the bus studies. On the supply side, there are covariates related to the static network characteristics and dynamic operational factors. The static network factors modelled across the studies include: 1) station characteristics such as station type, and access and egress distances 2) route properties including the number of interchanges, route length, inter-station lengths, signalling block lengths, and the number of stops, and 3) rollingstock characteristics such as rollingstock model type, and door width. The dynamic operational factors include headway, timetabled and actual arrival/departure times, dwell times, and train speeds.

On the demand side, the variables can be grouped into those related to passenger volume, and those related to individual passenger characteristics. Variables to capture the effect of passenger volumes are modelled at multiple points, including volumes at boarding, alighting, and on-train. The time of day is also included as a covariate to capture the temporal fluctuations in passenger volumes. In their journey time model, Krygsman et al. (2004) include individual passenger characteristics including gender, age, number of children per household, and the urban form surrounding the individual's residence.

In terms of variables to capture unpredictable events, Uniman (2009) models the effect of incidents on the variance of journey times using a dummy variable to denote the occurrence of an incident event. Unlike the bus studies, the rail studies reviewed here do not include a covariate to model the effect of weather. This could be attributed to the fact that the rail systems studied are typically underground metro systems, and so there is a perceived lack of influence on train operations from weather conditions. Any potential changes in passenger demand as a result of weather are left to be captured in the passenger volume covariates.

2.5.3 Train running times

The in-vehicle time of a passenger, referred to as the train running time in train operations literature, is determined by the planned train network schedule which specifies the planned arrival and departure time of every train at each station. The main input of a train schedule is the train speed profile along the route. The train speed profile is in turn developed using the fundamental mechanics of train motion. Here the key principles of train motion are briefly summarised to highlight the underpinning factors that affect train running times.

For forward propulsion, the train motor providing the forward tractive force must overcome a number of internal and external resistive forces including: 1) train running resistance which captures the mechanical and aerodynamic resistive forces of the train unit, 2) curve and grade resistance which captures the forces on the train as a result of the horizontal and vertical geometry of the track, and 3) the inertia resistance which is caused by the acceleration required to bring the train into motion from the stationary position or at any other acceleration point when the speed profile of the train is not constant (Profillidis, 2014; Luthi, 2009; Lukaszewicz, 2001).

Based on empirical work and building on earlier empirical formulations by Schmidt (1910) and Strahl (1913), the seminal equation for train running resistance was proposed by Davis (1926) (Luthi, 2009). Running resistance is specified as a quadratic in train speed; the general form is given in equation 2.6. This equation is still widely applied to evaluate train movement, however, the coefficients of speed have been shown to vary widely across different rail systems. The coefficients are derived via full-size or scaled testing. Profillidis (2014) provides an overview of the coefficients used by France, Germany, and the US.

$$R = A + \frac{B}{W_{train}} + CV + \frac{KV^2}{W_{train} n_{axles}} \tag{2.6}$$

where R is the unit resistance of the train (dimensionless), W_{train} is the train carriage weight per axle, n_{axles} is the number of axles, V is the train speed, and K is a drag coefficient which captures the aerodynamic effects of the train face shape, the longitudinal shape, and skin friction. The parameters A, B, C are empirical constants specific to the rollingstock and rail network under consideration.

Accounting for the resistive forces, the train speed profile can then be developed between each stop along the route. The complete network timetable (i.e arrival and departure times at each station) can then be specified. The inputs include the train speed profiles along the entire route, dwell times at stations, the operating headways, the minimum safe separation distance between trains, and an allowance for a buffer time to absorb possible fluctuations in the running and dwell times (Kittelson & Associates Inc et al., 2013; Glover, 2013). After the base timetable is developed for the network, parameters can be altered in order to achieve the required operating conditions. There are two broad streams of research relating to the analysis and optimisation of the timetable: numerical optimisation methods and simulation methods. The optimisation methods use the empirical relationships of train motion as the basis, with boundary conditions applied to capture operational limitations (e.g. speed, headways, dwell times). Train cost functions are then developed, and the objective is to either minimise total run time or energy usage (Howlett, 2000; Khmelnitsky, 2000; Liu, 2003; Albrecht, 2013). The simulation methods involve the development of a model of the entire rail network using the empirical train motion equations, along with inputs of the network and operational characteristics and boundaries (Kim, 1997; Goodman, 1998; Kim, 2010). The simulation methods are used to analyse and compare different timetabling and/or network schemes, and can be used to develop the most optimal timetable under different conditions, or to evaluate the impact of upgrades or changes to the operations and/or network.

There is an emerging field of research that makes use of regression methods to analyse train running times, however work is currently limited. Table 2.7 provides a summary of the studies. The objectives and applications of the regression analyses are diverse. Murali et al. (2010) and Hansen et al. (2010) use the regression model to obtain improved estimates of train running times for application in simulation models. Kecman and Goverde (2015) analyse train run time and then use the results from the linear regression model as a base to compare the performance of predictive numerical regression tree and regression forest algorithms. Gibson et al. (2002) use the model to estimate congestion charges based on values of delay for operators to use the country-wide rail network in England. Gorman (2009) performs the regression modelling for the purpose of quantifying and comparing how different operational factors affect run times across eight freight rail systems in the US.

Table 2.7: Regression studies on train running times

Location and source	Model	Response variables	Significant explanatory variables
Rail, England, (Gibson et al., 2002)	Linear reg. - OLS	Delay i.e. difference between observed vs scheduled departure/arrival time	Route ID, route section ID, capacity utilisation index (i.e. ratio between time taken to operate optimal timetable to actual timetable), time of day
Rail, 8 locations in western US (Gorman, 2009)	Linear reg. - OLS	In-vehicle time (total train running time)	Direction, delay at start of route, meets (when trains in opposite directions cross), overtakes (train overtaking another in the same direction), no. trains in preceding 24hr period, total train hours, headway
Rail, simulation data from Los Angeles, US (Murali et al., 2010)	Linear reg. - OLS	Delay	Route length, speed, no. of crossings, crossing spacing, no. of trains travelling in opposite direction, no. of trains in section
Rail, Rotterdam and The Hague, Netherlands (Hansen et al., 2010)	Linear reg. - OLS	In vehicle time	Departure delay
Rail, Rotterdam and The Hague, Netherlands (Kecman and Goverde, 2015)	Linear reg. - least trimmed squares (LTS)	In-vehicle time	Dummy for peak hour, departure delay, headway, distance to next station, distance from previous station, block length, rollingstock type

The dependent variable in the models is set as either train running time or a measure of delay. In Hansen et al. (2010), Kecman and Goverde (2015), and Gorman (2009), the run time is measured at a line or signalling block level to include stops and dwell times at all intermediate stations. Gibson et al. (2002) and Murali et al. (2010) use delay as the dependent variable, which is defined as the difference between the observed and scheduled arrival and/or departure time.

In terms of the independent variables, the factors modelled relate more to service supply rather than passenger demand. This is as expected; the train operations are the primary determinant of train run times. Demand is accounted for via the operating parameters of headway and dwell time at stops. As previously described, the service supply related variables can be categorised into static and dynamic factors. The static factors are time-invariant and capture fixed properties of train service. The studies model static factors including those related to the route such as route length and the number and spacing of crossings (Murali et al., 2010), and those related to the rollingstock factors such as rollingstock type (Kecman and Goverde, 2015). The dynamic factors reflect the time-varying operating conditions, including: headway, speed, the delay at the start of the route run, the number of meets (when trains travelling in opposite directions cross), the number of passes (when a train overtakes another), the number of overtakes (when a train is overtaken) (Gorman, 2009; Murali et al., 2010; Hansen et al., 2010; Kecman and Goverde, 2015). Gorman (2009) additionally finds that the number of trains in the preceding 24 hour period, along with the total train hours in operation are significant in the model of train run times for US freight railways.

2.5.4 Train dwell times

The dwell time is defined as the time a train stops at a station platform for the loading and unloading of passengers. More specifically, dwell time comprises of the time taken from wheel stop to door opening, the time for passenger boarding/alighting movements, any residual door-open time without passenger movement, and the time for the doors to close to wheel-start of the train. The dwell time can be a determining factor in line capacity planning; long dwell times, or high levels of variation result in the requirement of longer headways between trains and fewer services. The US Transit Capacity and Quality of Service Manual (TCQSM) states that dwell times are typically specified based on values observed from past experience, and range from 30 to 50 seconds for rail stations operating near capacity (Kittelson & Associates Inc et al., 2013).

For the specification of dwell times for new rail lines based on empirical analysis, the US TCQSM refers to Parkinson and Fisher (1996), who propose a series of equations derived from linear regression analyses of dwell times observed on 10 urban rail systems in the US and Canada. The equations estimate passenger flow times, i.e. the time for a passenger to move through a train door, as a function of boardings, alightings, and the number of on-board standing passengers. The flow time is then multiplied by the peak flow rate at the most heavily used door to obtain an estimate for the dwell time. An earlier and widely-cited study by Weston (1989) estimates train dwell times for the London Underground. Weston (1989) proposes a more detailed formula where dwell time is a function of the number of boarding and alighting passengers, through passengers (i.e. those on the train who do not alight), the number of train seats, and a ratio of the peak door to average door to capture door loadings (see equation 2.7).

$$SS = 15 + 1.4\left[1 + \frac{F}{35}\left(\frac{(T-S)}{D}\right)\right]\left[\left(\frac{FB}{D}\right)^{0.7} + \left(\frac{FA}{D}\right)^{0.7} + \left(0.027\frac{FB}{D}\frac{FA}{D}\right)\right] \tag{2.7}$$

where SS is station stop time in seconds (or dwell time), A is the number of alighting passengers, B is the number of boarding passengers, D is the number of train doors, F is a door factor and is the ratio between peak door to average door, T is the number of through passengers, and S is the number of seats on the train. The constant of 15 seconds represents the "function time" for door opening/closing operations and the time to train departure after the closure of doors.

Harris and Anderson (2007) perform a verification of Weston's formula using data from 31 metros around the world. They conclude that with some minor adjustments to boarding and alighting parameters to capture system-specific conditions, the formula for dwell time is valid and appropriate for estimating dwell times for different metros. An earlier study by Harris (2006) tests Weston's formula at high levels of passenger flow, using observed data from some of the busiest stations along London's Southwest overground rail line, however concludes that the formula does not hold at high levels of passenger demand ($\geq$ 50 passenger movements).

Along with the two widely-cited approaches discussed, researchers have investigated the impact of other aspects of train service and operation on dwell times. Across the studies, three main groups of factors are identified: factors related to passenger movement and volumes, rollingstock characteristics, and station characteristics including the train/platform interface. A selection of the most recent work is summarised in Table 2.8. The most commonly modelled covariates are those relating to passenger volumes; more specifically, the number of boardings, alightings, and through passengers standing on-board the train. The consensus across the studies is that an increase in passenger volumes corresponds to longer dwell times, supporting the formulations previously established in the earliest analyses of dwell times by Weston and the TCQSM. Some studies include covariates that may directly or indirectly influence passenger volumes, such as the time of day to capture peak vs off-peak effects (Kecman and Goverde, 2015; van den Heuvel, 2016; Christoforou et al., 2017), and station classification based whether the station is located in a major hub or smaller suburban area (Kecman and Goverde, 2015; Christoforou et al., 2017).

Another field of research focuses on quantifying different characteristics of passengers, and using these to provide a more detailed analysis of passenger movement during dwell times. The most widely cited works include Wiggenraad (2001) and Heinz (2003). Wiggenraad (2001) further defines different types of passenger movement categories, distinguishing between clustered and non-clustered passengers and their movements. Heinz (2003) uses the underlying theory on the physiological mechanisms of walking to evaluate the impact of door width and step height, and in turn, dwell times.

Other studies investigate the effect of rollingstock-specific characteristics, including the rollingstock model type, the number of train doors, door width, the distance between train doors, car area, the type and number of seats, and the presence of additional train car furniture such as steps or hold poles. Another subset of work also investigates the effect of station-specific characteristics, in particular, the horizontal and vertical step distances at the platform-train interface, platform widths, and train positioning along the platform. Similar to the methods used by (Thoreau et al., 2016), a number studies are undertaken using laboratory methods, where a model of a train carriage is constructed to investigate the effect of rollingstock and station platform factors such as door width, step height, train furniture, and platform management techniques (Seriani and Fernandez, 2015; Holloway et al., 2016; Karekla and Tyler, 2012).

Table 2.8: Empirical studies on train dwell times

Location and source	Model	Response variables	Significant explanatory variables
Light rail, Massachusetts, US (Lin and Wilson, 1992)	Linear reg. - OLS	Dwell time (several models capturing same set of covariates)	No. boardings, alightings, no. standing passengers arriving and departing
Metro, Hong Kong (Lam et al., 1998)	Linear reg. - OLS	Dwell time	No. boardings, alightings
Metro, Boston, Massachusetts, US (Puong, 2000)	Linear reg. - OLS	Dwell time	No. boardings, alightings, and through-standees (i.e. standing passengers on train who do not alight)
Rail, Sydney, Australia (Douglas Economics, 2012)	Linear reg. - OLS	Dwell time	No. boardings, alightings, number of standing through passengers
Urban rail, international (Harris et al., 2014)	Multivariate fractional function	1) Boarding rate (passengers per sec) 2) Alighting rate (passengers per sec)	1) No. alightings, ratio alightings to boardings, ratio through passengers to train capacity, vertical stepping distance from platform to train, seat density, distance between train doors 2) No. boardings, ratio alightings to boardings, ratio through passengers to train capacity, ratio train car area to door width, standback distance, dummy for steps on train, platform width
Rail, Rotterdam and The Hague, Netherlands (Kecman and Goverde, 2015)	Linear reg. - least trimmed squares (LTS)	Dwell time	Dummy for peak hour, arrival delay, rollingstock type, station type, scheduled time
Rail, Amsterdam, Netherlands (van den Heuvel, 2016)	T-tests	Dwell time	Peak vs non-peak times of day, train position,
Metro, experimental data based on Korea metro (Oh et al., 2016)	Linear reg. - OLS, Least Absolute Deviations (LAD)	Dwell time	No. boardings, alightings, standing, door width
Rail, London, UK - Laboratory experiment (Thoreau et al., 2016)	Experimental, empirical relationship	Boarding rate, and alighting rate	Door width, seat type, horizontal gap between train and platform
Light rail, Nantes, France (Christoforou et al., 2017)	Linear reg. - OLS	Dwell time	Time of day, line, direction, if station is in CBD area or not, if station is close to any points of interest, no. boardings, alightings, vehicle type

2.5.5 Platform wait times

Fundamental properties of wait times

As mentioned in section 2.3.1, the static wait time at the platform is one of the most inconvenient parts of a rail journey in terms of passenger disutility. The earliest work on estimating wait times for rail transit was based on work undertaken for bus transit, where researchers made the assumption that passengers arrive randomly, and the resulting wait times follow a uniform distribution (Holroyd and Scraggs, 1966; Osana and Newell, 1972). The simplified assumption of uniformly distributed wait times is still currently used as the basis for the estimation of platform wait time on the London Underground by Transport for London (see section 2.3.1 for further details).

O'Flaherty and Mangan (1970), Seddon and Day (1974), and Jolliffe and Hutchinson (1975) were among the first researchers to observe that although some passengers do indeed arrive randomly, others use knowledge of schedules via published timetables or experience to adjust their arrival to minimise wait times. In subsequent work on wait times, researchers generally adopt the two-category typology of passenger wait times proposed by Turnquist (1978), which can be summarised as follows:

1. Passengers arrive randomly when service frequencies are high and resulting expected wait times are low.
2. Passengers arrive non-randomly and time their arrival to minimise wait time when service frequencies are low and train schedule information is known.

The transition point between random and non-random arrivals is found to occur at different values of headway depending on the transit mode and location. The US Transit Capacity and Service Quality Manual currently quotes the transition at 10 minute headways, though some of the most recent research quotes values as low as 5 minutes (Kittelson & Associates Inc et al., 2013; Ingvardson et al., 2018). When passenger arrivals are perfectly random with a uniform distribution, the expected wait time $E[w]$ is half of the headway, i.e. $E[w] = 0.5h$, where h is the headway between consecutive services. When passenger arrivals are a mix of random and non-random, it follows that the expected wait time can be assumed to be less than half of the headway, and this is found to be the case in early work on bus transit.

O'Flaherty and Mangan (1970) characterise bus wait times in Leeds, UK, as a linear function of headway, where the expected wait time is defined as $E[w] = 1.79 + 0.14h$. Substituting $E[w] = 0.5$, the distinction between random and non-random arrivals occurs at 5 minute headways. Seddon and Day (1974) analyse bus wait times in Manchester, UK, and also observe a linear function where $E[w] = 2.34 + 0.26h$ and the transition from random to non-random arrivals occurs at approximately 10 minute headways. In more recent work, Nygaard and Torset (2016) study bus travel in Trondheim, Norway and observe non-random arrivals behaviour for bus services with 10, 15, and 20 minute headways. For the 10 minute headway service, the authors specify a linear model for the expected wait time, where $E[w] = 0.32h$. In their study of bus wait times in Austin, Texas, Fan and Machemehl (2009) analyse services with headways ranging from 8 to 60 minutes and also specify a linear model for wait times, where $E[w] = 2 + 0.3h$. The transition from random to non-random arrivals behaviour is reported to occur at 11 minute headways, and it is reported after a headway of 38 minutes, all arrivals can be considered as non-random. Guo et al. (2011) study the passenger arrivals behaviour for passengers transferring from rail services to bus services in Beijing, China. The passenger arrival rate is found to follow a log normal distribution for passengers who transfer from rail services, while a gamma distribution is found for passengers who do not connect from the rail service. Gong et al. (2016) perform a similar analysis in Beijing, China, and conclude the same findings; arrival rates follow a log normal for trips transferring from rail services, and a gamma distribution for those that are not connected.

Two of the most recent studies on rail travel by Luethi et al. (2007) and Ingvardson et al. (2018) use two-part mixture distributions to capture the random and non-random arrivals. Luethi et al. (2007) analyse platform wait times for bus, tram, and rail in Zurich Switzerland, and propose a mixed uniform and Johnson SB distribution to account for random and non-random arrivals. The transition from random to non-random arrivals is observed to occur at approximately 5 minute headways. Ingvardson et al. (2018) study wait times for metro and regional rail lines in Copenhagen with headways ranging from 2 to 60 minutes, and propose a mixed uniform and beta distribution of passenger arrivals. This distribution is tested against the Luethi et al. (2007) mixed uniform and Johnson SB function, and is found to provide a better fit to the wait time data. In terms of the passenger arrivals process, the authors find that the proportion of passengers that exhibit non-random arrival behaviour timed to minimise wait times is estimated at 43% at 5 minute headways, and 52% at 10 minute headways, and so it can be inferred that the transition between random and non-random arrivals occurs between 5 and 10 minute headways.

Factors that influence wait times

In line with the seminal formula for wait times for high frequency services widely adopted by operators (see equation 2.1 in section 2.3.1), empirical research verifies that the degree of headway regularity is one of the main determining factors in the passenger arrivals process and resulting distribution of wait times. Most studies report that at higher degrees of service irregularity, passenger arrivals are observed to be more random, resulting in longer wait times. Furth and Muller (2006) propose a formula for the transition between random and non-random arrivals based on the reliability of headway deviations and schedule deviations. Using six different operational scenarios, it is shown that passengers are more likely to time their arrival with the schedule when service is more consistent. Bowman and Turnquist (1981) further confirm that for observed bus wait times in Chicago, passengers are more sensitive to the regularity of headways rather than headway itself. Frumin and Zhao (2012) assess wait times on the London Overground rail lines, and find that passengers arrive more randomly when the schedule deviation of the service is high. Nygaard and Torset (2016) assess bus wait times in Trondheim, Norway, and show that passengers experience longer wait times as the deviation between actual and scheduled arrival increases.

Apart from the influence of headway regularity, empirical work shows that passenger wait times are also influenced by factors including the time of day, the availability of train timetable information, and other passenger-specific behaviours. Here a brief summary of the some of the most widely-cited and recent work in this area is presented. Ingvardson et al. (2018) assess rail travel in Copenhagen, Luethi et al. (2007) study bus, tram, and rail modes in Zurich, Frumin and Zhao (2012) study the London Overground rail lines, and Csikos and Currie (2007) investigate rail travel in Melbourne, Australia. In their analysis of passenger wait times, the four studies investigate the impact of the time of day and all four report that a higher proportion of passengers arrive randomly in off-peak hours than during the peak. It is argued that those who travel in the peak are more likely to be regular travellers who have knowledge of the train departures, and are therefore more likely to time their arrivals.

Ingvardson et al. (2018) also find that a greater proportion of passengers arrive randomly for the metro service, which operates to a set service frequency rather than to a scheduled timetable. In comparison to the random metro arrivals, they observe that passengers time their arrival to minimise wait times for the rail lines where published timetables are available. This illustrates the impact of passenger information on wait times - where information is available, passengers will make use of it to time their arrivals in order to reduce wait times. Ingvardson et al. (2018) also find that station characteristics have an impact on wait times; passengers have longer wait times at stations with complicated access routes than those who have direct paths to the platform. These stations also tend to be the larger stations servicing large passenger volumes, and so station congestion could also have an influence in these cases.

Csikos and Currie (2008) extend on their earlier work on rail wait times in Melbourne, Australia and conduct a longitudinal study to analyse the arrival behaviour of regular passengers who complete the same route for at least 10 weekdays. Four types of arrivals behaviour is defined: 1) those who have the greatest consistency in arrival time and arrive within a narrow window (1-3 minutes before scheduled service), 2) those who arrive randomly but within a consistent wider window of arrival time, 3) those who arrive consistently within a given window, but also display outlying arrival times, and 4) random arrivals. The authors find that passengers with the most consistent and narrow arrival window also exhibit the shortest offset to the train arrival time (i.e. passenger arrival coincidences very closely to train arrival), and tend to travel earlier in the peak period of the day.

Fan and Machemehl (2009) analyse wait time for buses in Austin Texas and perform a linear regression analysis using OLS fitting to investigate the influence of headway, headway variance, location, time of day, gender, ethnicity, and access mode. The variables found to be significant at 95% are headway, headway regularity, the dummy variable for the off-peak period, and the dummy variable for ethnicity type. All significant variables have positive coefficients and so increases in the value of all variables lead to longer wait times.

2.5.6 Summary of factors that influence urban rail journey times

From the preceding review of empirical research, the key factors that influence urban rail journey times by journey time component are summarised in Table 2.9. The factors are further categorised under the groups of regular variations, which include the sub-categories of supply vs demand factors, and irregular variations as per the framework defined in section 2.5.1. This summary table is used as an input to identify the covariates that are used in the analysis of journey times in the thesis.

Table 2.9: Factors that influence journey time variance

Journey time component	Definition	Regular variations: Demand	Regular variations: Supply	Irregular variations
General	Across all components	Regular daily variation in passenger volumes between peak and off-peak times, and seasonal variation	Planned closures or modifications to service schedules	Incident events (from passenger demand and service supply causes), weather, other unpredictable events that cause fluctuations in passenger demand/service supply
Access time	From passenger tap-in at gates to train wheel stop at origin station. Includes walk time from the gates to platform, and wait time at platform.	Passenger volume entering station, at platform, and boarding train Individual passenger movement characteristics	Service frequency/ headway between trains Origin station capacity, including: Walk distance to platform Vertical transport to platform level Platform area	
Dwell times	From train wheel stop to wheel start at origin/destination stations	Passenger volume entering/exiting trains, on train, and at platforms	Service schedule	
On-train time	From train wheel start to wheel stop between origin and destination stations		Route distance Train speed	
Egress time	From train wheel start to passenger tap-out at destination station	Passenger volume alighting train, on platform, and exiting at station Individual passenger movement characteristics	Destination station capacity, including: Platform area Walk distance to exit Vertical transport to exit level Number of exit gates	

2.5.7 Evaluation of regression methods used in empirical work

For the rail and bus studies that use regression methods to analyse journey times, multivariate linear regression based on ordinary least squares (OLS) fitting is the most commonly used approach (see Tables 2.6, 2.7, 2.8, and A.1). The main drawback of the OLS models is that strict assumptions apply for the model to be valid, namely adherence to the Gauss-Markov conditions to ensure that the parameter estimators are unbiased and efficient (Greene, 2018). In cases where the data cannot adhere to the governing conditions, some studies make use of alternative fitting methods to address issues including correlation between covariates, heteroskedasticity, and non-normality of underlying covariate distributions. Ma et al. (2015) use the seemingly unrelated regression equations (SURE) method for the analysis of bus transit in Brisbane, Australia. The SURE method was first applied in transit journey time research by Martchouk et al. (2011) for highway car travel, and the method works by correcting for correlation between different equations in the model. Yetiskul and Senbil (2012) use maximum likelihood estimation (MLE) to account for the potential non-normal distribution of covariates in the model of bus journey times in Turkey.

For the rail transit studies, Uniman (2009) develops a linear regression model to analyse journey time variance on the London Underground on 800 different origin-destination routes, and uses two methods to correct any unequal variance in the residuals arising from the different route properties. The first method involves the use of a feasible generalised least squares (FGLS) structure to allow for unequal variance in the residuals, and the second model makes use of FGLS with first-order autoregressive error structure (AR1) to correct for correlation between observations within each OD route group. Both methods perform better than a base OLS model, and the FGLS model with AR1 errors performs best. Kecman and Goverde (2015) apply a least-trimmed squares (LTS) estimation method in the analysis of rail journey times in the Netherlands. This fitting technique identifies and excludes outliers in the model, giving more robust parameter estimates. In the model of metro dwell times in Korea, Oh et al. (2016) use the least-absolute deviations (LAD) fitting method. The LAD method is similar to the MLE method in that the OLS assumption that residuals must be normally distributed is relaxed to accommodate non-normal distributions.

Although the non-OLS models can adequately address some violations of Gauss-Markov conditions, in all studies reviewed, a basic linear structure is used, and so the relationship between the independent and dependent variables is restricted a priori. The use of linear relationships simplifies the interaction between the variables of interest, and important information can be omitted in the process. Given the quantity of data available from AFC sources, it is possible to generate flexible, non-linear relationships from the data points via semiparametric regression. All information between the independent and dependent variables is retained, and so this method defines the relationship between variables with a higher degree of fidelity to the data compared to methods that require linear a priori assumptions on functional form. As such, there is a gap in the literature for scope to apply more flexible analytical methods to quantify the factors that contribute to journey time variance with a higher degree of accuracy.

2.5.8 A note on machine learning methods

Aside from regression methods, another method increasingly applied in the analysis of large data sets is machine learning. Rather than generating predictions based on explicit parametric relationships that specify how a response systematically varies with respect to given explanatory variables, the predictions are generated based on the magnitude and dispersion of the data points representing the response. The main distinction between machine learning methods and conventional multivariate regression analysis is that machine learning is predominately utilised to make predictions of the response variable, whereas regression methods are more concerned with description, i.e. quantifying the degree of influence of different explanatory variables on the response variable (Matloff, 2017; Hastie et al., 2009). The underlying mechanism of machine learning methods is that the predictive ability of the algorithm improves through "learning" from the response data points.

There are many forms of machine learning algorithms available; the general categorisation in the literature is to distinguish between supervised and unsupervised learning algorithms. For the supervised algorithm subset, a sample or "training" data set is used with known values of the explanatory variables (termed "features" in machine learning terminology) and the response variable. The training data are used in a first run of the algorithm to estimate the model features using known values of the response. The model's predictive ability is then verified through a second sample or "prediction" data, where the algorithm's predictions are compared against the known true value of the response (Matloff, 2017; Hastie et al., 2009). Within the supervised learning subset, the algorithms can be further categorised based on the form of the response variable; if the response is continuous, the algorithms are termed "regression algorithms" and if the response is categorical, the algorithms are termed "classification algorithms" (Matloff, 2017; Hastie et al., 2009). For the unsupervised algorithms, only the features (explanatory variables) are known and the value of the response is unknown. The unsupervised algorithms group or cluster the data based on the features in order to generate predictions (Matloff, 2017; Hastie et al., 2009). Further details on the different algorithms can be found the textbooks by Hastie et al. (2009) and Matloff (2017).

The application of machine learning for transport applications is an emerging field. Goldberg (2018) provides a useful summary table of the specific algorithms that are used in transport modelling applications, including descriptions of their advantages and disadvantages. The type of response generated by the machine learning based studies is varied; for example, studies focus on estimating transit demand through predictions of traffic flows and ridership (Liu and Chen, 2017; Guo et al., 2012; Wei and Chen, 2012; Leshem and Ritov, 2007), estimating responses to disrupted conditions (Goldberg, 2018; Guo et al., 2012), and a number of studies also focus on estimating travel times (Wen, 2017; Kecman and Goverde, 2015; Han, 2012). As mentioned earlier, the objective of applying machine learning methods is to generate predictions rather than quantifying the degree to which the explanatory variables influence the response of interest, and so the studies that adopt machine learning methods do not typically include detailed analyses of the factors that influence the response.

This thesis is not concerned with the explicit prediction of journey times, but rather decomposition and quantification of the degree to which different factors contribute to journey time variance. Conventional econometric regression methods remain a time and resource efficient way of undertaking this type of analysis. It should be noted, that machine learning algorithms can in theory be re-arranged to set the outcome of the algorithm to predict the value of the coefficient or elasticity of an explanatory variable rather than predicting the response. In cases where the effect of many explanatory variables are required to be estimated however, this process is likely to be unduly time and resource-inefficient. On this basis, the application of machine learning methods is excluded from the scope of the thesis, and is instead recommended as a direction for future work, for example, to compare the estimates of covariate elasticity and fixed and random effects generated by the semiparametric regression models in Chapters 6 and 7.

2.6 Summary of literature review and research contributions

The preceding chapter has presented a review of the research areas that are addressed in Chapters 4, 6, and 7 of the thesis. As mentioned previously, the literature covering the assignment of passengers to trains is reviewed separately in Chapter 5. The specific contributions of the thesis as presented in Chapters 4, 6, and 7 are summarised as follows:

- In terms of the metrics used to quantify journey time performance, the metrics currently used by the London Underground are more focused on quantifying performance from an operator perspective rather than a passenger perspective, and this is a key limitation. This gap is addressed in Chapter 4 of the thesis; building on the most recent work on passenger-oriented performance metrics, three new metrics are proposed to evaluate journey time variance from a passenger perspective.

- Compared to the work on road based transit, the literature on the functional definition of rail transit journey times is limited. The thesis contributes to this field by undertaking an analysis to fit theoretical distributions to rail journey time data on the London Underground in Chapter 4. The form of the journey time distributions is mathematically defined and the moments of the distributions are used to provide a more detailed understanding and definition of the fundamental characteristics of rail journey times under different operating conditions.

- In the review of empirical research on the underlying drivers of journey time variance, the majority of research on rail journey times is segmented and focuses on analysing components of journey time individually. This is useful in analysing factors that influence individual components of journey time considered in isolation. However, this does not enable practical conclusions to be drawn in terms of quantifying the relative influence of factors on both total journey times and the components of journey time for a single system. From the review of the empirical research, the most influential variables are identified and using this, a more unified approach to analyse the effect of various supply and passenger demand factors on journey times on the London Underground is proposed in Chapters 6 and 7. A series of regression models are developed to directly compare the relative effect of different factors on access, on-train, and egress times, as well as total journey times. A ranking of which factor/s have the most influence both overall and at different points within a typical rail journey is then produced.

- Along with the empirical contribution of new results on the factors that influence rail journey times in Chapters 6 and 7, a methodological contribution is also made. The majority of the regression-based studies on journey times involve the application of linear regression methods. This is a simplified approach which limits the relationship between the independent and dependent variables, and can result in the omission of important data. To correct for this, semiparametric regression is applied. This method retains all data information, and is able to model flexible relationships between the variables based on the data points themselves. Parameter estimates are then able to be generated with a higher degree of accuracy compared to the conventional linear models.

- The influence of train headways on passenger wait times is an important area of research relevant in transit planning applications, and the majority of work on quantifying passenger wait times is undertaken either using linear regression methods, or descriptive distribution fitting methods. For the linear regression methods, wait time is set as the dependent variable, and the influence of observed train headways and other factors such as the regularity of train headways, the time of day, and other passenger movement behaviours are nominated as independent variables. In Chapter 6 of the thesis, the semiparametric regression model of access time is used to compute marginal wait times by taking the derivative of access times with respect to train headways. Conclusions are able to be drawn regarding passenger arrival patterns under different train operating frequencies. To current knowledge, this is the first application of advanced regression techniques coupled with the mathematical derivation of marginal platform wait times from passenger access times, while conditioning for other service supply and passenger demand factors a priori.

- The semiparametric regression models focusing on network supply and demand characteristics in Chapter 6 are extended to quantify the effect of individual passenger characteristics on journey times in Chapter 7. Literature on passenger characteristics focuses on capturing effects that are descriptive of passenger attributes, rather than at an individual-specific level. In empirical work on total transit journey times, dwell times, and platform wait times, examples of modelled passenger categories include: gender, age, ethnicity, number of persons per household, and passenger walking behaviours. Quantifying heterogeneity at an individual passenger level using semiparametric regression methods has not been previously undertaken, and this is the key contribution of Chapter 7.

Note: The contributions to the literature for the proposed passenger to train assignment method are outlined separately after the review of the existing assignment methods in Chapter 5.

Chapter 3

Data

The following chapter provides a summary of the data sets that are used in the analyses presented in Chapters 4, 5, 6, and 7 of the thesis. The data include information on passenger trips, train movements, and incident events on the London Underground collected from automated data systems, as well as information on network infrastructure from station plans and network databases. The chapter begins with a definition of the geographical analysis boundaries on the London Underground network, followed by a summary of the data on passenger trips, train movements, incidents, and other network properties. Descriptive statistics are presented, along with a discussion of the limitations of each data set.

A number of plots are also included in the chapter to illustrate the distributional properties of the data. As discussed in the literature review, the plots are generated through non-parametric kernel density estimation. The method involves the application of a kernel density function corresponding to a normal distribution and a smoothing parameter to capture the trade-off between the degree of matching the data points and the smoothness of the function (Scott, 2015).

3.1 Definition of London Underground study area

The analysis of journey times on the London Underground focuses on trips in central London's zone 1 undertaken on four lines: 1) the Victoria line, 2) West Acton to Oxford Circus on the Central line, 3) Bond Street to North Greenwich on the Jubilee line, and 4) the Waterloo and City line. The geographical boundaries of the analysis are illustrated in Figure 3.1, and a full map of the London Underground is included in Appendix B for reference. Within these extents, only single-line trips are assessed, i.e. trips that originate and terminate on the same line. Two important exclusions are made: 1) the trips selected for the analysis do not include interchange movements, and 2) the trips do not require consideration of route choice, as the origin-destination (OD) pairs are selected such that there are no other probable alternative routes under normal operating conditions. Table 3.1 summarises the number of stations and unique OD pairs that are included in the analysis by line and direction. In Chapter 4 of the thesis, all four lines are analysed; however, in Chapters 5, 6, and 7, the Waterloo and City line is not able to be included as a consequence of incomplete train movement data (this is discussed further in section 3.3).

Table 3.1: Study area - number of stations and unique OD pairs

Line	Direction	No. stations	No. OD pairs
Central	East & westbound	12	66
Jubilee	East & westbound	10	45
Victoria	North & southbound	16	120
Waterloo & City	East & westbound	2	1
	Total	40	464

3.2 Passenger trip data

3.2.1 Data summary

The Oyster smart card is a medium of automated payment administered by Transport for London (TfL) for travel on bus, rail, and ferry public transport systems in London. On the London Underground metro system, automated card readers are installed at the entries and

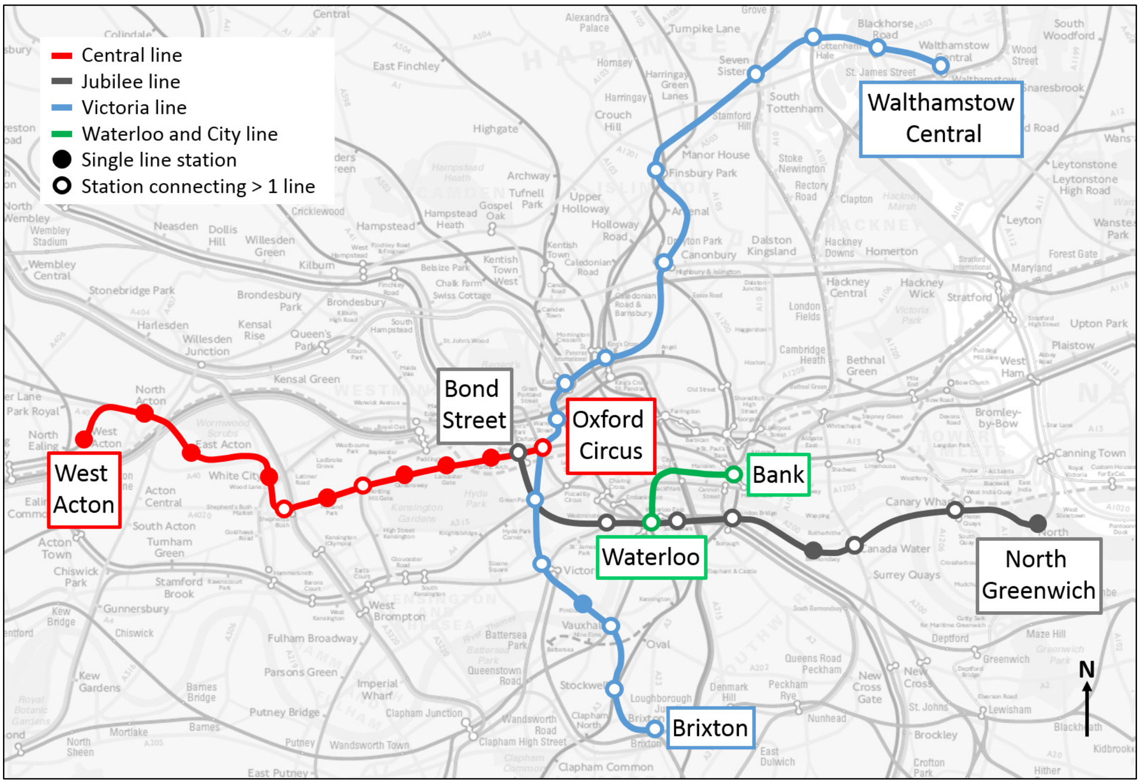

Figure 3.1: Study area, London Underground (adapted from Transport for London (2014a))

exits of each station. To travel on the system using the Oyster card, passengers must tap-in at the card readers when entering at the origin station, and tap-out at the card readers when exiting at the destination station. The majority of the 270 stations that comprise the network have gated entries and exits, and the card readers are mounted on the gates. A total of 48 stations typically located in the outer suburban areas of the network are not gated, but have stand alone card readers either at the platforms or close to the station entry/exit points (Chan, 2007).

The current pay-as-you-go format for the Oyster system was first introduced in 2004 (Verma, 2010). This payment format treats the Oyster card as a digital purse, where money can be added to the card at any time prior to travel. The fare payment is taken as the passenger taps out at the destination station, and the fare amount is governed by the distance of travel under the TfL zonal system (refer to Appendix B for a full zonal map of the London Underground). TfL collects and retains records of all transactions made with the Oyster card. In this research, a data set covering a seven week time period from 27th October 2013 to 14th December 2013 is used. The data set includes all transactions made on the London Underground, as well as transactions on the London bus network, the Docklands Light Rail network, and the National Rail network within the Greater London region. In the raw data files, a total of 85 attributes are recorded, including both trip based and non-trip based transaction data. The non-trip records include transactions where value has been added to the Oyster card, card balance enquiries, and incorrect card use, amongst others. In this research, information is required on the trip based transactions only, where complete tap-in and tap-out records at the origin and destination stations are provided. After filtering the data for complete trips on the London Underground, an average of approximately 4 million journeys on weekdays and 1-2 million journeys on weekends are observed across the network.

In the analysis, the following attributes recorded by the Oyster system are used: the tap-in and tap-out dates, times, and locations for each trip, and passenger-specific information including the passenger age category, card type, and discount type. In addition to the raw data records, a number of parameters are derived to further categorise the data and to enable compatibility with the other data sets used in the analysis. The raw and derived passenger trip attributes are summarised in Table 3.2.

Table 3.2: Trip data set attributes

Attribute name	Unit	Description
Card number	Number	Unique Oyster card number
Card type	Number	Category denoting card type
Age category	Number	Category denoting passenger age category associated with card
Discount type	Number	Category denoting any discounts applied to card
Date	Number	Date in numeric TfL format
Origin station	Number	TfL NLC code to identify entry station
Entry time	Mins after midnight	Time of tap-in at entry station in TfL format of minutes after midnight
Destination station	Number	TfL NLC code to identify exit station
Exit time	Mins after midnight	Time of tap-out at exit station in TfL format of minutes after midnight
OD	Number	Code assigned to identify the origin-destination pair
Line	Number	Code assigned to identify which line is taken
Direction	Number	Code assigned to identify direction of travel along the line
Journey time	Mins	Journey time calculated as $(exit\ time) - (entry\ time)$

In Table 3.2, the tap-in and tap-out timestamps are recorded in an internal TfL format of minutes after midnight, and the date codes are also in an internal TfL numerical format ranging from a value of 12354 representing 27 October 2013 to 12400 representing 14 December 2013. The stations are also assigned unique numerical identifiers, which are documented in the TfL National Location Codes (NLC) database. A list of the codes for the 40 stations included in the analysis is provided in Appendix C for reference. For consistency, the numerical format for the date, time, and station fields are retained and used across all data sets in the analysis.

Along with the recorded data, three additional fields are derived to further categorise the trips: at an origin-destination pair level, at a line level, and at a direction level. Unique numerical identifiers are defined and assigned for each possible origin-destination pair within the analysis extents. TfL uses internal numerical codes to define the lines and direction of travel along the lines. For consistency, the same codes are used to identify travel on the four lines by direction; these are documented in Appendix C for reference. An additional field of data is also included to record the journey time between OD pairs. The passenger journey time is defined as the total travel time between tap-in to tap-out, and it is computed by subtracting the recorded tap-out time at the destination station from the tap-in time at the origin station.

Passenger-specific characteristics are also recorded in the Oyster system. In the analysis, information on passenger age, card type, and discount type is used to further understand passenger travel behaviour. TfL uses numerical codes to define the passenger categories within each of these fields. A summary of the different categories for the passenger-specific attributes is given in Table 3.3.

3.2.2 Descriptive statistics

Within the analysis extents along the four lines defined in section 3.1, trips made during weekdays only are analysed. In order to reflect the typical operating conditions of the network, the data are further filtered to capture trips undertaken between the core operating hours of 6am and 12am. The data are then cleaned to eliminate erroneous records, including duplicate rows, trips with negative journey times, and trips where the tap-in and tap-out locations reference the same station. A total of 308 trips (0.003%) are discarded through the cleaning process,

Table 3.3: Passenger data categories

Passenger age	Card type	Discount type
Adult	Oyster Card (Retail)	No discount/NA
Child	Staff Pass	Under 11
Non-Concession Child	Staff Nominee Pass	16+ (Legacy)
Youth	Bus Operator Pass	Young Persons railcard
16-17	Bus Operator Nominee Pass	Student
Sixty Plus	Retired Pass	Full Time Education
	Disabled Freedom Passes	Senior railcard
	Elderly Freedom Passes	Concession
	Oyster Photocard	HM Forces Railcards
	Visitor Oyster Card	Privilege
	Barclaycard OnePulse	Disabled persons railcards
	Barclays Debit	Bus & Tram Discount
		NR Railcard
		Network railcard
		Travelcard
		Annual Gold card
		Group
		Visitor
		Family
		Family railcard

leaving a total of approximately 9,665,000 trips. Descriptive statistics of the passenger journey times by line and direction are given in Table 3.4.

Table 3.4: Summary statistics for passenger trips

			Journey times (mins)			
Line	Direction	No. trips	Min.	Max.	Mean	s
Central	Eastbound	544451	2.0	132.0	14.1	5.8
	Westbound	507094	2.0	142.0	14.4	5.8
Jubilee	Eastbound	1359457	2.0	157.0	13.2	5.9
	Westbound	1534780	2.0	136.0	13.3	5.6
Victoria	Northbound	2596675	2.0	256.0	16.0	8.1
	Southbound	2884964	2.0	259.0	16.0	8.4
Waterloo & City	Eastbound	82941	3.0	84.0	9.0	3.0
	Westbound	154641	3.0	87.0	7.5	2.9
Total		9665003	2.0	259.0	14.8	7.4

Note: s is the sample standard deviation.

As shown in Table 3.4, the average trip time across all lines is 14.8 minutes. Trips on the Waterloo to City line have the shortest average journey times, while journey times on the Victoria line are on average longest. Figure 3.2 shows the distribution of passenger entry times by time of day. In terms of passenger fare differentiation, TfL defines the official peak operating periods as 6.30-9.30am for the morning peak and 4-7pm for the afternoon peak. It should be noted that in terms of passenger demand, TfL uses 7am-10am as the morning peak time of demand, and 4-7pm as the afternoon peak time of demand. Across all lines, Figure 3.2 shows that trip entries are indeed concentrated in the times that correspond to the morning and afternoon peak periods. Compared to the other lines, the Waterloo and City line and Central line westbound appear to have more direction-based peak periods. The majority of passengers on the Waterloo and City line travel east in the morning, and west in the afternoon. Passengers travelling westbound on the Central line are also concentrated in the afternoon peak. The commercial centre of the city, the City of London, is located to the east of the city, and so these trips are likely capturing regular commuters travelling for work purposes.

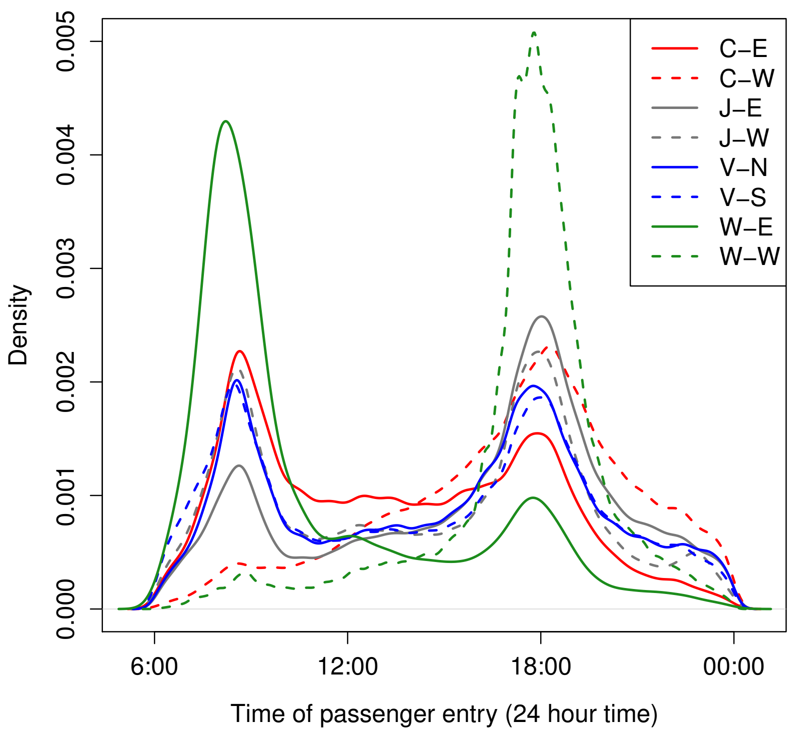

Figure 3.2: Passenger entry times

3.2.3 Limitations

The Oyster data provide a rich source of information on passenger trips, however, there are some limitations, which are discussed in the following section.

One of the most prominent limitations relates to the accuracy of the timestamps. As previously mentioned, the tap-in and tap-out times are recorded in the format of minutes after midnight. The times are truncated at a minute level, eliminating the decimal numbers which represent seconds. Under this format, the true tap-in/tap-out time can occur at any time between HH:MM:00 and HH:MM:59. And so, as a result of the truncation, the timestamps at stations are accurate within a window of 0 to 59 seconds. Considering the level of uncertainty in the timestamps at the origin and destination for a complete journey, the total journey time has a level of accuracy within a window of 0 seconds to 1 minute 58 seconds. This implies that for all trips recorded, the actual journey times may be up to almost 2 minutes longer. For the analyses that involve merging multiple time-based data sets, discrepancies in the differing levels of timestamp accuracy arise. The way in which this limitation is treated is further discussed in the relevant analysis chapters.

In meetings with TfL and consulting past academic work using the Oyster data, it has been raised that there may be potential errors in the time synchronisation of the card readers at stations across the network. As a result, the recorded tap-in and tap-out timestamps may be inaccurate, with Paul (2010) reporting that the error may be in the range of several minutes. The exact degree of the errors and exact locations of the non-synchronised readers is not known; and so these potential measurement errors would be captured in the random error component of the statistical models developed in the analyses.

As other ticketing mediums such as single paper tickets and periodic travel card passes were available in 2013, it should be noted that the Oyster card transactions capture only a proportion of all trips made on the London Underground. Although statistics are not available for 2013, in 2007, TfL estimated that approximately 70% of all trips on the London Underground were made via Oyster card (Chan, 2007). Another consideration to take into account is that persons who use the Oyster card would more likely be regular users who may move quicker through the system as a result of route familiarity. As a result, the Oyster trip data should be considered as a sample of more regular users of the London Underground, rather than a sample that accurately represents the entire population of users.

Another commonly raised limitation of the Oyster data and other similar tap-in/tap-out fare collection systems is that the only known route information is the location of the origin and destination nodes. If a passenger travels on a route that involves an interchange, this information is not recorded as the Oyster system does not require additional tap-in at intermediate stations between the origin and destination stations. And so, for routes that involve interchange or for OD pairs that have more than one possible routing option, the true route taken between the origin and destinations cannot be known. Therefore, the analysis extents have been deliberately chosen to eliminate the possibility of ambiguity in routing. All OD pairs within the extents of the analysis are not likely to involve an interchange and are not likely to have more than a single feasible routing option under normal operating conditions.

3.3 Train movement data

3.3.1 Data summary

Data on train movements are recorded in the TfL Network Management Information System (NetMIS) database. The records are extracted from the London Underground signalling system and contain information on every train movement at each station platform as registered by the track circuits. The track circuits closest to the station platforms are used to derive the arrival and departure times of the train at the platform. The track circuit system records the time at which the train is occupying the section of track adjacent or closest to the station platform, and then a given time offset is added or subtracted to determine the exact time the train arrives at and departs from the platform (Paul, 2010). The NetMIS database records each time a train arrives and departs at a station platform; this series of movements is termed a train event. In the raw database files, a total of 37 different attributes are recorded for each train event. In the analysis, information on only six fields is required. The relevant data attributes include the station, line, and direction of travel, as well as the date, the arrival and departure timestamps, and unique identifiers for the train number and trip number of the train.

Data cleaning and processing is undertaken to derive the attributes required for the analysis and to ensure compatibility with the Oyster trip record data. Table 3.5 provides a summary of the relevant raw and derived attributes for the train movement data. The station names are recorded as three letter codes in the NetMIS database, however, the letter codes are transformed to the numeric codes allocated for each station as described in section 3.2. The arrival and departure timestamps for each train event are recorded is HH:MM:SS format, whereas in the Oyster trip data, the timestamps are recorded in the internal TfL format which measures time in minutes after midnight. For compatibility, the NetMIS timestamps are transformed to the number of minutes after midnight. It is important to note that in order to maintain accuracy, the seconds information in the transformation is retained.

Table 3.5: Train data set attributes

Attribute name	Unit	Description
Date	Number	Date in numeric TfL format
Line	Number	Code assigned to identify which line is taken
Direction	Number	Code assigned to identify direction of travel along the line
Train number identifier	Number	Code assigned to identify the train
Train trip identifier	Number	Code assigned to identify the trip number of the train (i.e. the first train run is 1, and this incrementally increases by 1 every time the train turns around at the end of the line)
Origin station	Number	TfL NLC code to identify entry station
Origin departure time	Mins after midnight	Time at which train departs from the origin station platform in TfL format of mins after midnight
Destination station	Number	TfL NLC code to identify exit station
Destination departure time	Mins after midnight	Time at which train departs from the destination station platform in TfL format of mins after midnight
OD	Number	Code assigned to identify the origin-destination pair
On-train time	Mins	Train run time calculated as $(destination\ departure\ time) - (origin\ departure\ time)$
Headway	Mins	Headway between train calculated as $(origin\ departure\ time) - (origin\ departure\ time\ of\ previous\ train)$

The headway between trains is a key input in the train assignment analysis, and this is able to be computed from the NetMIS data. For each possible OD pair, the time at which the train arrives and departs at the origin and destination stations is identified by matching the train number and train trip identifiers at each station. To keep track of the matching origin and destination station timestamps, the timestamps are grouped and the unique OD code for the pair of stations is allocated. The data are then filtered by OD, line, and direction in order to obtain all train movements for the given OD pair. The train event records are ordered by date and time to reflect the exact order in which trains arrive at the origin station. The headway between consecutive trains is then computed for each train event by subtracting the train departure timestamp for the preceding train event from the train departure timestamp for the given train event under consideration. For the first train event of the day, the headway is set as “NA”. The train run time, herein referred to as “on-train time”, is also computed for each OD route by subtracting the departure timestamp at the destination station from the departure timestamp at the origin station.

3.3.2 Descriptive statistics

Complete train movement data for the required fields are available for the Victoria, Central, and Jubilee lines. The train movement records for the Waterloo and City line, however, are incomplete and as a result, the line is not able to be included in the analyses that involve passenger to train assignment inputs. Further details of the incomplete records are given in the discussion of the data limitations in section 3.3.3.

Summary statistics for on-train times and headways by line and direction for the typical network operating hours of 6am to 12am are given in Table 3.6. Figure 3.3 shows the distribution of train departures by time of day for each line and direction. The distributions have distinct peaks corresponding to the morning and afternoon peak travel times, showing that a higher proportion of trains are in operation to meet higher levels of passenger demand at these peak times (refer to distributions of passenger trips by time of day in Figure 3.2).

Table 3.6: Summary statistics for train movement data

			On-train times (mins)				Headways (mins)				No. trains per hr	
Line	Direction	No. trips	Min.	Max.	Mean	s	Min.	Max.	Mean	s	Min	Max
Central	Eastbound	787869	0.02	525.2	8.2	5.5	0.08	244.2	3.1	3.2	2	35
	Westbound	809826	0.02	44.1	8.4	5.5	0.25	241.8	3.0	3.2	2	35
Jubilee	Eastbound	614748	0.03	886.6	7.4	5.5	0.05	249.6	2.7	3.1	2	30
	Westbound	608895	0.02	717.9	7.3	5.3	0.05	244.4	2.8	3.0	2	29
Victoria	Northbound	1247256	0.02	725.6	11.3	8.6	0.02	747.2	3.1	6.5	1	33
	Southbound	1387354	0.20	178.7	11.2	7.5	0.03	691.1	2.8	5.2	2	33
Total		5455948	0.02	886.6	9.5	7.0	0.02	747.2	2.9	4.7	1	35

Note: s is the sample standard deviation.

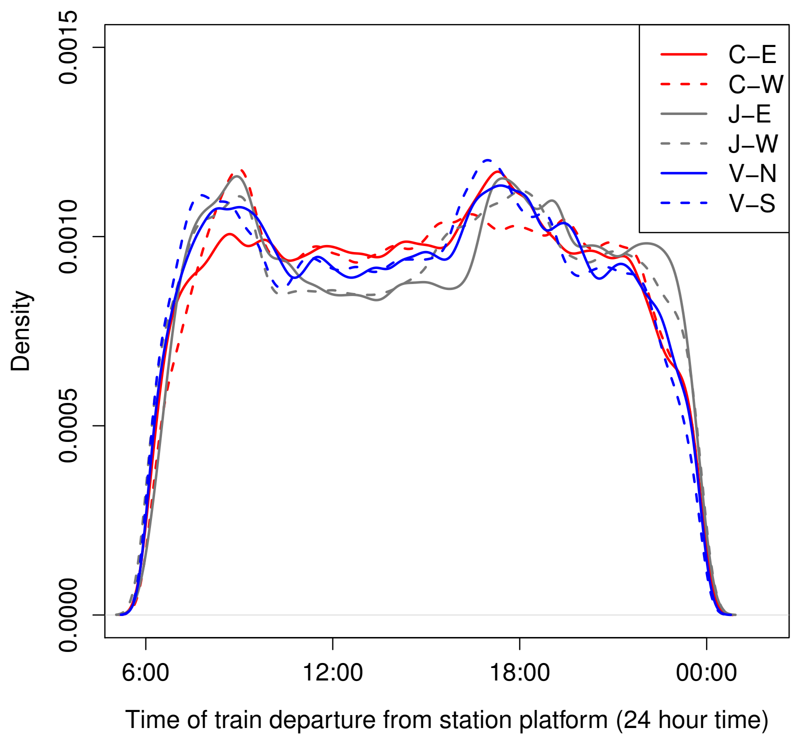

Figure 3.3: Train departure times

Although timetabled information on the exact arrival times of each train is not available, the published operating train frequencies for each line are available from the TfL Working Train Timetables for October to December 2013. The minimum and maximum train frequency per hour is summarised in Table 3.6 by line and direction. The Central line has the highest train frequencies, with 35 trains scheduled per hour in the morning peak. This corresponds to a minimum operating headway from 1.67-1.71 minutes between trains, depending on whether trains arrive on the hour or one headway later. The lowest frequency across all lines is 1 train per hour on the Victoria line northbound, implying a maximum operating headway between 30 and 60 minutes, depending on when the train is timetabled. As seen in Table 3.6, the minimum and maximum headways from the train movement data exceed the published minimum and maximum operating headways. To further investigate the distribution of train headways, a heat-map of the train headways aggregated across all lines by time of day is shown in Figure 3.4, and the empirical density function of all headways aggregated across the lines and directions is plotted in Figure 3.5.

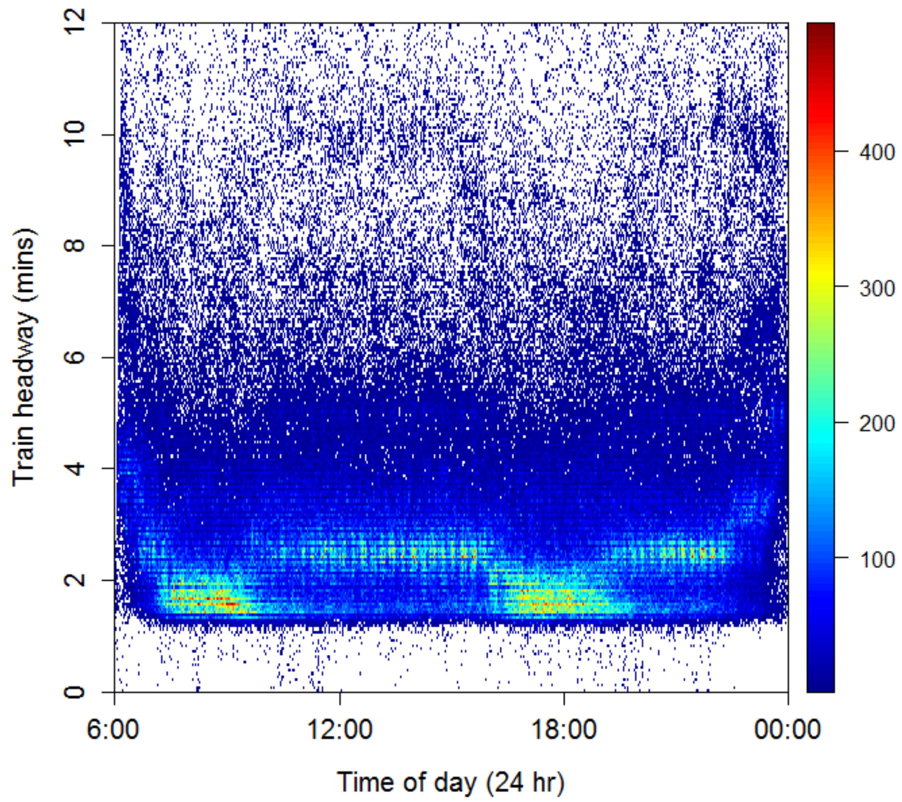

Figure 3.4: Train headways by time of day

Figure 3.4 shows that the majority of trains across the three lines operate between approximately 1.6 and 6 min headways across the day, however, there are outlying values as indicated by the more sparse data points in the plotting regions below a headway of 1.6 minutes and above 6 minutes. In Figure 3.5, the distribution of headways is further illustrated for the entire data set, and the minimum published operating headway of 1.67 minutes is also illustrated by the vertical dotted line. The plot shows that although the majority of trains operate at headways at or greater than 1.67 minutes, there are a number of trains than operate at headways shorter than the minimum published headway. Out of all train events, approximately 15% of trains operate at headways less than the minimum published headway of 1.67 minutes. This implies that either there are errors in the train movement records, or the trains may be operating at shorter headways due to passenger congestion. Under congested conditions, trains may have longer dwell times at platforms in order to accommodate higher volumes of boarding and alighting passengers. A longer dwell time from a preceding train leads to a shorter headway

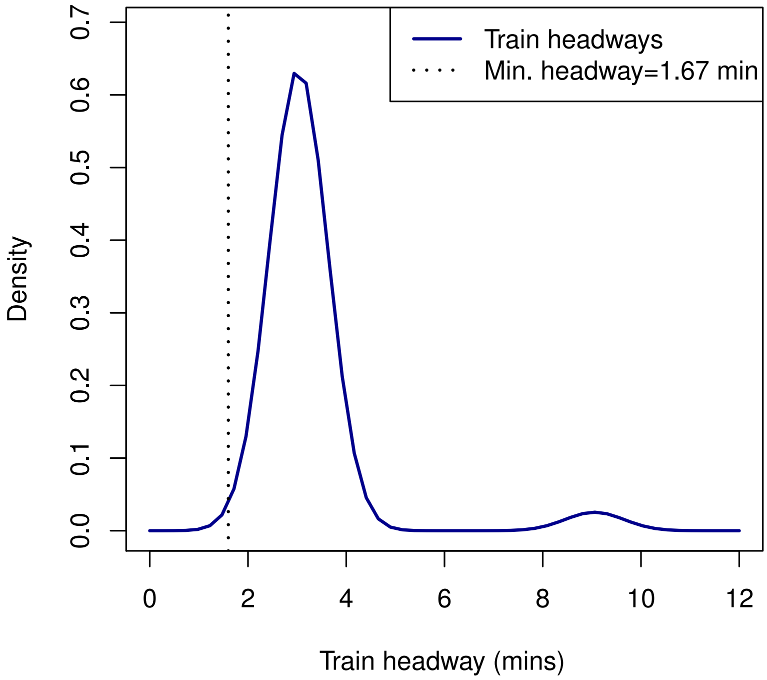

Figure 3.5: Distribution of train headways

between trains if the next train arrives as per the published schedule. In highly congested conditions, multiple trains may have longer platform dwell times, and this can therefore lead to train bunching, where multiple trains operate at headways shorter than the planned headway (Kittelson & Associates Inc et al., 2013).

In terms of the upper bound of headways, although the published maximum operating headway is between 30-60 minutes, 99.5% of all train routes operate at headways of 20 minutes or less. As shown in Table 3.6, there are outlying values where trains operate at headways longer than 60 minutes, however, this represents only 0.3% of the train event records. These long values may arise out of either erroneous data recording or the system may be operating in a delayed state as a result of severe incident event/s.

In normal operating conditions in line with the published train frequencies, the majority of trains operate at headways between 1.67 and 10 minutes. As shown in Figure 3.4, a high concentration of trains operate at headways between 1.6 and 2 minutes during the morning and afternoon peak times. During the early morning, midday inter-peak, and late evening times, the majority of trains operate at slightly longer headways between 2 and 3 minutes.

3.3.3 Limitations

Compared to the Oyster trip data, the train movement data records are less complete, and the missing data fields are the main limitation of the train movement data set. In past work using the London Underground train movement data, researchers have documented that the missing records are a result of ageing infrastructure. The servers that collect the train occupancy information from the track circuits are not always synchronised between track sections, and so information is lost (Freemark, 2013; Paul, 2010). As mentioned in section 3.3.1, the Waterloo and City line is not able to be included in the analysis as a result of missing information. The majority of the train numbers and train trip identifiers are not recorded. These attributes are required to identify and group the origin and destination stations belonging to the unique train route. As the values are missing, the line must be excluded from the analyses that require train movement data.

Another field of missing information is the train arrival time. Timestamps are more consistently provided for the departure times of trains at platforms, however, the majority of the train arrival times are missing. In order to maintain consistency in the analysis methods, only the departure timestamps are used. As a consequence, the dwell times of the trains at platforms are not known. The implications of this are further discussed in the passenger to train assignment method described in Chapter 5.

A final point on the limitations of the train data relates to the method in which the train timestamps are obtained. As mentioned in section 3.3.1, the timestamps are not recorded at the exact time a train reaches a platform. Rather, the timestamps are a derived quantity where the train arrival and departure times are inferred based on the time at which the train passes the track circuit closest to the platform, and then an offset time is added or subtracted to obtain the arrival/departure time at the platform. The values of the offsets used are not available, and so any errors in these are captured in the random error component of the statistical models that are developed in the analyses.

3.4 Incident data

3.4.1 Data summary

Operational delays caused by incidents on the London Underground are documented in the Contract Performance Information Database (CuPID). For each logged incident event, information is recorded on the following: the date, time, and location of the incident, a detailed description of the incident event, the amount of operational delay incurred in minutes and Lost Customer Hours, categorisation of the incident into a cause category, and attribution of incident responsibility to the appropriate department within TfL.

The duration of delays from incident events and the Lost Customer Hours metric are calculated metrics which are based on historical incident data, historical passenger flow data, and the London Underground Train Service Model (refer to section 2.3.3 in Chapter 2 for a more detailed discussion). The initial amount of delay is measured by the amount of operational disruption (in minutes) to the relevant components of the network. The duration of incident delay represents the time for the system to recover to normal operating conditions, and this quantity is estimated based on calculations of passenger flow and/or train movement by incident type, time of day, date, and location. The Lost Customer Hours metric represents the overall severity of the incident by summing the total amount of delay across the total number of passengers that are affected. The incident cause and disruption type categories are listed in section 2.3.3 in Chapter 2.

To ensure compatibility with the passenger trips and train movement data, the incident data are cleaned and filtered to obtain the required information. Table 3.7 summarises the relevant attributes of the resulting data set. As with the train movement data, the incident data timestamps are recorded in HH:MM:SS format. This format is transformed to the same format used in the Oyster trip data, where time is recorded as minutes after midnight. As with the train movement data, the seconds information is retained to maintain accuracy. If the incident occurs at a station, the location is recorded as the station name in lettered characters. To be compatible with the other data sets, the station names are transformed to the numeric codes used by TfL in their station National Location Code (NLC) database. As previously mentioned, the total amount of operational delay incurred as a result of an incident event is recorded in two components: the initial delay and the remaining duration of delay. The total delay of the incident event is computed as a sum of the initial delay and delay duration to obtain an aggregate value of delay.

Table 3.7: Incident data set attributes

Attribute name	Unit	Description
Incident number	Number	Unique code to identify the incident event
Date	Number	Date in numeric TfL format
Line	Number	Code assigned to identify which line is taken
Direction	Number	Code assigned to identify direction of travel along the line
Time	Mins after midnight	Time to denote the beginning of the incident event TfL format of mins after midnight
Initial delay	Mins	Measured initial disruption to network operations
Delay duration	Mins	Estimated time for network to recover to normal operating conditions
Total delay	Mins	Total amount of delay calculated as (*initial delay*) + (*delay duration*)
LCH	Hours	TfL metric of calculated Lost Customer Hours
Cause category	Number	TfL code to identify incident cause category
Disruption type	Number	TfL code to identify disruption type category

3.4.2 Descriptive statistics

Incident data is recorded for all four lines in the analysis. The summary statistics for the total delay and Lost Customer Hours by line and direction are given in Table 3.8. As shown in Table 3.8, the Victoria line has the highest number of incident events out of all lines, though the average severity of the incidents as measured by the total delay and Lost Customer Hours metrics does not rank highest. On average, incidents that occur at the stations on the Waterloo and City line cause the greatest amount of total delay. In terms of the Lost Customer Hours metric, on average incident events have the most severe effect at the stations on the Jubilee line. To further investigate the severity of the incidents, Figure 3.6 shows the distribution of total delay by line and direction. As shown by the peaks of the distributions, most incident events across all lines cause a total delay between 3 and 5 minutes.

Table 3.8: Summary statistics for incident events

Line	Direction	No.	Total delay (mins) Min.	Max.	Mean	s	Lost Customer Hours (hours) Min.	Max.	Mean	s
Central	Eastbound	49	2.0	15.0	4.6	2.6	106.0	6162.2	1060.5	1177.5
	Westbound	33	2.0	259.0	24.4	55.2	28.0	2036.0	387.7	464.0
	At station	30	0.0	182.0	7.4	33.5	4.0	8063.0	357.0	1467.1
Jubilee	Eastbound	49	0.0	13.0	3.9	2.4	0.0	7308.0	1243.6	1551.9
	Westbound	38	2.0	15.0	4.6	2.9	22.0	12503.0	1700.5	2680.8
	At station	67	0.0	57.0	1.2	7.5	0.0	7293.0	412.4	1332.6
Victoria	Northbound	121	0.0	67.0	4.5	8.4	0.0	7064.0	634.7	907.6
	Southbound	101	0.0	113.0	7.9	17.4	0.0	11364.0	1050.6	1619.2
	At station	51	0.0	277.0	31.5	56.9	0.0	9850.0	450.1	1462.1
Waterloo & City	Eastbound	14	2.0	14.0	4.7	3.5	15.0	2456.0	565.6	768.9
	Westbound	6	0.0	6.0	3.3	2.2	4.0	325.0	141.3	140.1
	At station	14	6.0	181.0	59.4	58.9	1.0	2613.0	297.2	701.3
Total		573	0.0	277.0	9.7	28.4	0.0	12503.0	780.9	1451.2

Note: s is the sample standard deviation.

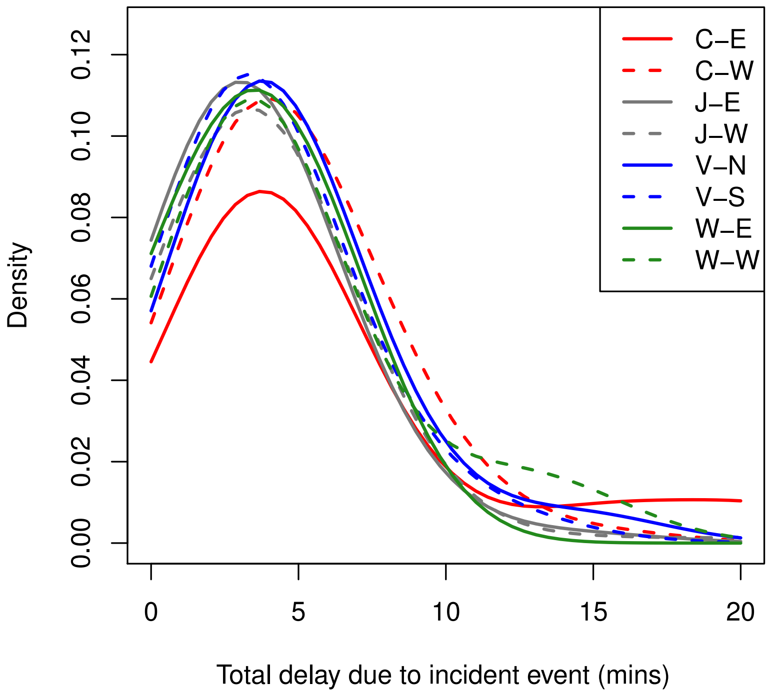

Figure 3.6: Total delay due to incident events

Figure 3.7 shows the distribution of incident events by time of day. Across all lines, incident events are most concentrated during the morning and afternoon peak operating times. At these times of the day, both passenger demand is highest and trains operate at their highest frequencies. Due to the greater amount of activity at these times, it is expected that there is a corresponding higher likelihood of incident events occurring within the peak travel periods.

3.4.3 Limitations

The main limitation of the incident data set is that the total delay and Lost Customer Hours metrics are calculated quantities based on historical incident and passenger volume records. The actual effects on passenger journey times and train run times in real-time are not measured and recorded in the CuPID database. As a consequence of this, the derived quantities of incident delay and Lost Customer Hours can only be treated as indicators that represent the relative severity of the incidents, rather than the absolute magnitude. The main application of the

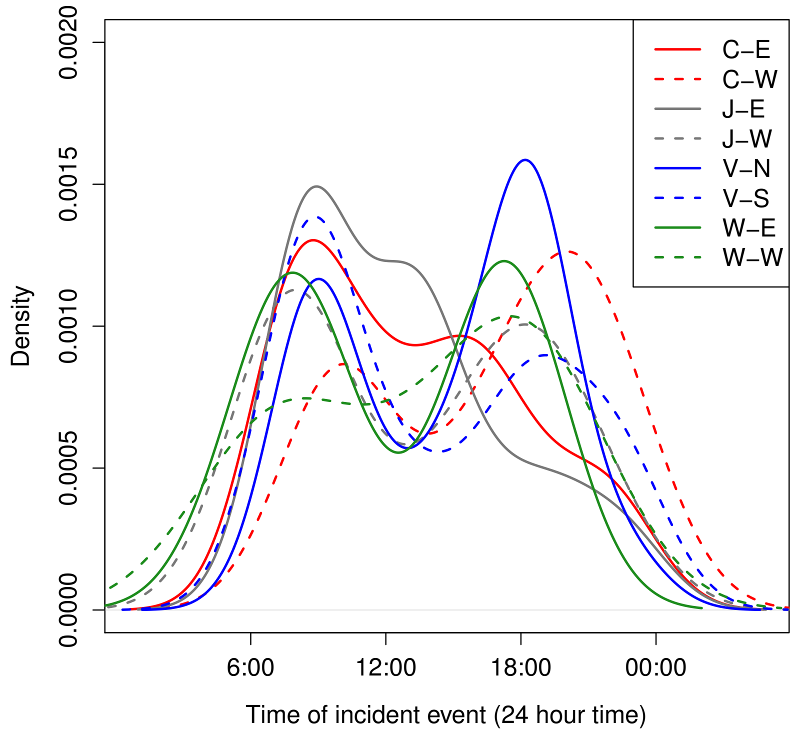

Figure 3.7: Time of incident event

incident database in the analysis is to identify the times and locations at which incident events occur. A separate assessment is then performed of whether there is an impact on passenger journey times.

As with the train movement data, the timestamps of the incident events are recorded to a seconds accuracy. In terms of cross-data compatibility, therefore, the incident data is consistent with the train movement data and the two data sets can be merged as is. The passenger trip data, however, is recorded to a minute level of accuracy. As a result, when matching the incident events with the passenger trip data, allowances must be made for the discrepancies in the level accuracy of the timestamps. The way in which this is undertaken is further discussed in the analysis of the journey times of trips affected by incidents in Chapter 4.

3.5 Additional network data

In the semiparametric regression models presented in Chapters 6 and 7, additional information is required on the physical and operational characteristics of the London Underground at a network and station level. The following sections provide a brief summary of the relevant additional data sets.

3.5.1 Rolling Origin Destination survey data

As discussed in section 3.2, the Oyster trip data set does not represent the total number of passengers that use the London Underground system. For the analysis time period from October to December 2013, ticketing mediums other than the Oyster card were available, and it is conservatively estimated that approximately 30% of all trips were made using fare payment methods other than Oyster (Uniman, 2009; Paul, 2010). Therefore it is not appropriate to use the Oyster data set to estimate total passenger volumes within the network.

An alternative source of passenger volume data is the Rolling Origin and Destination survey (RODs) database. The RODs database combines multiple data sources including passenger travel surveys and manual and automated station gate-line counts. The passenger travel surveys are undertaken at a number of stations within the network, and randomly selected passengers are asked to provide information on the following: the trip purpose, the final destination, the route that will be taken between the origin and destination stations, the time of entry and exit, the ticket type, and other personal travel attributes and preferences. The travel survey gives valuable information on the route choice of passengers. The other set of data in the RODs database is the counts of passengers entering and exiting at stations. For stations that are fully gated at the entry and exit points, the gate counts are directly extracted from the automated gate detection systems. At stations where gates are not installed, manual counts are performed to estimate passenger entries and exits.

The passenger travel survey data and gate counts are then combined to develop a full OD matrix for the entire network, with estimates of passenger flows for each possible and feasible route between all OD pairs on the network. The manual surveys are labour-intensive and expensive to disseminate and process, and so they are not undertaken regularly. To estimate the passenger flows for the most current year of operation, multiple years of data are combined. Therefore, the RODs database can give a good approximation of passenger flows year by year, but cannot be used as an absolute representation of passenger flows in real-time. In the analysis, the RODs data are used to provide an indication of passenger flows within and between stations.

3.5.2 Physical network and station characteristics

In the semiparametric regression models, covariates are included to capture the physical features of the stations and OD routes. Information on these characteristics is obtained from a number of different data sources, including station layout plans, network maps, and attribute-specific databases. From the station layout drawings, manual counts have been performed to determine the following attributes at each station: the number of fare gates at the entry/exit points of the stations, the number of stairs, escalators, lifts, and the number of platform entry/exit points. Additional attributes noted include whether the station platform is located underground or above ground, and whether platform screen doors are installed along the platform. From information recorded in internal TfL station databases, additional information is obtained on platform lengths, and average access and egress walk distances between the midpoints of the entry/exit gatelines and the station platforms.

From the London Underground network diagram, the following attributes are recorded: the fare zone in which each station is located, the number of lines that the station connects to, whether the station is a terminus station of a line, and whether the station connects to the National Rail system or not. The number of intermediate stops and the route distance between each OD pair are calculated from the TfL database that records station chainage locations by line. Table 3.9 provides a list of all data fields along with summary statistics.

Table 3.9: Summary statistics for station and network characteristics

Data attribute	Description	Unit	Min	Max	Mean	s
Gates	No. of gates available for entry/exit	Numeric count	3.0	55.0	16.5	15.2
Escalators	No. escalators available for entry/exit between platform and gates	Numeric count	0.0	20.0	3.7	4.0
Lifts	No. lifts available for entry/exit between platform and gates	Numeric count	0.0	5.0	1.2	1.3
Stairs	No. stair steps for entry/exit between platform and gates	Numeric count	0.0	4.0	1.7	1.1
Platform points	No. platform entry/exit points	Numeric count	1.0	11.0	4.4	2.5
Platform length	Length of platform	Metres	122.0	153.0	136.9	5.0
Walk distance	Distance between midpoints of entry/exit gatelines and platforms	Metres	19.3	400.6	143.8	56.9
Station age	Maximum age of station	Years	15.0	174.0	101.3	45.8
Station zone	TfL zone where station is located	Numeric	1.0	3.0	1.7	0.7
Connecting lines	Number of lines that station connects to	Numeric count	1.0	6.0	1.7	1.1
Open/closed platform	Dummy variable for whether station platform is open air or enclosed underground	Dummy, value 1 or 0				
Platform screen doors	Dummy variable for presence of platform screen doors	Dummy, value 1 or 0				
Terminal station	Dummy variable for whether station is a terminal station	Dummy, value 1 or 0				
Connection to National Rail	Dummy variable if station is connected to national rail or not	Dummy, value 1 or 0				
Intermediate stops	No. of intermediate stops between the origin and destination stations	Numeric count	1.0	15.0	5.0	3.3
OD route distance	The distance between the origin and destination stations	Kilometres	0.4	21.3	6.4	4.6

Note: s is the sample standard deviation.

Chapter 4

Characterising journey time distributions

4.1 Introduction

There has been increasing recognition in the transport industry of the need for performance metrics that capture journey time reliability from a passenger perspective, as opposed to the traditional operator-oriented indicators. As presented in the literature review, the most commonly reported metrics relating to journey time include metrics to measure schedule adherence, headway regularity, platform wait times, and measures of delay (Transport for London, 2019a, 2015; Anderson et al., 2013; Barron et al., 2013; Kittelson & Associates Inc et al., 2013; Furth et al., 2006; Furth and Muller, 2006; Kittelson & Associates Inc et al., 2002). The metrics are typically calculated using data on train movements, and provide either average or aggregate measures of performance from an operational perspective relative to specified target levels of service. As detailed in the literature review, the operator indicators are rarely able to reflect journey time performance from the passenger perspective.

The journey time distribution on a route provides a better representation of the passenger experience. A journey time distribution can be constructed by collating multiple travel times between a given origin and destination; the form and moments of the distribution provide useful information about the travel conditions. The form of the distribution can be influenced by a number of factors including: passenger demand levels, service attributes such as frequency, infrastructure provisions, the occurrence of service delay incidents, and the level of temporal and spatial aggregation over which travel times are analysed (Ma et al., 2016; Susilawati et al., 2013; Li et al., 2006; Taylor, 1999). In recent academic literature, metrics based on the distribution of journey times have been nominated as improved alternatives to the operator-oriented metrics, including measures of statistical range (Van Lint and Van Zuylen, 2005; NCHRP, 1997), buffer time metrics (Wood, 2015; Schil, 2012; Uniman et al., 2010; Uniman, 2009; Chan, 2007; Cambridge Systematics Inc. and Texas Transport Institute, 2005), tardy trip metrics (Lomax et al., 2003), and probabilistic measures (Van Lint et al., 2008).

As demonstrated in the literature review, much of the research on journey time distributional form has focused on road travel, with very limited work on rail transit. In general, both road and rail transit journey times typically follow a unimodal right-skewed form under regular operating conditions (Buchel and Corman, 2018; Li et al., 2017; Duran-Hormazabal and Tirachini, 2016; Guessous et al., 2014; Kieu et al., 2014; Fosgereau and Fukuda, 2012; Taylor and Susilawati, 2012; Taylor, 1999; Dandy and McBean, 1984; Taylor, 1982). In congested conditions, and depending on the degree of spatial and temporal aggregation, other forms including the normal distribution and multimodal distributions have been observed (Rahman et al., 2018; Ma et al., 2016; Kieu et al., 2014; Susilawati et al., 2013; Barkley et al., 2012; Jun, 2010; Li et al., 2006).

In this chapter, the gaps in empirical research on metro journey time distributions are addressed. A new data set from the London Underground is used, and journey time distributions are parametrically defined under both regular and incident-affected operating conditions on different lines. The passenger-based metrics of journey time performance presented in the literature are built upon, and new measures are proposed that both capture the passenger experience and provide useful information to operators regarding service quality. The objective of the chapter is to characterise and quantify journey times and journey time variance at a line level on the London Underground, under regular and incident-affected operational conditions. A two part analysis is presented as follows:

1. First, empirical journey time distributions are generated and parametrically defined to determine whether different forms are observed over different lines and directions, over different times of the day, and under regular and incident-affected operations.

2. Second, new metrics are proposed based on the moments of the distributions to provide improved measurement of journey time performance.

The chapter is organised as follows: a brief summary of the study area, the passenger trip data, and incident data is provided in section 4.2. The outline of the methods used in the distribution fitting process and definition of the proposed performance measures are given in section 4.3. Results for the regular and incident-affected journey time distributions are presented in section 4.4, and conclusions are summarised in section 4.5.

4.2 Study area and data

Journey times on four lines on the London Underground are analysed: the Central, Jubilee, Victoria and Waterloo and City lines. The study area is defined as follows: the entire length of the Victoria, and Waterloo and City lines, the central portion of the Central line from West Acton to Oxford Circus, and the central portion of the Jubilee line from Bond Street to North Greenwich (see Figure 3.1, Chapter 3 for geographical extents). Journey times are analysed by line and bidirectionally, and only single line trips that originate and terminate on the line of interest with no other probable routes under normal conditions are selected.

Weekday travel data from Monday to Friday are analysed over a seven week period from October to December 2013. Automated fare collection (AFC) data from the Oyster smart card system provide the tap-in and tap-out times and locations of individual passenger trips. To understand how incidents affect the distributional properties of journey times, incident data from the CuPID database covering the same time period and spatial extents as the regular journey time analysis is used. The incident data set includes a record of the date, time, and locations of each incident, as well as the duration of delay, incident severity, attribution of incident cause, and a detailed description of each incident event. In the analysis, incident events are categorised by incident severity, and this is further discussed in section 4.3.

As mentioned in Chapter 3, there is a discrepancy between the accuracy of the passenger trip timestamps and incident data timestamps. The incident data are accurate to 1 second in the form HH:MM:SS, while the passenger trip data are accurate to 1 minute in the form HH:MM. In order to be conservative, the assumption is made that the passenger could be affected by the incident event any time within the minute. This assumption therefore leads to an increase in the possible number of trips that are considered incident-affected. According to the truncated trip timestamps, a passenger can arrive at the origin/destination station at any time between HH:MM:00 and HH:MM:59. If the incident event timestamp is recorded at some time within this range at HH:MM:XX, the assumption is made that if a passenger timestamp coincides with the hours and minutes HH:MM of the incident event, the trip is included in the set of incident-affected trips. Further detailed information on the specifications and limitations of the trip data and incident data are presented in Chapter 3.

The kernel density distributions of passenger trips and incident events by time of day are illustrated in Figure 4.1. As shown in the plot, the occurrence of incident events coincides with the morning and evening peak times of travel, with more incident events observed during the morning peak. The number of passenger trips and incidents by line and direction are summarised in Table 4.1. In this analysis, trips are considered to be incident-affected if the trip commenced within an interval of 15 minutes from the beginning of the incident event. Spatially, the region of influence of an incident is taken from the recorded origin location of the incident to the nearest adjacent station and all downstream stations. The definition of the incident-affected distributions is further discussed in section 4.3. Overall, approximately 0.5% of all trips are considered to be incident-affected. The Waterloo and City line eastbound has the highest percentage of incident-affected trips at 1.05%, and trips on the Central line westbound are the least affected with 0.17% of trips considered to be incident-affected.

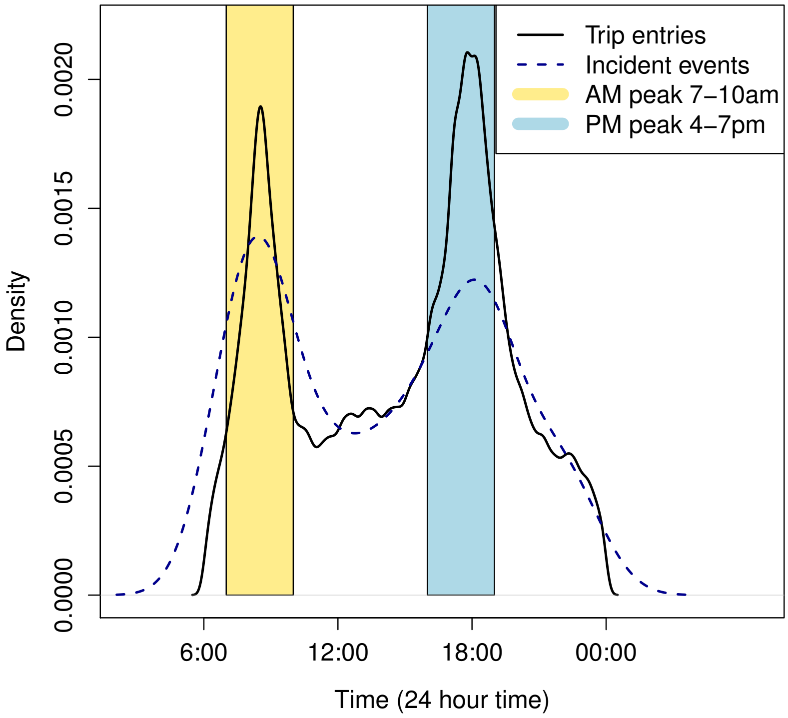

Figure 4.1: Passenger trips and incident events by time of day

Table 4.1: Number of passenger trips and incident events by line and direction

Line	Direction	No. trips	No. incident events	No. incident-affected trips	% incident-affected trips
Central	Eastbound	544451	79	1786	0.33%
	Westbound	507094	63	884	0.17%
Jubilee	Eastbound	1359457	116	6034	0.44%
	Westbound	1534780	105	9783	0.64%
Victoria	Northbound	2596675	172	21118	0.81%
	Southbound	2884964	152	11186	0.39%
Waterloo & City	Eastbound	82941	28	869	1.05%
	Westbound	154641	20	812	0.53%
	Total	9665003	735	52472	0.54%

4.3 Methods

There are two parts to the analysis. First, the journey time distributions are defined by fitting theoretical distributions to empirical data observed over different lines at varying times of the day, under regular and incident-affected conditions. Second, a set of passenger-oriented metrics are proposed to measure the degree to which journey times vary spatially and temporally, to provide operators with information on how well their systems perform.

4.3.1 Operating scenarios

As mentioned in the literature review, the form of a journey time distribution is influenced by the levels of temporal and spatial resolution. In a sample of journey times which are representative of regular uncongested operating conditions on a route, the journey time distribution can be expected to be unimodal and right-skewed (Taylor, 1999; Fosgereau and Fukuda, 2012; Taylor and Susilawati, 2012; Susilawati et al., 2013). For the same route, if samples over a shorter time interval and a shorter route distance are taken, it is likely that fewer passenger trip observations would be available to construct a journey time distribution, and so these journey times are more likely to be randomly distributed in a symmetrical normal form (Li et al., 2006; Mazloumi et al., 2010; Ma et al., 2016). Over longer time intervals, there is a greater opportunity to observe mixed travel conditions with higher proportions of both fast and slow trips, leading to a multimodal distributional form (Susilawati et al., 2013; Jun, 2010; Ma et al., 2016). At the longest route distances and time periods of analyses, literature indicates that the journey time distribution converts back to a normal form in keeping with the central limit theorem, whereby journey times are more influenced by random variations rather than systematic route and time-specific causes of variation (Mazloumi et al., 2010; Rahman et al., 2018).

For this analysis, the purpose is to assess the distribution of journey times within a time period that represents typical travel conditions. Temporally, the time period must be selected such that it is long enough to represent the typical travel conditions that a passenger would experience on a single trip. For network passenger demand analyses and station design considerations, Transport for London (TfL) adopts a 15 minute time period of temporal resolution. Over 15 minutes, travel conditions are deemed to be relatively uniform and representative of the passenger experience. Moreover, during this interval, a sufficient number of observations are available to estimate a continuous probability distribution. As such, a 15 minute interval is adopted for the level of temporal resolution in this analysis. In terms of spatial resolution, all possible combinations of origin-destination routes within the study extents are considered.

This includes routes with no intermediate stops where the origin station is directly adjacent to the destination station, up to routes that span the entire length of a line. The specific extents of the number of journey time distributions generated by line and direction are defined in the following sections.

Aggregate journey time distributions

The analysis of all trips is herein referred to as the aggregate data set. For each unique OD pair within the study area, journey time distributions are generated at 15 minute intervals over the standard operating hours of the London Underground from 6am to 12am. This results in a total of 72 $\times$ 15 minute journey time distributions per OD pair. In terms of spatial scale, all possible OD pairs on each line and direction are analysed, and this results in a total 9144 unique distributions on the Central line, 6480 on the Jubilee line, 17280 on the Victoria line, and 144 on the Waterloo and City line. Table 4.2 provides a summary of the number of journey time distributions by line.

Table 4.2: Number of journey time distributions by line

Line	Direction	No. stations	No. OD pairs	No. 15 min distributions
Central	East & westbound	12	66	4572
Jubilee	East & westbound	10	45	3240
Victoria	North & southbound	16	120	8640
Waterloo & City	East & westbound	2	1	72
	Total	40	464	33408

The results of the distribution fitting are reported by line and direction and by time period, which are defined as follows:

i. AM peak – 7am to 10 am

ii. Midday inter-peak – 10 am to 4 pm

iii. PM peak – 4 pm to 7 pm

iv. Total – Total operating period from 6am to 12am

The AM and PM peak periods are the peak travel periods as defined by TfL based on passenger demand.

Incident-affected journey time distributions

Incident-affected distributions are generated for the OD pairs affected by incident events at 15 minute intervals, measured from the beginning of the incident event recorded in the TfL CuPID incident database. The spatial scale extends from the origin location of the incident to the closest adjacent station and all downstream stations. For example, if an incident occurred between station A and B, OD pair AB is considered, as well as all OD pairs downstream which have station A as the origin.

To report the results, the incidents are categorised by incident severity based on the TfL metric Lost Customer Hours (LCH), which is defined as the sum of the total extra journey time experienced across all affected passengers due to service disruptions of 2 minutes or more. Table 4.3 defines the different incident categories.

Table 4.3: Incident band definition

Band	Lost Customer Hours range	No. trips	% of trips
8	≥ 10000	38	4.82E-04
7	≥ 5000 & ≤ 10000	1161	0.01
6	≥ 1000 & ≤ 5000	5664	0.07
5	≥ 500 & ≤ 1000	5895	0.07
4	≥ 100 & ≤ 500	8364	0.11
3	≥ 50 & ≤ 100	931	0.01
2	≥ 0 & ≤ 50	8621	0.11
1	0	1139	0.01
0	No incident	7850950	99.60
	Total	7882763	100.00

The results of the incident-affected distributions are reported by severity and by time period, i.e. by AM peak, midday inter-peak, PM peak, and the total daily operating period from 6am to 12 am, as previously defined for the analysis of the aggregate journey time distributions.

4.3.2 Unimodal distribution fitting

As mentioned previously, under regular uncongested travel conditions, journey time distributions are unimodal and typically skewed to the right. In the literature, a number of right-skewed distributions have been put forward as providing the best fit for transit journey times. Guided by the findings in the literature, the following distributions are trialled: log normal, gamma, log logistic, and Weibull (Herman and Lam, 1974; Richardson and Taylor, 1978; Polus, 1979; Kieu et al., 2014; Duran-Hormazabal and Tirachini, 2016; Li et al., 2017; Buchel and Corman, 2018). In cases where the distributions may have heavier right tails, the more flexible, multi-parameter distributions including the Burr and Generalised Pareto distributions are also trialled. These distributions are more appropriate for capturing outlying values in the tail of distributions (Taylor and Susilawati, 2012; Susilawati et al., 2013; Guessous et al., 2014). In congested conditions, passenger arrivals tend to be more random, and as a result, the journey time distributions tend to follow a normal distribution (Kieu et al., 2014; Li et al., 2006). The normal distribution is therefore also included in the analysis to capture such possible conditions, which are most likely to arise on OD pairs with high levels of passenger demand and at peak times of the day.

The mathematical forms of the theoretical functions are given in Table 4.4. For the unimodal distribution fitting, the R statistical package "fitdistrplus" is used along with the "actuar" package, which includes the specification of the log logistic, Burr, and Generalised Pareto distributions (Delignette-Muller and Dutang, 2015; Dutang et al., 2008). It is assumed that the observed journey times are identically and independently distributed. The functional parameters of the theoretical distributions are estimated via maximum likelihood estimation, which is a commonly used statistical fitting method. The parameters of the theoretical function are calculated by maximising the likelihood function for the given set of empirical journey time observations (Hastie et al., 2009). The general form of the likelihood function is shown in 4.1.

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta) \tag{4.1}$$

where

x represents journey time,

n represents the total number of individual journey time observations x_i,

$f(x_i|\theta)$ is a theoretical probability density function as per Table 4.4 fitted to the empirical journey time data points x_i ,

θ represents the set of model parameters used in the estimation of the theoretical probability density function, and

$L(\theta|x)$ is the likelihood function.

Table 4.4: Theoretical probability density functions

Distribution	Form
Normal	$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Log normal	$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} exp[-\frac{(\ln x-\mu)^2}{2\sigma^2}]$
Gamma	$f(x;k_g,\theta_g) = \frac{x^{(k_g-1)} e^{-\frac{x}{\theta_g}}}{\theta_g^k{}_g \Gamma(k_g)}$
Log logistic	$f(x;\alpha_{ll},\beta_{ll}) = \frac{(\beta_{ll}/\alpha_{ll})(x/\alpha_{ll})^{\beta_{ll}-1}}{(1+(x/\alpha_{ll})^{\beta}_{ll})^2}$
Weibull	$f(x;\lambda_w,k_w) = \frac{k_w}{\lambda_w}(\frac{x}{\lambda_w})^{k_w-1} exp^{-(x/\lambda_w)^k_w}$
Generalised Pareto	$f(x;\xi,\mu,\sigma) = \frac{1}{\sigma}(1+\frac{\xi(x-\mu)}{\sigma})^{(-\frac{1}{\xi}-1)}$
Burr	$f(x;\lambda_b,c_b,k_b) = \frac{c_b k_b}{\lambda_b}(\frac{x}{\lambda_b})^{c_b-1}[1+(\frac{x}{\lambda_b})^c_b]^{-k_b-1}$

where

x represents journey time,

μ is the mean, σ is the standard deviation, and σ^2 is the variance of journey time,

θ_g is the scale parameter and k_g is the shape parameter for the gamma distribution,

α_{ll} is the scale parameter and β_{ll} is the shape parameter for the log logistic distribution,

λ_w is the scale parameter and k_w is the shape parameter for the Weibull distribution,

ξ is the shape parameter for the Generalised Pareto distribution, and

λ_b is the scale parameter and c_b and k_b are the shape parameters for the Burr distribution.

For illustrative purposes, an example plot of journey time data for trips from Victoria to Oxford Circus in the 15 minute period from 8:45-9:00am is given in Figure 4.2. The plot shows the empirical journey time observations in an histogram format, while the form of the seven theoretical distributions fitted to the data are shown as lines of varying type and colour.

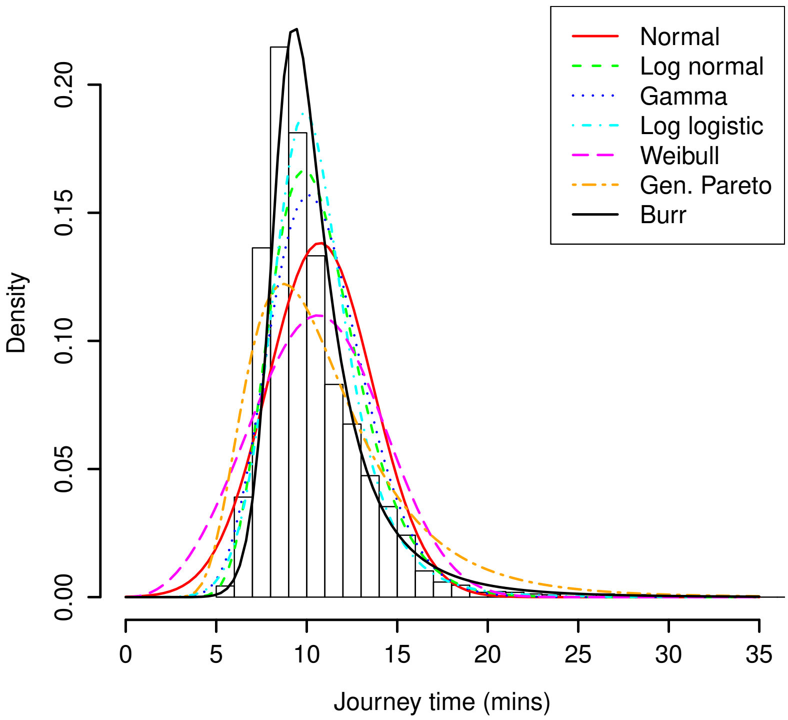

Figure 4.2: Example of theoretical probability distribution functions - Victoria to Oxford Circus, 8:45-9:00am

4.3.3 Evaluation criteria for unimodal distributions

The goodness-of-fit of the unimodal distributions are assessed through log-likelihood based indicators and goodness-of-fit hypothesis testing. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are indicators that assess the goodness-of-fit of a model based on its log-likelihood value and the number of parameters. The general forms of the indicators are given in equations 4.2 and 4.3. Both indicators penalise models with high degrees of complexity (i.e. if there are many model parameters to be estimated). The BIC penalises complexity to a greater degree than the AIC. Smaller values of the AIC and BIC indicate a better fitted model (Hastie et al., 2009).

$$AIC = -2\log(L) + 2p \tag{4.2}$$

$$BIC = -2\log(L) + \log(n)p \tag{4.3}$$

where

p is the number of estimated parameters in the model,

n is the number of observations used to generate the model, and

$\log(L)$ is the log-likelihood of the model, where the likelihood is calculated via the likelihood function as defined in equation 4.1.

Three types of goodness-of-fit tests are also used to assess the fit of theoretical distributions: the Kolmogorov-Smirnov test, and extensions to this test known as the Cramer-von Mises test and Anderson-Darling test. The null hypothesis of this family of tests is that the fitted theoretical distribution function $f(x)$ is uniformly close to the empirical distribution function $f_n(x)$, i.e. $H_0 : f(x) = f_n(x)$. The alternative hypothesis is that the functions differ, i.e. $H_1 : f(x) \neq f_n(x)$. The test statistics compute a measure of the discrepancy between the empirical and theoretical distribution functions; large values of the test statistics beyond a critical level lead to a rejection of the null hypothesis (Lehmann and Romano, 2005). The general form of the test statistics for the three types of tests are given in equations 4.4, 4.5, and 4.6.

Kolmogorov-Smirnov

$$KS = sup|f_n(x) - f(x)| \tag{4.4}$$

Cramer-von Mises

$$CvM = n \int_{-\infty}^{\infty} (f_n(x) - f(x))^2 dx \tag{4.5}$$

Anderson-Darling

$$AD = n \int_{-\infty}^{\infty} \frac{(f_n(x) - f(x))^2}{f(x)(1 - f(x))} dx \tag{4.6}$$

where

x represents journey time,

n is the number of journey time observations,

$f_n(x)$ is the empirical distribution function, and

$f(x)$ is the theoretical probability distribution function as per Table 4.4.

The three types of tests are conducted, as each places different emphasis on different aspects of the distribution. Literature suggests that the Kolmogorov-Smirnov and Cramer-von Mises tests are more powerful in terms of detecting deviations from the median range of the data (Lehmann and Romano, 2005; Delignette-Muller and Dutang, 2015). The Anderson-Darling test gives relatively more emphasis to the tail of the distributions (Delignette-Muller and Dutang, 2015).

Further supporting information to verify the shape of the distributions can be obtained by computing measures of dispersion, namely, the skewness and kurtosis of the distributions (Delignette-Muller and Dutang, 2015). The skewness indicates the degree of symmetry of a distribution. The general equation for skewness is given in equation 4.7 (Casella and Berger, 2002). The value of skewness for the symmetrical normal distribution is 0. A negative value of skewness indicates that the distribution has a longer left tail, and a positive value indicates a longer right tail. The kurtosis of a distribution provides information on the tails of the distribution. The general form of the equation for kurtosis is given in equation 4.8 (Casella and Berger, 2002). A positive value indicates that the distribution is more heavily tailed with a more pronounced peak, while a negative value indicates that the distribution is more lightly tailed with a peak that is less discernible.

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{s^3} \tag{4.7}$$

$$\text{Kurtosis} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{ns^4} \tag{4.8}$$

where

x_i represents the ith observation of journey time in the sample data,

n is the total number of observations of journey time in the sample data,

$\bar{x}$ is the mean of journey times in the sample, and

s is the standard deviation of journey times in the sample.

4.3.4 Multimodal distribution fitting

The literature indicates that for distributions that capture mixed travel conditions from free-flowing to congested conditions, more than one peak may be observed (Jun, 2010; Ma et al., 2016; Barkley et al., 2012). If the level of temporal resolution is appropriate, extending this theory to the case of incident-affected conditions, multiple peaks are likely to be observed in the journey time distribution. The first peak is likely to capture trips that are unaffected by the incidents and are therefore distributed around a shorter journey time, while subsequent peaks at longer journey times capture trips that are affected by delays as a result of the incident event.

Distributions that exhibit more than one peak can be modelled using a multimodal distributional form, and this form is trialled in the analysis of incident-affected distributions. The fitting process involves first confirming the presence of multiple peaks in the distribution, and then fitting a mixture model with an appropriate number and form of component distributions. Hartigans' dip test is used to establish the presence of multimodal peaks in the distributions.

The null hypothesis for the test is that the distribution is unimodal against the alternative hypothesis that the distribution is not unimodal i.e. multimodal. To compute the test statistic, a unimodal distribution function is generated which minimises the maximum difference between the empirical data and the unimodal function. The test statistic is then calculated via an iterative algorithm to find the maximum difference between the empirical distribution and the unimodal distribution function (Hartigan and Hartigan, 1985; Hartigan, 1985). The R statistical package "diptest" is used to undertake the test (Maechler, 2016).

Once the presence of multimodal peaks is established, a mixture model is fitted using the expectation maximisation technique. This fitting technique can be considered as a specific case of maximum likelihood estimation, where the empirical data set is assumed to contain missing latent variables (or unobserved observations), and the estimation of these missing observations in turn facilitates the estimation of the underlying model parameters (Benaglia et al., 2009; Hastie et al., 2009). The algorithm is an iterative process comprising of two steps. The first step, termed expectation, involves calculating the expected value of the log-likelihood of the model given an initial estimate of the model parameters (see equation 4.9). The second step, maximisation, re-calculates the parameters of the model through maximisation of the log-likelihood from the expectation step (see equation 4.10). The expectation step is then repeated with the updated parameter values, which are used to compute the distribution of the latent variables. The process is repeated until convergence in the parameter values is achieved.

$$\text{Expectation step: } Q(\theta|\theta^{\tau}) = E[\log(L_{\theta}(C \mid x, \theta^{\tau}))] \tag{4.9}$$

$$\text{Maximisation step: } \theta^{\tau+1} = \arg\max Q(\theta|\theta^{\tau}) \tag{4.10}$$

where

x represents the given (known) observations of journey time,

C represents the complete data set of journey times, which comprises of the vector of known observations of journey time x and the vector of unknown observations of journey time z, such that $C = (x, z)$,

θ represents the model parameters,

L_θ represents the likelihood function of the multimodal distribution, evaluated for the model parameters θ, and

τ is the iteration number, while θ^τ represents the value of the model parameters at iteration τ.

In the literature, the component distributions are typically specified as the normal distribution; the resulting multimodal model is termed the Gaussian mixture model (Susilawati et al., 2013; Ma et al., 2016). The normal distribution is therefore adopted for the component distributions in the mixture models trialled in the analysis. The general form of the multimodal mixture model is given in equation 4.11. The R statistical package "mixtools" is used for the multimodal distribution fitting (Benaglia et al., 2009). An example illustration of a bimodal (two peak) distribution is shown in Figure 4.3 for reference.

$$f(x|\lambda_m, \theta) = \sum_{k=1}^{K} \lambda_m k f_k(x|\theta_k) \tag{4.11}$$

where

x represents journey time,

λ_m is a vector of mixture coefficients,

θ is a vector of model parameters, and

$f_k(x|\theta_k)$ is the normal probability distribution function for the kth component distribution. For bimodal distributions k=2, and for three peaks $k = 3$.

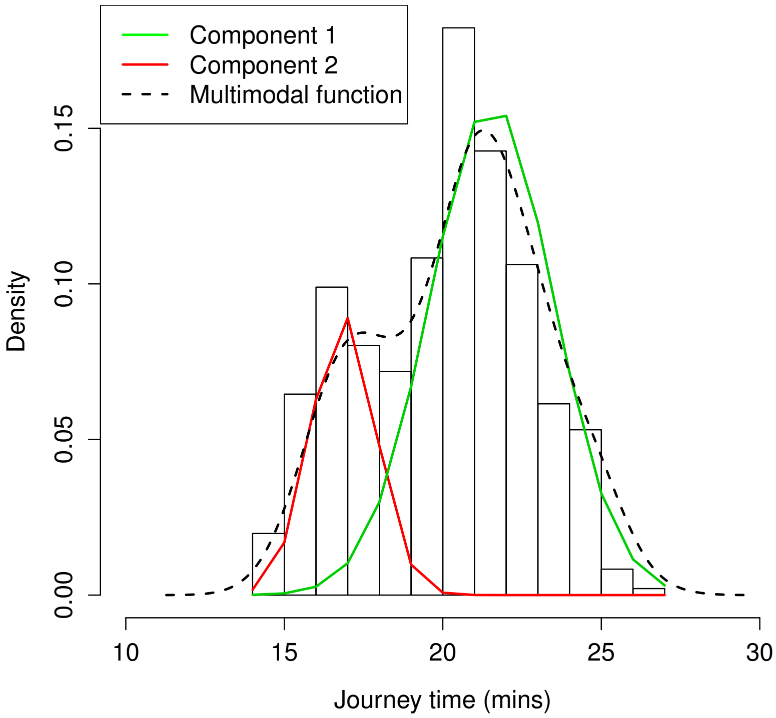

Figure 4.3: Example of multimodal probability distribution function - Brixton to Oxford Circus, 8:00-8:15am, 13 November 2013

4.3.5 Journey time performance metrics

Responding to the need to provide better measures of the passenger experience, the existing work on passenger based metrics is built upon, and three new journey time performance metrics are proposed. In the following sections, the criteria for the development of the metrics is first detailed, and the proposed metrics are then defined.

Criteria for metric selection

Criteria for the metrics developed in this thesis build on the criteria defined by Chan (2007), Uniman (2009), and Schil (2012) in their work on developing journey time reliability metrics for the London Underground. The main criteria for the metrics are defined as follows:

1. *Passenger-oriented* - As discussed in the literature review in Chapter 2, the metrics currently in use to measure journey time performance on the London Underground are more

operations-oriented, focusing on average or aggregate measures of service performance using data from the train signalling system. The main objective of the metrics developed in this thesis is that they must be reflect the passenger experience, and capture the full set of travel conditions that a passenger is exposed to. As such, the metrics must be derived from the distribution of passenger journey times generated from the Oyster AFC data.

2. *Comparable* - The metrics must be of a form that enables direct comparisons of performance to be made across routes of differing length. A direct comparison of absolute values may give misleading information about the performance of a given route relative to another. For example, a longer route may be prone to a higher degree of variance in absolute terms compared to shorter routes, simply by virtue of the route being of a longer length. As such, an appropriate normalising value should be integrated within any measure proposed.

3. *Able to be aggregated* - Following on from the requirement of comparability, the metrics must additionally have the ability to be aggregated at different spatial and temporal levels. The lowest spatial resolution of the metrics is at an individual origin-destination route level. As discussed in the review of metrics used by the London Underground in Chapter 2, the metrics are reported at a network level to reflect the overall performance of the system. The metrics developed must therefore have the ability to be spatially aggregated from an OD route level to a network level. In terms of temporal scale, the London Underground reports on network performance at monthly and yearly intervals and additional segmentation is undertaken to distinguish between peak and off-peak periods across the day (Transport for London, 2019a, 2018a). The metrics must therefore also have the ability to be aggregated at different temporal levels.

4. *Accessible* - The methodology involved in calculating the metrics must be straight-forward enough to be adopted by operators, as repeat measurement at regular intervals to track performance over time is typically required for input into network performance reports. The resulting values of the metrics must also be able to be easily interpreted by operators

for regular performance reporting (as mentioned previously, the London Underground reports on network performance at monthly and yearly intervals (Transport for London, 2019a, 2018a)).

Definition of journey time metrics

As mentioned in the first criterion, the objective of the metrics is to provide a measure of the full set of travel conditions that a passenger may experience. As detailed in the literature review in Chapter 2, the probability distribution function of all observed passenger journey times represent the full range of travel conditions on a route over the given time period of observation (Abkowitz et al., 1978; Lomax et al., 2003; Van Lint and Van Zuylen, 2005; Chan, 2007; Uniman, 2009; Schil, 2012; Wood, 2015). In the literature, the mean-variance approach is typically adopted, whereby the full distribution of journey times is characterised by measures of the mean and variance on a route (Taylor and Susilawati, 2012; Lomax et al., 2003; Noland and Polak, 2002). In the development of the performance metrics here, the measure of mean journey time is directly derived from the distribution of journey times in the conventional manner. Rather than adopt a singular measure of the total variance on a route, the framework for journey time variance introduced by Wong and Sussman (1973) is used as a guide. In their framework, journey time variance is split into three sources: 1) Regular condition-dependent variations, 2) Irregular condition-dependent variations, and 3) Random variations. The three sources of variance can be measured from the moments of the journey time distribution. The first represents the day-to-day reliability of the system excluding outliers to reflect the "typical" variance in journey times that a passenger experiences. The second represents the unpredictable outlying journey times that may arise when the system is operating under disrupted or highly congested conditions. The third random source of variance is more difficult to measure outright, but can be captured in measures of total journey time variance on a route.

On the basis of the literature on characterising journey time distributions, the development of the passenger oriented metrics is based on the modified mean-variance framework described above. Three metrics are proposed using measures of the mean and spread of the journey time distributions generated from the AFC Oyster data. The first metric provides a measure of average travel conditions on the route derived from a measure of the mean of the distribution. The second and third metrics capture sources of variation. The second metric is used to quantify the first source of variance identified in the Wong and Sussman (1973) framework, and provides a measure of the typical travel conditions on a route excluding outliers. The third metric provides a measure of the total variance of journey times on a route including outliers from disrupted and/or congested travel conditions as well as random variation; this metric therefore captures the second and third sources of variance presented in the Wong and Sussman (1973) framework. Between the three metrics, the full distribution of journey times are captured, representing the full range of travel conditions that a passenger experiences.

The specific moments of the journey time distribution that are used in the calculation of the journey time metrics are illustrated in Figure 4.4 for reference. To meet the comparability criteria of the metrics, each journey time metric is normalised by the free flow time on the route in order to enable comparison across routes of differing lengths. A methodology for aggregation of the metrics is also proposed, to enable the calculation of the metrics at different spatial and temporal levels as outlined in the third criteria for the metrics. It is important to note that the metrics proposed here are based on total passenger journey times, and so records of both the passenger tap-in at the entry station and tap-out at the exit station are required. The metrics are defined in more detail in the following sections.

1. Normalised mean

$$\bar{x}_{od,t}/FF_{od} \tag{4.12}$$

Where $\bar{x}_{od,t}$ is the mean value of journey times $x_{i,od,t}$ observed for OD pair od over $t = 15$ minutes. FF_{od} is the free flow time calculated as the 10th percentile journey time in the aggregate journey time distribution of the OD pair over all times. The mean is calculated as follows:

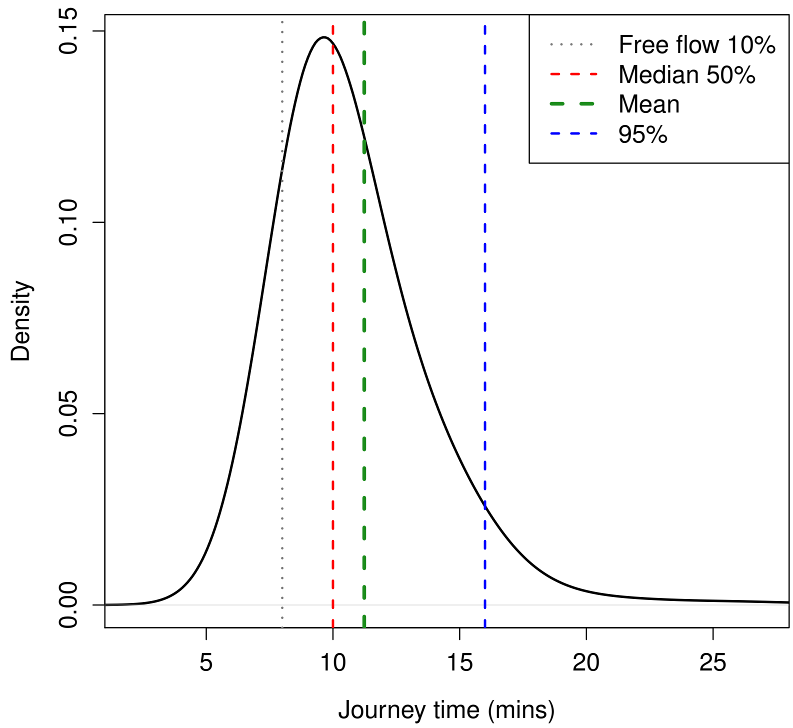

Figure 4.4: Moments of a journey time distribution used in the calculation of performance metrics

$$\bar{x}_{od,t} = \frac{\sum x_{i,od,t}}{n_{od,t}} \tag{4.13}$$

Where $n_{od,t}$ is the number of journey times observed for the OD pair over 15 minutes.

The normalised mean provides an indication of the mean journey time between a given OD pair over a time interval of 15 minutes. As mentioned previously, the mean of journey times is one of the most fundamental measures used to reflect typical travel conditions on a route (Kittelson & Associates Inc et al., 2013; Taylor, 2013; Lomax et al., 2003; Transportation Research Board Committee on Highway Capacity and Quality of Service, 2010). It is calculated as the sum of all observations, divided by the total number of observations (Wonnacott and Wonnacott, 1990). The free flow time per OD pair is used as the normalising value; it represents the time taken to travel from the origin to the destination in base uncongested conditions without delays (Transportation Research Board Committee on Highway Capacity and Quality of Service, 2010). The free flow time is defined as the 10th percentile of the aggregate journey time distribution for the OD pair of interest over all times. The advantage of taking the 10th percentile value is that it is not influenced by outlying journey times from the fastest individuals travelling in ideal situations e.g. a person running through an empty station to board the train just as it arrives. The normalised mean indicator proposed here can be used to determine how close the average travel conditions are to free flow conditions, and in turn, inferences can be made regarding the levels of congestion and/or delay on the route over the given time period. Higher values of the indicator imply that travel is more congested or delayed.

2. Normalised reliability buffer time

$$RBT_{od,t}/FF_{od} \tag{4.14}$$

Where $RBT_{od,t}$ is reliability buffer time for OD pair od over $t = 15$ minutes, defined as the 95th percentile journey time subtracted by the median (50th percentile) journey time as follows:

$$RBT_{od,t} = (95\%x - 50\%x)_{od,t} \tag{4.15}$$

Where x is a vector of journey times x_i observed on the OD pair over a 15 minute interval. FF_{od} is the free flow time calculated as the 10th percentile travel time for the OD pair.

The reliability buffer time metric provides a measure of the compactness of the tail end of the distribution from the 50th to the 95th percentile. The metric disregards extreme outliers, and so provides a measure of the variance of journey times during "typical" operating conditions as experienced by the passenger. As detailed in the literature review, the RBT metric was first developed and proposed for use by the US Federal Highway Administration for road based traffic, with the lower statistical threshold set as the mean journey time (Lomax et al., 2003; Cambridge Systematics Inc. and Texas Transport Institute, 2005). Since then, researchers have refined the definition for the lower statistical threshold to replace the mean value with the median journey time, the rationale being that the median is not affected by outlying journey times, and therefore provides a better representation of typical journey times on a route (Chan, 2007; Uniman, 2009; Uniman et al., 2010). For the upper statistical threshold, the 95th percentile journey time is seen as the most appropriate bound to optimise the balance between the amount of extra buffer time to be budgeted and the likelihood of arriving late at the destination, as discussed in the literature review in Chapter 2 (Cambridge Systematics Inc. and Texas Transport Institute, 2005; Chan, 2007; Uniman, 2009; Uniman et al., 2010). The RBT can be interpreted as a measure of passenger journey time reliability; it is the time that a passenger would typically need to budget above the median journey time in order to arrive on time at the destination at a confidence level of 95% (19 out of 20 trips). A higher value implies higher levels of journey time variance, while lower levels indicate a more narrow spread of times.

3. Normalised standard deviation

$$s_{od,t}/FF_{od} \tag{4.16}$$

Where $s_{od,t}$ is the standard deviation of journey times $x_{i,od,t}$ observed for OD pair *od* over $t = 15$ minutes, and FF_{od} is the free flow time calculated as the 10th percentile travel time for the OD pair. The standard deviation is calculated as follows:

$$s_{od,t} = \sqrt{\frac{\sum(x_{i,od,t} - \bar{x}_{od,t})^2}{(n_{od,t} - 1)}} \tag{4.17}$$

Where $\bar{x}_{od,t}$ is the mean of the journey times, and $n_{od,t}$ is the number of journey times observations for the OD pair over 15 minutes.

As much as it is important to capture the "typical" performance on a route to assess journey time reliability, it is equally important for operators to understand where journey times significantly deviate. In reality, extreme events such as incidents, excessive levels of congestion, and outlying individual travel behaviour are observed, and can greatly impact passenger journey times. In order to quantify the effects of the extreme outliers, the use of the normalised standard deviation indicator is recommended. The standard deviation of journey times on a route is typically adopted as the conventional measure to capture the total variation in journey times (Taylor, 2013; Lomax et al., 2003; Noland and Polak, 2002). The standard deviation measures the total spread of a given distribution including outliers (Wonnacott and Wonnacott, 1990). Higher values of the indicator represent higher levels of journey time variability, whereas lower levels show greater consistency in travel times.

4. Aggregation of metrics

$$M_{ld,t} = \frac{\sum_{ld} m_{od,t}.n_{od,t}}{\sum_{ld} n_{od,t}} \tag{4.18}$$

Where $M_{ld,t}$ is the weighted value of metric m for line l, direction of travel d, over time period t. $m_{od,t}$ is the unweighted value of the metric m for OD pair od over time t, and $n_{od,t}$ is the number of trips undertaken on the OD pair over the time period.

To compare performance at a greater spatial and temporal scale, the metrics presented above can be aggregated over many OD pairs to indicate the average performance by line and direction over a given time period of interest. To do this, a weighted mean value is used. The number of passengers per OD pair under the time period of interest is defined as the weighting value in the majority of studies on developing journey time metrics, and so this weighting value is also adopted for the metrics developed here (Chan, 2007; Uniman, 2009; Schil, 2012). It should be noted that the aggregation of trips here considers single line trips only where the origin and destination are on the same line; trips that have transferred from other lines have not been accounted for in this case.

The three performance metrics are evaluated separately for the set of regular trips, and for the set of incident-affected trips. The two sets of indicators can be directly compared by taking the ratio of the values of each metric under incident-affected conditions against the values of the metric under regular conditions, as specified in equation 4.19. The I/R ratio quantifies how the incident-affected trips perform in comparison to regular trips for each metric, and is of value to operators in explaining how resilient different lines are insofar as the extent to which incidents affect journey times.

$$(I/R)_M \text{ ratio} = \frac{I_{M,ld,t}}{R_{M,ld,t}} \tag{4.19}$$

Where I refers to the incident-affected trips, R refers to the regular set of trips, M represents the weighted value of the metric under consideration for line l, direction d, over time interval t.

4.4 Results

The following sections summarise the results for the two parts of the analysis beginning with the results of the distribution fitting, followed by the results for the performance metrics. In each of the two parts, the results representing regular travel are presented first, followed by the results for the incident-affected trips.

4.4.1 Distributional form

Regular travel

For each OD pair per line and direction, journey times are analysed at 15 minute intervals over the standard operating hours of the London Underground from 6am to 12am. It should be noted that distributions have not been fitted in intervals where there is a lack of or insufficient data; this is mostly observed at the extremities of the day i.e. early in the morning at 6am at the beginning of operations, and late in the evening close to 12am at the end of the day. As mentioned in the discussion of the analysis methods, for each empirical distribution generated, seven theoretical distributions are fitted, and five performance indicators are calculated, namely: the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling test statistics, and the AIC and BIC indicators. Each of the five indicators are given equal weighting, and the distribution that performs best across all five criteria is selected as the best-fitting distribution for the empirical data. Table 4.5 summarises the results of the best fitting distributions by line and direction, and Table 4.6 provides the results by time period. The tables show the number of empirical journey time distributions that adhere to each of the seven theoretical forms. For additional supporting information on the shape of the distributions, the aggregated values of skewness and kurtosis are also reported.

Table 4.5: Aggregate data - number of empirical distributions per distributional form by line and direction

Line	Direction	Normal	Log normal	Gamma	Log logistic	Weibull	Gen. Pareto	Burr	Total	Skew.	Kurt.
Central	Eastbound	81	354	83	1569	130	113	1478	3808	3.54	34.97
	Westbound	58	306	73	1603	101	116	1520	3777	3.47	31.75
Jubilee	Eastbound	22	163	37	997	28	50	1850	3147	4.69	54.45
	Westbound	20	184	44	1083	20	53	1775	3179	4.62	54.78
Victoria	Northbound	144	526	158	3543	116	193	3689	8369	4.73	58.02
	Southbound	121	442	112	3580	103	148	3855	8361	4.53	54.78
Waterloo	Eastbound	1	14	7	20	3	3	24	72	3.16	45.33
& City	Westbound	0	15	5	41	1	1	9	72	6.50	111.21
	Total	447	2004	519	12436	502	677	14200	30785		
	%	1.5%	6.5%	1.7%	40.4%	1.6%	2.2%	46.1%	100.0%		

Note: Skew. refers to skewness, Kurt. refers to kurtosis, and Gen. denotes Generalised.

Table 4.6: Aggregate data - number of empirical distributions per distributional form by time period

Time period	Normal	Log normal	Gamma	Log logistic	Weibull	Gen. Pareto	Burr	Total	Skew.	Kurt.
AM peak	73	398	109	2106	73	128	2107	4994	3.53	41.55
Inter-peak	119	676	121	4451	111	188	4879	10545	4.89	56.38
PM peak	27	167	28	1957	26	54	3185	5444	5.16	65.29
All	447	2004	519	12436	502	677	14200	30785	4.53	54.10
%	1.5%	6.5%	1.7%	40.4%	1.6%	2.2%	46.1%	100.0%		
Skew.	0.19	2.00	0.89	4.69	0.05	1.58	4.75			
Kurt.	3.20	28.39	7.26	58.00	2.63	16.59	55.50			

Note: Skew. refers to skewness, Kurt. refers to kurtosis, and Gen. denotes Generalised.

Beginning with a general discussion of the indicative measures of skewness and kurtosis, Table 4.5 shows that the average values of skewness for each line and direction range from approximately 3.2-6.5, and the average values of kurtosis range from approximately 32-111. The skewness of a perfectly symmetrical normal distribution is 0. The results for skewness are greater than 0 in magnitude and positive, therefore indicating that on average, the journey time distributions across all lines tend to be right-skewed. For the normal distribution, the value of kurtosis is 3. The average values of kurtosis are consistently higher than 3 across all lines and directions, which indicates that the journey time distributions tend to have long tails.

Based on the results of the five goodness-of-fit evaluation criteria, the Burr distribution performs best, fitting approximately 46% of all distributions, and the log logistic distribution performs similarly well, capturing a share of approximately 40% of all distributions. The values of skewness and kurtosis associated with the Burr and log-logistic distributions are the highest among all distributions, supporting the theory that these distributions are best at capturing long tails. The dominance of the Burr and log logistic distributions therefore indicates that there is a significant percentage of longer outlying journey times across all lines. This could be a result of high levels of passenger demand leading to longer journey times, delays from incident events, and individual-specific behaviour, amongst other factors.

In terms of comparing the shapes of the distributions by line and direction, Table 4.5 shows that journey times on the Waterloo and City line eastbound and the Central line are the least skewed with the lowest degree of kurtosis. This indicates that the journey times on these lines are comparatively lighter tailed, indicting fewer outlying journey times compared to the other lines. The skewness and kurtosis are highest on the Waterloo and City line westbound, indicating that journey time distributions are the most skewed and heavily tailed. This is likely reflecting high levels of congestion on the Waterloo and City line westbound.

Comparing the measures by time of day, Table 4.6 shows that the skewness and kurtosis of the distributions are highest during the afternoon peak period. Referring back to the distribution of passenger trips by time of day in Figure 4.1, it can be seen that the PM peak is the busiest time of the day. As a result, it therefore follows that the journey time distributions are the most skewed and heavily tailed during the PM peak. Interestingly, the results show that although there is a higher volume of passenger trips in the AM peak compared to the inter-peak periods, the skewness and kurtosis of the distributions are lowest during the AM peak. This indicates that although passenger volumes are higher, the journey time distributions during the AM peak are comparatively less skewed with fewer outliers. It is likely that average journey times during the AM peak are longer than the inter-peak period as a result of the higher levels of passenger demand, while the variance in journey times is lower. It should be noted that measures of the mean and variance of journey times are presented in the second part of the results section to further quantify the relative performance of the lines by time period.

Although most distributions are found to be right-skewed, Tables 4.5 and 4.6 show that 447 data sets (1.5% of all distributions) adhere to a symmetrical form, which is represented by the normal distribution. The literature suggests that travel conditions can be interpreted as being more congested if the journey times are distributed symmetrically around a higher mean value, compared to distributions which are right-skewed and distributed around a lower mean value (Taylor, 1999). The data support this theory: on average, the mean journey times of OD pairs with normal distributions are higher than the mean journey times of the same OD pairs during time periods when the journey time distributions are right-skewed. The Victoria line has the highest number of normal distributions out of all lines. In terms of the temporal scale, the normal distribution is observed approximately four times more during the morning peak compared to the afternoon peak (see Table 4.6).

Incident-affected travel

For the incident data, the distributions are tested for the presence of multimodality via Hartigans' dip test at 15 minute intervals from the start time of the incident event. The results are summarised by line and direction and by time period in Tables 4.7 and 4.8, respectively.

Table 4.7: Incident data - number of empirical distributions per distributional form by line and direction

	Incident band	1		2		3		4		5		6		7		8					
	Form	M	U	M	U	M	U	M	U	M	U	M	U	M	U	M	U	Total	% M	Skew.	Kurt.
Line	Direction																				
Central	Eastbound	0	0	17	49	3	6	14	78	5	71	17	79	0	6	0	0	345	16.2%	0.68	3.85
	Westbound	0	0	14	63	4	26	15	58	1	15	0	14	0	3	0	0	213	16.0%	0.39	2.79
Jubilee	Eastbound	13	18	69	68	9	17	20	43	19	21	18	28	1	12	0	0	356	41.9%	1.17	9.44
	Westbound	8	20	91	117	9	6	47	58	16	35	48	58	0	7	0	6	526	41.6%	0.88	5.67
Victoria	Northbound	11	16	43	121	69	72	252	297	62	89	39	131	1	5	0	0	1208	39.5%	0.92	6.26
	Southbound	2	12	46	108	10	50	105	239	84	81	34	106	15	25	0	0	917	32.3%	1.52	14.35
Waterloo	Eastbound	0	0	2	4	1	1	5	1	3	0	1	1	0	0	0	0	19	63.2%	1.18	7.60
& City	Westbound	0	0	0	3	2	0	2	0	0	0	1	0	0	0	0	0	8	62.5%	1.49	7.77
	Total	34	66	282	533	107	178	460	774	190	312	158	417	17	58	0	6	3592	34.7%		
	% M	34.0%		34.6%		37.5%		37.3%		37.8%		27.5%		22.7%		0.0%					

Note: "M" refers to multimodal distributions, "U" refers to unimodal distributions, Skew. refers to skewness, and Kurt. refers to kurtosis.

Table 4.8: Incident data - number of empirical distributions per distributional form by time period

Incident band	1		2		3		4		5		6		7		8					
Form	M	U	M	U	M	U	M	U	M	U	M	U	M	U	M	U	Total	% M	Skew.	Kurt.
Timeperiod																				
AM peak	0	0	41	67	34	34	173	171	89	138	71	157	8	16	0	3	1002	41.5%	0.94	7.02
Inter-peak	9	25	69	154	30	46	94	209	41	90	24	96	1	18	0	0	906	29.6%	0.65	3.89
PM peak	16	8	88	106	19	11	133	179	38	40	54	112	8	20	0	3	835	42.6%	1.65	13.81
All	34	66	282	533	107	178	460	774	190	312	158	417	17	58	0	6	3592	34.7%	1.07	8.22
% M	34.0%		34.6%		37.5%		37.3%		37.8%		27.5%		22.7%		0.0%					
Skewness	1.15	0.19	0.77	0.41	0.58	0.46	1.34	0.45	1.42	0.50	1.75	0.49	0.71	0.72	NA	0.25				
Kurtosis	7.33	2.54	5.14	3.16	3.93	3.14	11.29	2.95	11.73	2.91	12.34	2.96	4.16	3.87	NA	2.01				

Note: "M" refers to multimodal distributions, "U" refers to unimodal distributions, Skew. refers to skewness, and Kurt. refers to kurtosis.

When testing for the presence of multimodal peaks in the journey time distributions, approximately 35% of all trips that could be affected by incidents reject the null hypothesis of unimodality, indicating that at least two peaks are present (see Table 4.7). According to the literature, this indicates that within these distributions, mixed travel conditions, both congested and uncongested, are observed during incident events over an interval of 15 minutes from the start of the incident. Overall, the majority (65%) of the incident-affected distributions adhere to a unimodal form. This implies that within the 15 minute period of analysis, the distributions do not show clear evidence of mixed congested and non-congested travel times. Within these unimodal distributions, the effect of the incident is likely reflected in journey times being distributed around a singular higher mean value. This is shown to be the case in the next section of the results, where measures of the mean and variance of journey times are evaluated under incident conditions.

Comparing lines, Table 4.7 shows that the Victoria line northbound has the greatest number of incident-affected distributions (1208), and the Waterloo and City line westbound has the least (8). Of the incident-affected distributions, the Waterloo and City line has the highest percentage of multimodal distributions (approximately 63%), indicating that mixed congested and uncongested travel conditions are commonly observed during incident events. The Central line has the lowest proportion of multimodal distributions, with approximately 16% of the incident-affected distributions showing more than one peak.

Comparing across time periods, Table 4.8 shows that the morning peak has the highest number of incident-affected distributions (1002), followed by the inter-peak (906), and the afternoon peak period has the least amount of incident-affected distributions (835). The highest proportion of multimodal distributions are observed during the AM peak and PM peak (approximately 42%), while approximately 30% of the incident-affected distributions are multimodal in the inter-peak. This indicates that in the peak periods, mixed conditions are more likely to be observed compared to the inter-peak.

In terms of incident severity, Tables 4.7 and 4.8 show that multimodal distributions are more commonly observed amongst the mid incident severity ranges, i.e. incident bands 3, 4, and

5, which correspond to lost customer hours (LCH) between 50 and 1000 LCH. Multimodal distributions account for approximately 37% of the incident-affected distributions in this range of incidents. As incident severity increases past band 5 (1000 LCH), the amount of multimodal distributions observed decreases. At severe incident levels, more journeys are likely to be affected and so journey times are more likely to be distributed around a singular higher mean value.

Generally, it can be seen that the measures of skewness and kurtosis for the incident-affected distributions are much lower compared to the set of regular journey time distributions discussed in the previous results section. The values of skewness range from 0.39-1.52 across lines for the incident-affected distributions, whereas the values range from 3.2-6.5 for the regular journey time distributions. The kurtosis ranges from approximately 2.8-14.4 for the incident-affected distributions, and 32-111 for the regular distributions. This indicates that the incident-affected distributions tend to be more symmetrical, with lighter tails and fewer outlying values. The most likely explanation is that this form is reflecting a higher proportion of trips distributed around a higher mean value during incident-affected conditions. This notion is supported by the results in the next sections on measures of the mean and variance of journey times.

4.4.2 Journey time performance metrics

Regular travel

Table 4.9 summarises the journey time performance measures by line, direction, and time of day. It should be noted that over the course of the analysis, operator feedback indicated that the detailed tabular format may be cumbersome for operators to interpret in practice. As such, the metrics are also presented in a more accessible format consistent with the standard benchmarking style used by the international metro benchmarking groups administered by the Transport Strategy Centre at Imperial College. In Figures 4.5, 4.6, and 4.7, lines are ranked from best to worst in reference to the value of each metric aggregated over the total period of analysis ("Total" rows in Table 4.9).

Table 4.9: Aggregate data - journey time performance measures by line, direction, and time period

Line	Direction	Time period	Mean/FF	RBT/FF	s/FF
Central	Eastbound	AM peak	1.19	0.42	0.23
		Inter-peak	1.23	0.47	0.31
		PM peak	1.25	0.47	0.31
		Total	1.24	0.45	0.29
	Westbound	AM peak	1.23	0.55	0.34
		Inter-peak	1.22	0.48	0.31
		PM peak	1.26	0.51	0.30
		Total	1.38	0.49	0.31
Jubilee	Eastbound	AM peak	1.27	0.58	0.31
		Inter-peak	1.24	0.51	0.35
		PM peak	1.29	0.64	0.39
		Total	1.28	0.57	0.35
	Westbound	AM peak	1.25	0.47	0.28
		Inter-peak	1.23	0.46	0.32
		PM peak	1.29	0.57	0.36
		Total	1.34	0.51	0.34
Victoria	Northbound	AM peak	1.20	0.35	0.21
		Inter-peak	1.23	0.48	0.35
		PM peak	1.24	0.41	0.27
		Total	1.25	0.42	0.28
	Southbound	AM peak	1.24	0.40	0.24
		Inter-peak	1.21	0.43	0.32
		PM peak	1.27	0.47	0.30
		Total	1.32	0.45	0.29
W&C	Eastbound	AM peak	1.53	0.59	0.34
		Inter-peak	1.33	0.56	0.44
		PM peak	1.30	0.68	0.43
		Total	1.42	0.68	0.44
	Westbound	AM peak	1.30	0.53	0.51
		Inter-peak	1.51	0.66	0.69
		PM peak	1.50	0.65	0.52
		Total	1.55	0.71	0.59

Note: FF refers to free flow time, RBT is the reliability buffer time, and s refers to the sample standard deviation.

As mentioned in the section covering the definition of the indicators, the mean values of the indicators over the specified time intervals are calculated by applying weightings of passenger flows along each unique OD pair. Here the values of the indicators are aggregated over the time periods of the AM peak (7-10am), the inter-peak (10am-4pm), the PM peak (4pm-7pm), and over the total daily operating time period (6am-12am).

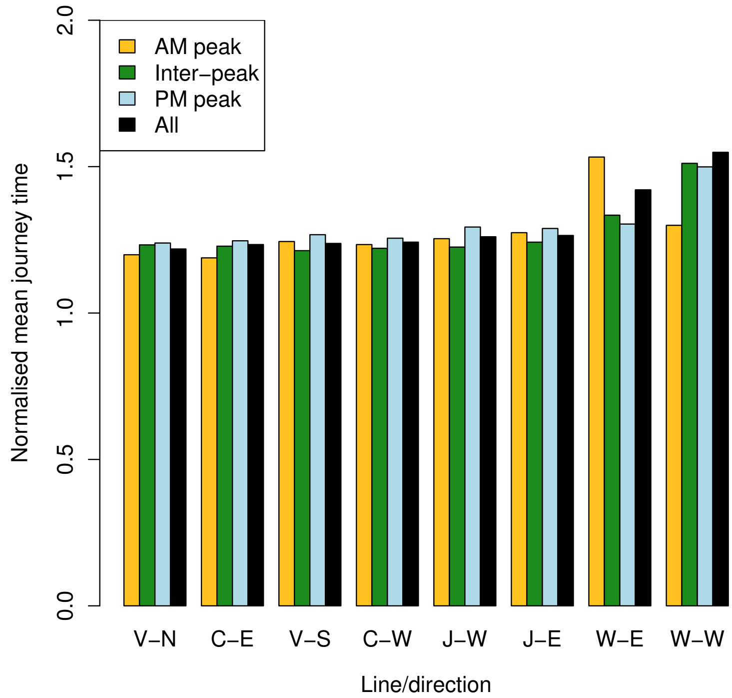

Figure 4.5: Benchmarking format - normalised mean of journey time by time period

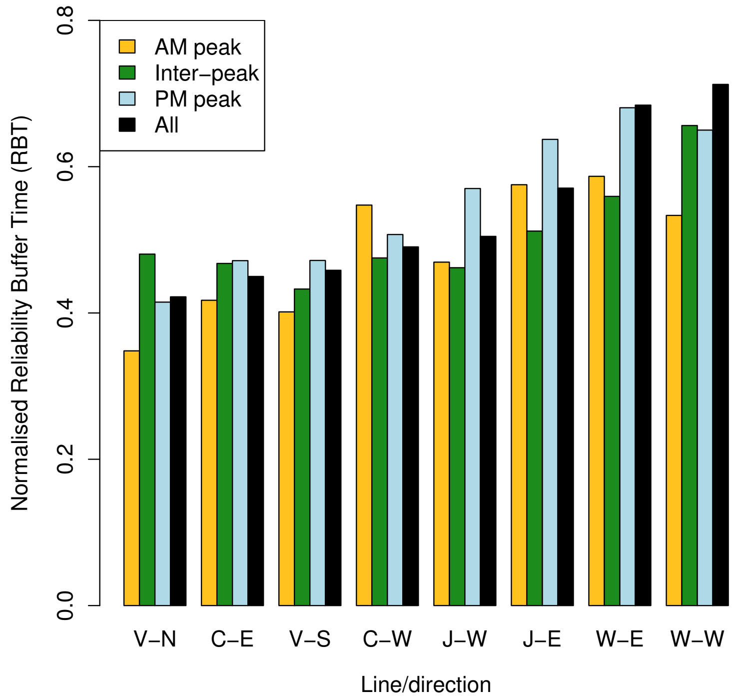

Figure 4.6: Benchmarking format - normalised reliability buffer time by time period

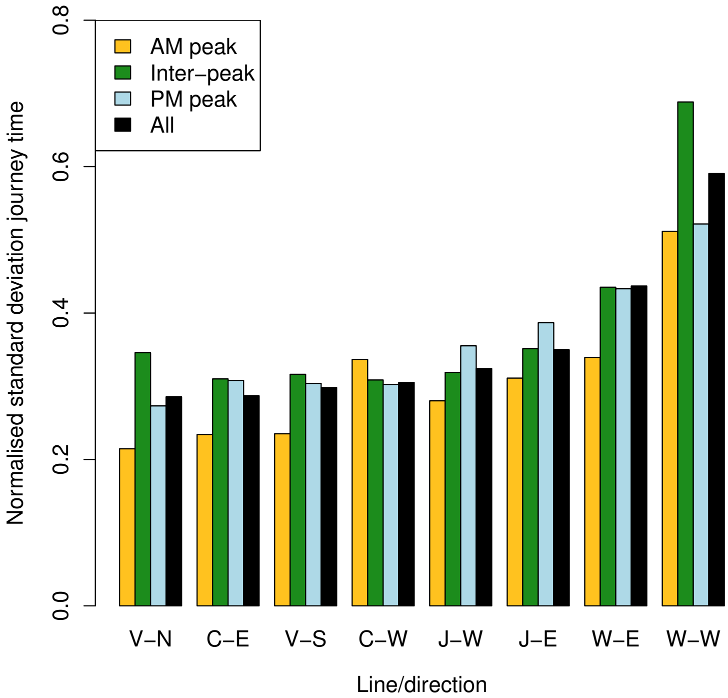

Figure 4.7: Benchmarking format - normalised standard deviation of journey time by time period

The first metric provides an indication of the mean journey times relative to the free flow times. The value of the measure is greater than 1 for all lines, indicating that mean journey times are longer than the free flow time across the entire day in all cases. In terms of the performance of the lines by the aggregated time periods, higher values are observed during the AM and PM peak periods as expected. The mean journey times are longest during the PM peak on most lines, on average 30% longer than the free flow time. The Waterloo and City line eastbound is the exception, where mean journey times are longest during the AM peak by a factor of 1.53 compared to the free flow time. Overall, mean journey times on the Waterloo and City line are longer compared to the other lines (see Figure 4.5).

In terms of the total daily values, the mean journey times on the Waterloo and City line are on average 42% and 55% higher than the free flow time in the east and westbound directions, respectively; while mean times on the other lines range from 25% to 38% higher than the free flow time (see Table 4.9). The results therefore indicate that the Waterloo and City line has the highest levels of passenger demand relative to the free-flow capacity provided, compared to the other lines. It is indeed a unique line serving only two stations, connecting Waterloo, a major national rail terminal, with the City of London. Infrastructure constraints limit train capacity to approximately half of that of the other lines, with a lower service frequency. Compared to the other lines, the higher degree of imbalance between the demand and capacity is the most likely cause of the higher values of the normalised mean indicator for the line.

The second and third performance measures give an indication of the spread of journey times. The RBT measure describes the variance of journey times at the tail end of the distribution (50th to 95th percentile) and is more suited to capturing the variance in journey times experienced by passengers on a “typical” day. The normalised standard deviation measure reports on the total spread of journey times including extreme outliers, and is arguably more useful to operators in identifying routes which are highly variable and susceptible to long delays. Both indicators are normalised by the free flow time to enable comparability across lines.

Unlike the mean journey time measure where common trends are observed across all lines, the normalised RBT appears to vary by line. Journey times on the Jubilee line, Victoria line southbound, and Waterloo and City line eastbound are most variable during the PM peak; journey times on the Victoria line northbound and Waterloo and City line westbound are most variable during the inter-peak; the Central line westbound has the most variable journey times during the AM peak; and journey times on the Central line eastbound are approximately equally most variable during the inter-peak and PM peak.

The normalised standard deviation measure gives similar results to the normalised RBT measure (see Table 4.9). For the Jubilee line, the normalised standard deviation is greater during the PM peak compared to other times of the day. For the Victoria and Waterloo and City lines, the inter-peak period is more variable in both directions. For the Central line eastbound, the inter-peak and PM peak are approximately equally most variable. On lines where journey times are more variable during the PM peak, congestion from high passenger demand levels and/or service irregularity are the most likely causes of the variance. On the Victoria line northbound and Waterloo and City line westbound where journey times are more variable in the inter-peak, fewer services are scheduled at these times compared to peak periods. This could lead to a greater degree of variability in platform wait times, and therefore a greater degree of variance in total journey times. Service irregularity could again have an effect during these times.

In terms of the overall daily journey time variance on the lines, Table 4.9 shows that the Waterloo and City line west and eastbound are most variable with average RBT values approximately 68% and 71% of the free flow time, and average standard deviations of journey time 44% and 59% of the free flow time, respectively. Journey times on the Victoria line northbound are least variable with an average RBT value 42% of the free flow time, and an average standard deviation 28% of the free flow time.

Incident-affected travel

Table 4.10 shows a comparison of the journey time performance for the incident-affected vs regular distributions by line and time period. The "I" (incident) and "R" (regular) values refer to the mean value of the performance measure aggregated at line level. The I/R ratio therefore reflects the degree to which incident events influence the values of the indicator, relative to regular conditions. The I/R ratio for each metric is also shown in graphical format in Figures 4.8, 4.9, and 4.10 by the AM peak, inter-peak, PM peak, and daily time periods, for easier interpretation by operators, as undertaken for the regular metrics.

Table 4.10: Incident data - journey time performance measures by line, direction, and time period

Line	Direction	Time period	Mean/FF I	Mean/FF R	Mean/FF I/R	RBT/FF I	RBT/FF R	RBT/FF I/R	s/FF I	s/FF R	s/FF I/R
Central	Eastbound	AM peak	1.26	1.19	1.06	0.25	0.42	0.59	0.18	0.23	0.75
		Inter-peak	1.31	1.23	1.07	0.25	0.47	0.53	0.24	0.31	0.78
		PM peak	1.28	1.25	1.03	0.22	0.47	0.47	0.19	0.31	0.61
		Total	1.35	1.24	1.08	0.41	0.45	0.90	0.34	0.29	1.17
	Westbound	AM peak	1.17	1.23	0.95	0.04	0.55	0.08	0.09	0.34	0.28
		Inter-peak	1.56	1.22	1.28	0.26	0.48	0.54	0.35	0.31	1.15
		PM peak	1.38	1.26	1.10	0.34	0.51	0.68	0.31	0.30	1.02
		Total	1.77	1.38	1.28	0.31	0.49	0.63	0.37	0.31	1.18
Jubilee	Eastbound	AM peak	1.27	1.27	1.00	0.27	0.58	0.46	0.29	0.31	0.93
		Inter-peak	1.36	1.24	1.09	0.49	0.51	0.97	0.29	0.35	0.83
		PM peak	1.32	1.29	1.02	0.43	0.64	0.68	0.27	0.39	0.70
		Total	1.71	1.28	1.33	0.71	0.57	1.25	0.51	0.35	1.47
	Westbound	AM peak	1.27	1.25	1.01	0.33	0.47	0.69	0.20	0.28	0.71
		Inter-peak	1.36	1.23	1.11	0.51	0.46	1.11	0.31	0.32	0.97
		PM peak	1.26	1.29	0.97	0.26	0.57	0.45	0.20	0.36	0.56
		Total	1.58	1.34	1.18	0.79	0.51	1.55	0.53	0.34	1.55
Victoria	Northbound	AM peak	1.24	1.20	1.03	0.25	0.35	0.72	0.16	0.21	0.73
		Inter-peak	1.34	1.23	1.09	0.35	0.48	0.74	0.28	0.35	0.81
		PM peak	1.28	1.24	1.03	0.32	0.41	0.78	0.21	0.27	0.77
		Total	1.62	1.25	1.30	0.76	0.42	1.79	0.51	0.28	1.79
	Southbound	AM peak	1.31	1.24	1.05	0.28	0.40	0.69	0.19	0.24	0.79
		Inter-peak	1.40	1.21	1.16	0.41	0.43	0.94	0.31	0.32	0.98
		PM peak	1.35	1.27	1.07	0.34	0.47	0.72	0.24	0.30	0.79
		Total	1.52	1.32	1.16	0.79	0.45	1.77	0.51	0.29	1.74
W&C	Eastbound	AM peak	2.37	1.53	1.54	0.99	0.59	1.69	0.73	0.34	2.16
		Inter-peak	NA	1.33	NA	NA	0.56	NA	NA	0.44	NA
		PM peak	1.84	1.30	1.41	1.17	0.68	1.72	0.65	0.43	1.50
		Total	1.95	1.42	1.38	0.82	0.68	1.20	0.55	0.44	1.25
	Westbound	AM peak	NA	1.30	NA	NA	0.53	NA	NA	0.51	NA
		Inter-peak	1.53	1.51	1.01	1.08	0.66	1.65	0.50	0.69	0.73
		PM peak	2.44	1.50	1.63	1.76	0.65	2.71	0.78	0.52	1.49
		Total	2.15	1.55	1.39	1.40	0.71	1.96	0.63	0.59	1.07

Note: FF is free flow time, RBT is the reliability buffer time, and s is the sample standard deviation, I denotes incident-affected trips, R denotes regular trips, I/R denotes ratio of incident-affected to regular trips, and NA indicates that no incidents are observed.

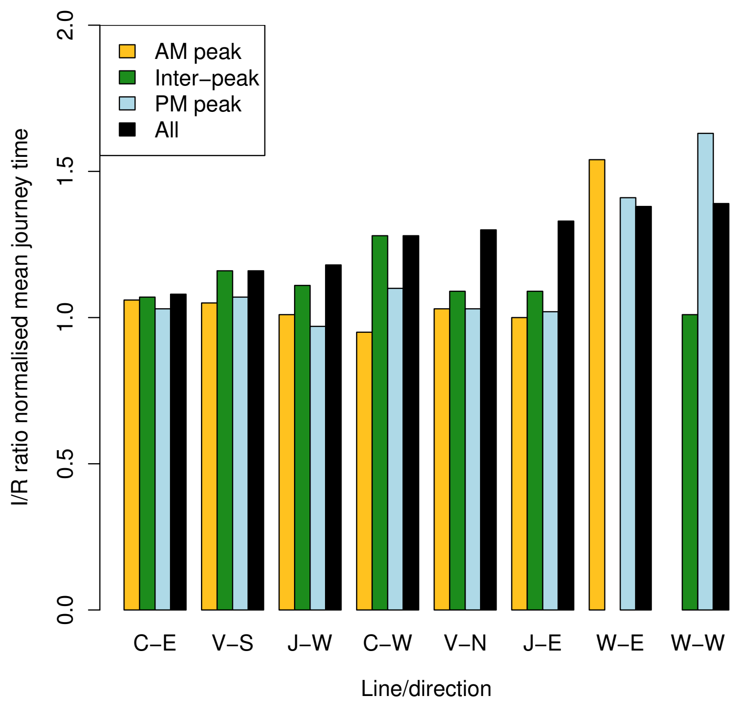

Figure 4.8: Benchmarking format - I/R ratio normalised mean of journey time by time period

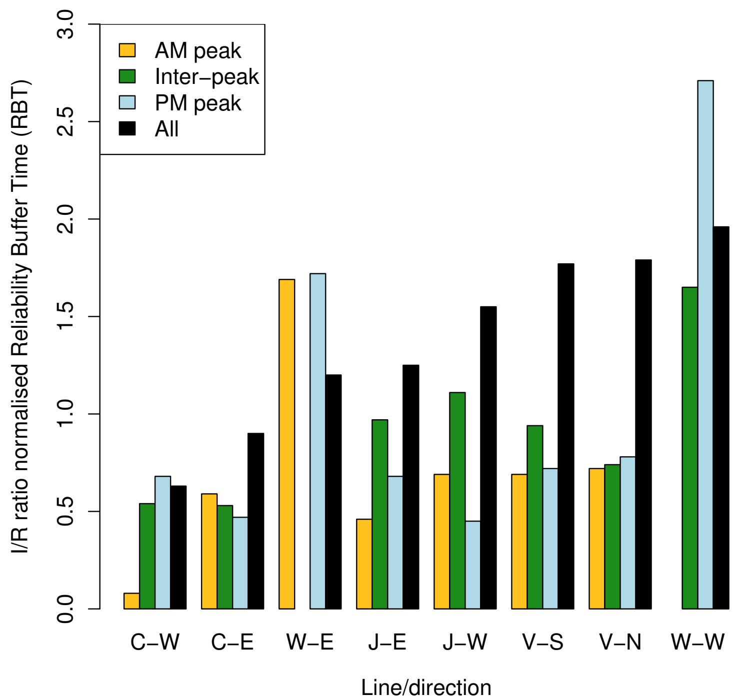

Figure 4.9: Benchmarking format - I/R ratio normalised reliability buffer time by time period

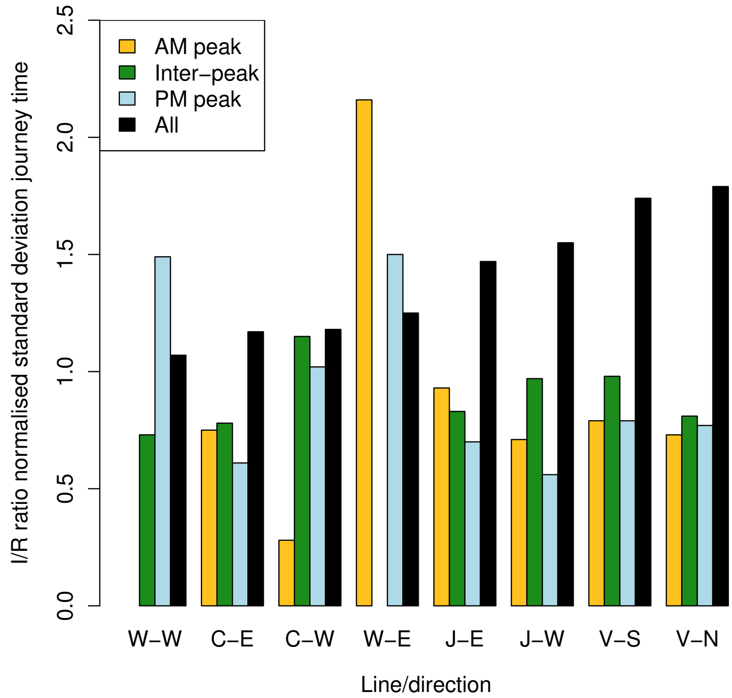

Figure 4.10: Benchmarking format - I/R ratio normalised standard deviation of journey time by time period

In terms of the normalised mean indicator, the results show that the incident-affected trips generally have higher mean journey times relative to the corresponding regular distributions across all lines, as expected. Only two observations in Table 4.10 have I/R ratio values lower than 1, namely the Central line westbound during the AM peak where the ratio is 0.95 and the Jubilee line westbound during the PM peak where the ratio is 0.97. Inspecting the incident data, on average only 4 trips are affected by incidents on the Central line westbound during the AM peak, and an average of 13 trips are affected during the PM peak on the Jubilee line westbound. The corresponding average number of passengers during regular conditions are 36 and 881 for the Central line westbound in the AM peak and Jubilee line westbound in the PM peak, respectively. It can therefore be inferred that the incident events only influence very few trips to a lower degree of severity, compared to the regular fluctuations in journey times across many more trips on these lines.

The I/R ratio for the normalised mean is greatest during the inter-peak for all lines except the Waterloo and City line. Incidents on the Waterloo and City line eastbound have the greatest effect during the AM peak with an I/R ratio value of 1.54, and in the westbound direction, incidents have the greatest effect during the PM peak with an I/R ratio value of 1.63.

Referring to the aggregated daily values, across all lines, the average value of the I/R ratio is 1.26 indicating that under incident-affected conditions, the mean journey times are 26% higher than the corresponding mean journey times under regular conditions. Incidents on the Waterloo and City line have the greatest effect, resulting in overall daily mean journey times 1.95 and 2.15 times higher than the free flow time in the east and westbound directions, respectively. In terms of the I/R ratio, this corresponds to mean journey times 38% and 39% higher during incident events compared to regular conditions in the east and westbound directions, respectively. As the Waterloo and City line is a single OD pair line and also rates as having the highest normalised mean journey time under regular conditions, this finding is as expected.

The normalised reliability buffer time (RBT) and standard deviation measures show more mixed results for the variance of journey times under incident conditions. Comparing across the AM peak, PM peak, and inter-peak time periods, the minimum value of the I/R ratio for the normalised RBT indicator is 0.08 for incident-affected trips on the Central line westbound during the AM peak, while the maximum value is 2.71 during the PM peak for the Waterloo and City line westbound. The majority of the indicator values (17 out of 22) by time period are less than 1, implying that the reliability buffer time associated with the incident-affected journeys is generally lower than the journey times in regular conditions. The normalised standard deviation indicator gives similar results, with 17 out of 22 of the I/R ratio values again being less than 1 across the three time periods. The minimum value of the normalised standard deviation I/R ratio is 0.28 for the Central line westbound during the AM peak, and the maximum value is 2.16 for the Waterloo and City line eastbound during the AM peak.

For the lines and time periods where the indicators of journey time variance under incident conditions are less than those for regular conditions, the corresponding mean journey times are generally higher under incident conditions. As discussed in the description of the data set, the number of passenger trips affected by incidents is relatively low with 0.5% of all trips considered as being incident-affected. The low number of affected trips may be attributed to fewer passengers being able to commence and/or complete their journeys under incident conditions, or that the incident events may be of low severity. Nonetheless, the fewer number of trip observations under incident conditions lead to a lower degree of variance in journey times compared to regular conditions, and the data show that these trips are distributed around higher mean journey times as a result of the incident events.

When aggregated over the total daily operational time period from 6am to 12am, the I/R ratios of the variance indicators show that incidents have a higher effect compared to indicators for the AM peak, PM peak, and the midday inter-peak. This indicates that the incidents that take place after the PM peak, and prior to the AM peak have a greater effect on the variance of journey times. Across all lines and directions, the mean value of the daily I/R ratio for the normalised RBT indicator is 1.38, while the daily I/R ratio for the normalised standard deviation indicator is 1.40. Therefore, the influence of incidents in the after-hours time periods prior to the AM peak and after the PM peak results in journey times being approximately 40% more dispersed compared to regular conditions.

4.5 Conclusions

In this chapter, journey time distributions are analysed to quantify the performance of the Central, Jubilee, Victoria, and Waterloo and City lines on the London Underground using data obtained from the Oyster smart card system. Journey times under regular conditions are analysed, as well as the impact of incidents on journey times.

The first part of the analysis defines the form of the journey time distributions. In line with recent literature, the long tailed Burr and log logistic distributions perform best, providing the best fit to approximately 86% of the distributions under regular conditions, jointly. In terms of the incident-affected distributions, 35% of the distributions adhere to a multimodal form, indicating the presence of both congested and non-congested travel conditions within a 15 minute time period of analysis.

Three performance measures are developed to assess the mean and variance of journey times.

The first indicator measures the 15 minute mean normalised by the free flow time. The second and third indicators provide measures of variance via the 15 minute reliability buffer time normalised by the free flow time, and the 15 minute standard deviation normalised by the free flow time.

Applying these indicators to the data, mean journey times are found to be longer during the PM peak, compared to the AM and midday inter-peak periods across all lines. In terms of line performance, journey times are longest (on average 49% longer) than the free flow time on the Waterloo and City line. The Central line eastbound performs best, with mean journey times 24% higher than the free flow time. In terms of variance, the indicators show that journey times tend to be more variable during the PM and midday inter-peak periods, depending on the line. Overall, journey times on the Waterloo and City line are most variable, and least variable on the Victoria line. For the incident data, the mean journey times are found to be higher under incident conditions compared to regular conditions, ranging from 8% to 39% higher depending on the line. The variance indicators also report higher levels of journey time variance under incident conditions across all lines when aggregated at a daily level.

The results presented here can be applied by urban rail operators to better understand, spatially and temporally, where journey time quality is relatively stronger or weaker, as experienced by their customers. Moreover, incident events have been shown to influence journey times to varying degrees, and such new information may help operators to better target their limited resources when attempting to improve service. Operators can use the methods presented here to benchmark customer travel time performance between lines, and the methods can be extended to benchmark at a network level to compare performance across different urban rail systems.

Chapter 5

Passenger to train assignment

5.1 Introduction

Travel by metro comprises of distinct phases involving different passenger movement activities. The stages of a typical metro passenger journey involves: 1) access time from the entry to the platform, including platform wait time at the origin station, 2) on-train time, 3) interchange time (if required), and 4) egress time from the platform to the exit at the destination station. It is known that passengers attach different levels of disutility to the different journey segments. For example, wait time has a higher value of time and is therefore more costly than uncrowded in-vehicle time (Transport for London, 2013; Wardman, 2004).

In order to evaluate the true economic impact of travel time reduction measures, the component of journey time which will be affected must be known. For this reason, modelling journey time decomposition is essential for demand modelling and investment appraisal. The journey times from the automated fare collection trip data (AFC) data set report the total journey times only. To enable the journey times to be split into the constituent components of access time, on-train time, and egress time, the itinerary of the train that the passenger boarded must be known. If information on train movements is known, the passenger trip data and train movement data can be merged, and passengers can be assigned to trains using probabilistic assignment methods.

A number of studies have been undertaken on the assignment of passengers to trains. The earliest assignment methods make use of published train frequencies and timetables (Nguyen and Pallottino, 1988; Spiess and Florian, 1989; Nuzzolo et al., 2001; Poon et al., 2004; Cepeda et al., 2006; Hamdouch and Lawphongpanich, 2008; Schmocker et al., 2008). Since the widespread availability of automated data sources, probabilistic assignment methods using AFC trip data and automated vehicle location (AVL) data have been proposed. In a number of methods, the main limitation is that the assumptions of passenger behaviour reflect more idealised versions of travel conditions. Kusakabe et al. (2010) assign passengers based on the shortest path that minimises wait times and lost times, however the possibility of missed boardings or longer wait times is not accommodated. Hong et al. (2016) and Sun and Schonfeld (2015) do make an allowance for missed boardings, but this is limited to the possibility of only one missed train and three missed trains, respectively. In reality, passengers experience a broader set of travel conditions, including exposure to highly congested situations, which can be observed during peak periods of demand in the morning and afternoon, and during service disruptions from incident events.

The assignment methods proposed by Paul (2010), Zhu (2014), and Zhu et al. (2017) do not prescribe as strict assumptions on passenger boarding behaviour, however, the main limitations of these methods is that additional input data from manual sources is required including passenger walk speed surveys, and detailed information of individual station layouts. To perform an assignment on an entire network, the collection and processing of this data is time-consuming and costly, and is not as efficient for operators to implement in practice. Hörcher et al. (2017) propose a method that addresses the main limitations as follows: 1) the full set of feasible train itineraries are considered, and no upper bound restrictions are applied on passenger wait times or missed boardings, and 2) the assignment is undertaken using the AVL and AFC data only. However, the main limiting factor of this assignment method is that the distribution of egress times from the set of unambiguously allocated passengers is used. A potential bias may be introduced whereby passengers tend to be allocated with quicker egress times as the upper bound of egress times is defined by the headway between consecutive trains.

In this chapter, the work undertaken by Hörcher et al. (2017) is built upon; an assignment method based on the unambiguous egress time distributions is proposed, making use of automated data sources only. The identified potential bias is addressed by using the distribution of unambiguous egress times in the off-peak period only, as headways between trains are longer during the off-peak, and the egress times of unambiguous passengers are therefore less restricted. Furthermore, an alternate method of calculating the egress time probabilities is proposed. Bayes' Theorem is applied, whereby the prior probability of the given egress time is updated by selecting and conditioning on a randomly sampled egress time from the observed unambiguous distribution in the off-peak. This gives a posterior probability for the egress time with a higher degree of accuracy. A series of statistical tests are also undertaken to more rigorously verify that the assignment is non-biased. The results from the passenger to train assignment algorithm presented in this Chapter are subsequently applied in Chapters 6 and 7 in the analysis of how different service supply and demand characteristics affect each component of journey time.

The chapter is structured as follows: a literature review of the existing assignment methods is presented in section 5.2, including the contributions of the proposed assignment method. The study area within the London Underground and a brief discussion of the data sets are then presented in section 5.3. The passenger to train assignment algorithm is presented in section 5.4. The analysis and discussion of the results are given in section 5.5, and the conclusions are summarised in section 5.6.

5.2 Literature review

The assignment of passengers to trains is a broad area of research, and the most commonly cited work is reviewed in the following sections. The review begins with a brief discussion on the earliest assignment methods, which make use of published train frequencies and timetables. This is followed by a review of the most current work on assignment methods, which use automated data sources to track passenger and train movements. It is important to note that the train assignment methods reviewed here are based on the assumption that all possible routes that a passenger can take between an origin-destination (OD) pair are known prior to the train assignment process. Consequently, a review of route choice methods is not presented here.

5.2.1 Prior to automated data availability

Prior to the widespread use of automated data sources, passenger to train assignment was initially estimated via average planned service frequencies, followed by methods using timetabled departure and arrival times for individual trains. The frequency-based methods produce average passenger loads by line section and time period (Spiess and Florian, 1989; Nguyen and Pallottino, 1988; Cepeda et al., 2006; Schmocker et al., 2008). The schedule-based methods produce a more disaggregate output, giving passenger loads for each individual train by time period according to the scheduled train timetables (Nuzzolo et al., 2001; Poon et al., 2004; Hamdouch and Lawphongpanich, 2008). Although the schedule based methods give a more granular measure of train loadings compared to the frequency methods, there is still potential for error, as unplanned train delays and other daily variations in train movements are not able to be accommodated.

5.2.2 Automated data methods

Automated vehicle location (AVL) data captures the real-time information of train arrivals and departures, and coupled with passenger automated fare collection (AFC) data, trains can be assigned to passengers while accounting for any irregularities in service. This produces results with a much higher degree of accuracy compared to the frequency and schedule-based methods. The next section begins with a summary of the general concept of train assignment using AVL and AFC data sources, followed by a review the empirical methods, and the main limitations of the methods.

General train assignment concepts

The general conditions of assignment for train j to be considered as a feasible option for passenger i are as follows:

1. Passenger i enters the system at the origin station platform before train j departs:

$$t_i^{entry} \leq DT_{oj} \tag{5.1}$$

2. Passenger i exits the system after train j arrives at the destination station:

$$t_i^{exit} \geq AT_{dj} \tag{5.2}$$

where t_i^{entry} and t_i^{exit} are the entry and exit times of passenger i, DT_{oj} is the departure time of train j at the origin station o, and AT_{dj} is the arrival time of train j at the destination station d.

By applying conditions 5.1 and 5.2, there are two possible outcomes:

1. Unambiguous trips - only one possible train itinerary.
2. Ambiguous trips - more than one possible train itinerary.

The general method of assignment is as follows: the set of unambiguous trips are first identified using the conditions of assignment described in equations 5.1 and 5.2. The set of ambiguous trips are then assigned using one of two general methods. The first option is that the ambiguous trips are assigned based on prescribed assumptions about passenger travel behaviour. The second option is that inferences about passenger travel behaviour are made based on the properties of the set of unambiguous trips. The set of ambiguous trips are then allocated to trains using probabilistic assignment algorithms.

Empirical assignment methods

The following section reviews the set of studies that allocate passengers to trains using prescribed assumptions about passenger behaviour, independent of the set of unambiguous trips.

Kusakabe et al. (2010) develop a method to assign passengers to trains using automatic fare collection data and the published train timetable for the urban rail system in Osaka, Japan. The main assumption is that passengers will minimise wait time at the origin station, and minimise lost time due to early arrival at the destination station. For cases where interchanges are required, the passenger will chose the route that minimises the number of transfers. A cost function for each possible train option is developed based on the wait times and number of transfers. Dijkstra's algorithm is then applied to select the shortest path, i.e. the train option where travel cost is minimised.

Paul (2010) uses train movement and smart card data to assign passengers to trains on the London Underground. The assignment is based on the assumption that access and egress walk times through stations are individual-specific. The distribution of access and egress times are generated for each station based on manual walk speed survey data collected by London Underground. Trains are then assigned such that the individual's access time percentile in the access time distribution is greater than or equal to the individual's egress time percentile in the egress time distribution, after allowing for an assumed wait time.

Sun and Schonfeld (2015) analyse the Beijing metro system using AFC data and train movement data. The key assumption of the assignment method is that access and transfer times are equivalent to the minimum walk times in free-flow conditions. Therefore, at times when the system is more congested, passengers are termed to be left behind as they cannot board the first available train that minimises their access and/or transfer walk times. In the application of the algorithm, the authors introduce an upper bound whereby the maximum number of times that a passenger can be left behind is three. Based on these assumptions, weightings are applied to the set of feasible train itineraries based on the probability of the passengers failure to board at the origin station and/or transfer stations. The train with the highest likelihood of boarding is then selected as the most feasible. Zhao et al. (2017) use a similar failure to board concept for their assignment method of passengers to trains using data from the Shenzen metro in China.

The following studies assign the set of ambiguous passengers by deriving passenger travel characteristics from the set of unambiguously allocated passengers.

Hong et al. (2016) develop an assignment algorithm for passengers on the Seoul metro in South Korea. Boarding is assumed to operate under a first-come-first-serve basis, and it is also assumed that passengers will either board the first or second feasible train. The set of unambiguous trips are referred to as the reference passenger set. With each iterative assignment of ambiguous trips, the access and egress times of the reference set are updated with each newly assigned passenger. As such, the assignment may begin with a small pool of unambiguous access and egress times, but the set increases as the assignment progresses. For the set of ambiguous trips, the probability of a passenger choosing the train is based on the proportion of reference passengers that chose that train. For the Seoul metro system, the minimum headway between trains is 3.5 mins, the mean egress times is 1.9 mins, and egress times are found to be independent of the time of day.

Zhu (2014) analyses the Hong Kong metro system, and proposes an assignment method for a subset of trips during the off-peak period where transfers and route choice are not considered. The assignment method is probabilistic, and Bayes Theorem is used to calculate the probability of the passenger boarding the train given their tap-out time. The probabilities are computed based on access and egress times for each feasible train. The access and egress time distributions are derived via maximum likelihood using the distribution of access and egress distances and walk speeds at each station, and access and egress time distributions from unambiguously assigned trips. Information on the distribution of passenger walking is obtained from manual surveys, and information on access and egress walk distances are obtained from station layout plans. Zhu et al. (2017) extends this method, and presents a more general assignment method applicable to both peak and off-peak travel, accounting for cases where passengers are left behind in congested conditions.

Hörcher et al. (2017) performs the assignment for the Hong Kong metro system using the AVL and AFC data sets only, without additional manual data inputs from surveys and station plans. This is an advantage for operators as the method can be implemented with readily available data. The method is applicable to trips that involve up to two transfers, and route choice is also incorporated. The assignment is performed on the Hong Kong metro system. First, unambiguous trips are assigned. From the unambiguous trips, an unambiguous egress time distribution is generated and used to assign passengers in ambiguous cases where there are no transfers involved. Probabilities are computed for each feasible train based on the likelihood of the given egress time, relative to the combined likelihood of egress times from all feasible trains (see equation 5.3). For cases where there are one or more transfers, the likelihood of a passenger boarding a given train is based on the access and egress time distributions generated from the unambiguous trips.

$$P(C_i|I) = \frac{P(C_i)}{\sum_{j \in S} P(C_j)} \tag{5.3}$$

where

C_i refers to the itinerary of candidate train i

I is the event of observing a total set S of feasible itineraries

5.2.3 Limitations of reviewed methods and thesis contributions

The passenger to train assignment methods reviewed require assumptions to be made regarding passenger travel behaviour. The limitations associated with these assumptions are briefly discussed in the following section, along with a summary of the contributions of the proposed assignment method presented in this chapter.

In a number of methods, the main limitation is that the assumptions of passenger behaviour reflect more idealised versions of travel conditions. In reality, passengers experience a broader set of travel conditions, including exposure to highly congested situations, which can be observed during peak periods of demand in the morning and afternoon. Kusakabe et al. (2010) assign passengers based on the shortest path that minimises wait times and lost times. No explicit consideration is given to the likely possibility of missed boardings or longer wait times, which is typical during peak periods. Hong et al. (2016) do make an allowance for missed boardings, however, this is limited to the possibility of only one missed train. Sun and Schonfeld (2015) allow for the possibility of up to three missed trains, and discard any additional train itineraries beyond this. Although it is less likely that a passenger would miss more than three trains, discarding feasible itineraries excludes the observation of extreme travel conditions that a passenger may be experiencing, for example, in extreme cases of service disruptions.

The assignment methods proposed by Paul (2010), Zhu (2014), and Zhu et al. (2017) require input data from manual surveys on passenger walking speeds and information on access and egress walk distances at an individual station level. From this data, distributions of walk speeds and distances are generated for each station. To perform an assignment on an entire network, the collection and processing of this data is time-consuming and costly. These methods are therefore not as efficient for operators to implement, compared to methods that solely make use of the widely available automated data sources.

Hörcher et al. (2017) propose a method that addresses the two previously discussed limitations: 1) the full set of feasible train itineraries are considered, and no upper bound restrictions are applied on passenger wait times or missed boardings, and 2) the assignment is undertaken using the AVL and AFC data only. However, as mentioned in Zhu (2014) and Zhu et al. (2017), the limiting factor of the assignment method is that the use of the distribution of egress times from the set of unambiguous passengers may introduce a bias. By definition, the upper bound of the set of unambiguous egress times is equivalent to the headway between consecutive trains. On systems where the minimum headway between trains is shorter than the typical egress time of passengers, the set of unambiguous trips represents the fastest passengers travelling through the system. In this case, therefore, the assignment of ambiguous passengers is biased towards allocating passengers with quicker egress times. Passengers that may move slower through the system are either given less emphasis, or excluded, depending on the value of the minimum headway.

The assignment method proposed in this chapter builds on the assignment method proposed by Hörcher et al. (2017), which is fully based on automated data sources on passenger trips and train movements. Three main contributions are made: first, the potential bias currently present in the algorithm is addressed through the use of egress times in the off-peak period; second, a Bayesian extension is applied to increase the degree of accuracy in calculating the egress time probabilities for feasible train itineraries; and third, additional statistical tests are undertaken to more rigorously verify that the underlying assumptions are unlikely to cause severe bias.

5.3 Study area and data

In this analysis, three lines on the London Underground are analysed bidirectionally: the Central, Jubilee, and Victoria lines. As mentioned in Chapter 3, the majority of the train movement timestamps for Waterloo and City line are incomplete, and so this line is not included in the passenger to train assignment analysis. The study area is therefore defined as follows: the entire length of the Victoria line (16 stations); West Acton to Oxford Circus on the Central line (12 stations); and Bond Street to North Greenwich on the Jubilee line (10 stations) (see Figure 3.1, Chapter 3 for geographical extents). Single line trips that originate and terminate on the line of interest are considered, and line sections are chosen such that there are no other probable routes between OD pairs under normal operating conditions.

Weekday data (Monday to Friday) over seven weeks from October to December in 2013 are analysed. Two data sets over this time period are used. The first is the automated fare collection passenger trip data which is administered and recorded through the TfL Oyster smart card system. The passenger trip data includes records of each entry and exit timestamp and location for each trip. The second data set is the automated vehicle location train movement data which is recorded in the TfL NetMIS system.

As mentioned in Chapter 3, the train departure timestamps are recorded for the majority of train movement records, however, the arrival timestamps of trains at each platform are not recorded for the majority of the train trips. Consequently, the train assignment can only be undertaken using the train departure timestamps. As a result, train dwell times at stations are not able to be accounted for explicitly, but are instead subsumed in other components of journey time as shown in Figure 5.2. At the origin station, dwell times are included in the access time component and at the destination station, dwell times are included in the on-train time component of journey time.

Another discrepancy lies between the accuracy of the train and passenger trip timestamps. As discussed in Chapter 3, the train movement data are accurate to 1 second in the form HH:MM:SS, while the passenger trip data are accurate to 1 minute in the form HH:MM. In order to be conservative, the assumption is made that the passenger could have entered at the origin station and exited at the destination station at any time within the minute. This assumption leads to an increase in the possible number of trains that the passenger could feasibly board. According to the truncated trip timestamps, the passenger can arrive at the origin station at any time between HH:MM:00 and HH:MM:59. If a train's departure timestamp at the origin is at some time within this range at HH:MM:XX, the conservative assumption is made that the passenger arrived at the earliest possible time close to or at HH:MM:00, and so the train itinerary is included in the set of feasible itineraries. Similarly, if a train departs at the destination station within the timestamp bounds, the conservative assumption is made that the passenger exits at the latest possible time closest to or at HH:MM:59. Under this assumption, the train is included as a feasible itinerary. For further information on the properties and limitations of the data, a detailed discussed is presented in Chapter 3.

The kernel density distributions of passenger trips and train departure times by time of day are shown in Figure 5.1. As shown in the plot, the number of passenger entries peak during the morning and afternoon, and the number of train departures correspondingly peak at these times to meet the higher levels of passenger demand. In the peak periods, trains operate at higher frequencies and shorter headways compared to the off-peak periods. The distinction between the two time periods is a key concept that is used in the development of the passenger to train assignment algorithm, and is discussed further in section 5.4.

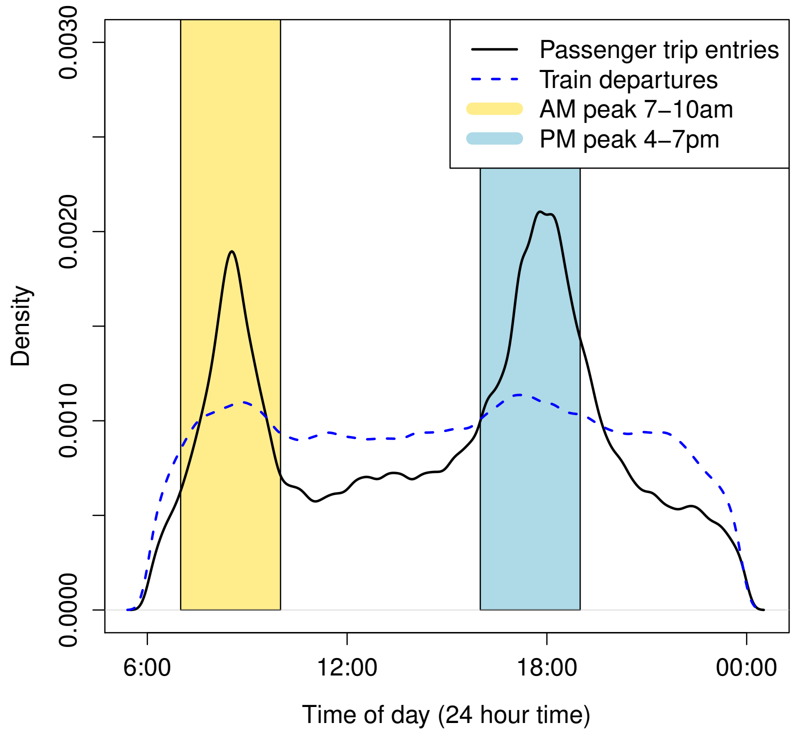

Figure 5.1: Distribution of passenger trip entries and train departures by time of day

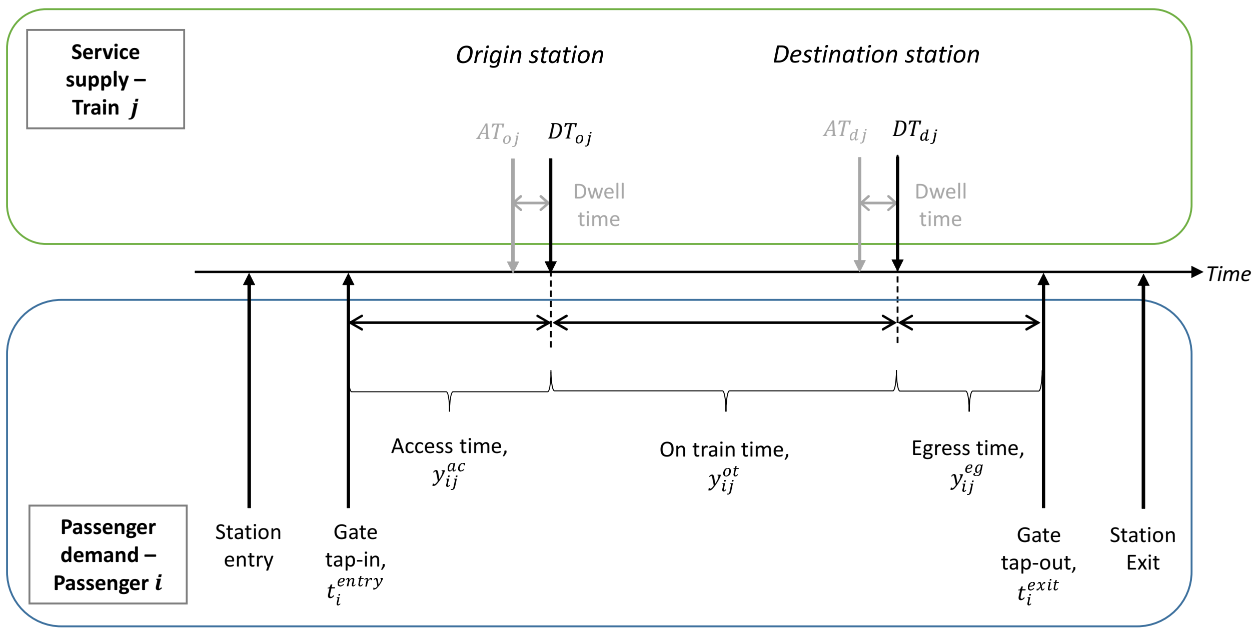

Figure 5.2: Journey time components using available train movement data

5.4 Methods

5.4.1 Proposed assignment algorithm

Passengers are able to be assigned to individual trains by merging the AFC trip data with the AVL train movement data. Based on the entry and exit times recorded by the AFC system, and the train departure timestamps at station platforms from the AVL data, the two sets of timestamps can be matched. In the following sections, the methodology for the assignment of passengers to trains is presented.

Based on the train departure timestamps and the passenger entry and exit timestamps, two general conditions must be satisfied for train j to be considered as a feasible option for passenger trip i as follows:

1. Passenger i must enter the system at the origin station platform before train j departs:

$$t_i^{entry} \leq DT_{oj} \tag{5.4}$$

2. Passenger i must exit the system after train j departs at the destination station:

$$t_i^{exit} \geq DT_{dj} \tag{5.5}$$

where

t_i^{entry} and t_i^{exit} are the entry and exit times for passenger trip i,

DT_{oj} is the departure time of train j at the origin station, and

DT_{dj} is the departure time of train j at the destination station.

By applying conditions 5.4 and 5.5, there are two possible outcomes:

1. Unambiguous trips - Train j is the only feasible train for passenger trip i. The set of unambiguous trips is herein referred to as $\mathcal{U}$.

2. Ambiguous trips - Passenger trip i can be feasibly allocated to more than one train $j = 1, .., J_i$, where J_i is the total number of feasible trains for passenger trip i. The set of ambiguous trips is herein referred to as $\mathcal{A}$.

For clarity, the application of train assignment conditions in equations 5.4 and 5.5 are illustrated in Figure 5.3. Passengers 1 and 2 both arrive at the origin station after the departure of train 1, but prior to the departure of trains 2, 3, and 4. Passenger 1 exits the destination station after the arrival of train 2, but prior to the arrival of train 3. Therefore, passenger 1 has only one feasible itinerary and is unambiguously allocated to train 2. In contrast, passenger 2 exits the destination station after the arrival of train 3 and before the arrival of train 4, and so this passenger could have feasibly boarded trains 2 and 3. The trip by passenger 2 is therefore classified as an ambiguous trip.

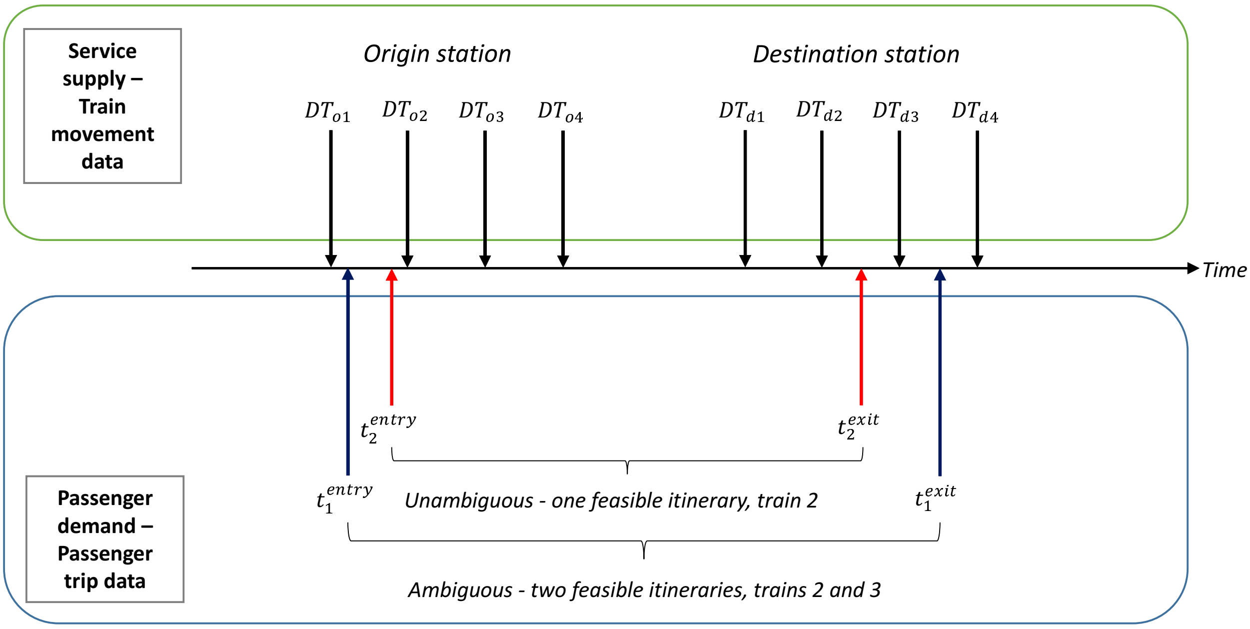

Figure 5.3: Example of train assignment - unambiguous and ambiguous trips

In order to allocate the ambiguous trips, a probabilistic algorithm is applied. The algorithm is an extension of the method presented by Hörcher et al. (2017), and is based on deriving the probabilistic likelihood of the egress time associated with each possible train itinerary using the AFC and AVL data sets only. The passenger is allocated to the train itinerary with the highest probable egress time.

The egress time probabilities are computed using the distribution of egress times from the set of unambiguously assigned passengers. The upper bound of egress times in the set of unambiguous trips is the headway between consecutive trains. As mentioned in the literature review, if headways are short, the set of unambiguous trips represent the fastest passengers only. The resulting assignment of ambiguous trips based on the set of unambiguous trips will therefore be biased toward the selection of the shortest egress time option. To reduce the likelihood of biased selection, the distribution of headways in the data set must be analysed. In peak times, train frequencies are higher in order to accommodate a higher level of passenger demand. Across all three lines on the London Underground, the peak times of train frequency extend from 7am to 10am in the morning and from 4pm to 7pm in the afternoon. Here, off-peak times are defined as all other operating times excluding the peak times, i.e. prior to 7am, 10am to 4pm, and after 7pm.

The distributions of headways between the peak times and the off-peak times of the day are compared. As noted in Chapter 3, the maximum timetabled train frequency on the three lines is 35 trains per hour, which corresponds to a minimum headway of approximately 1.67 minutes. In highly congested conditions, trains may operate at headways less than the scheduled minimum headway as a result of the train bunching phenomenon. In the raw data set, a minimum train headway of 0.02 minutes is observed, however, these are likely erroneous data. In the interest of reflecting realistic operational conditions, the headway data have been cleaned to eliminate itineraries where the headway is recorded as less than 1 minute (0.07% of the data set). An upper bound of 20 minutes is also applied, as only approximately 0.4% of the data are recorded with headways greater than 20 minutes. The summary statistics for cleaned train headway data are given in Table 5.1 and the distribution of headways by time period are shown in Figure 5.4. In peak times, it is shown that the headway between trains is shorter compared to the off-peak times in terms of the mean and median values and the range of headways.

Table 5.1: Train headways by time period

	Summary statistics (mins)					Mann-Whitney U test	
Time period	Min.	Median	Mean	Max.	s	Test statistic	p-value
Peak	1.00	2.08	2.50	19.98	1.30	4.59E+12	2.20E-16
Off-peak	1.00	2.67	3.01	19.82	1.42		

Note: s refers to the sample standard deviation.

To establish whether the difference in the distribution of headways between the two time periods is statistically significant, an unpaired non-parametric Wilcoxon rank-sum test, also referred to as the Mann-Whitney U test, is performed. The test evaluates whether there is a

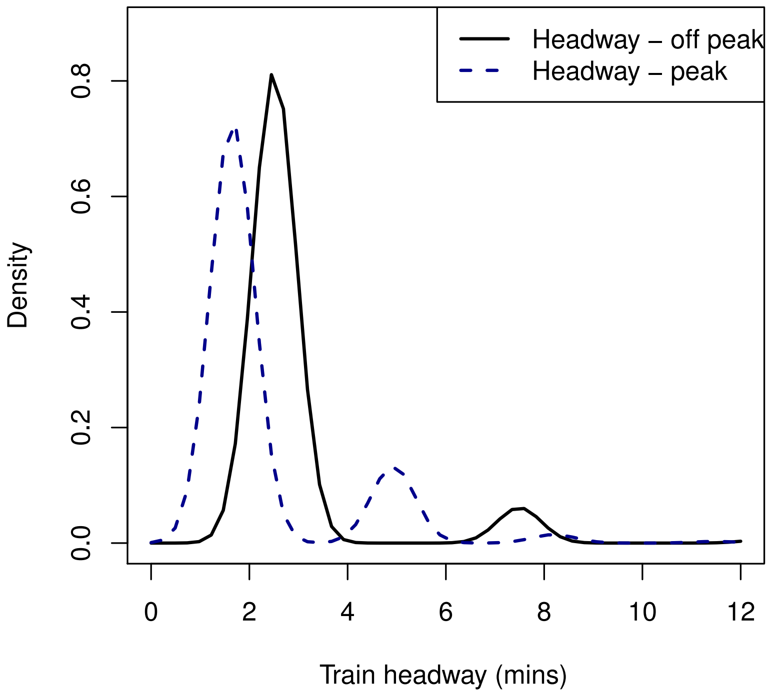

Figure 5.4: Distribution of train headways by time period

difference between two independent distributions. Unlike other tests of this form, assumptions regarding the underlying theoretical distribution are not required to be made. The headways are non-normally distributed, rather they appear to follow a right-skewed form, and so the use of the non-parametric Mann-Whitney test here is appropriate. The observations from both distributions are ordered by magnitude from lowest to highest, and then rankings are assigned to each observation. The degree of difference between the two distributions is assessed by computing a test statistic based on the number of observations in each distribution and the sum of the rankings for each distribution as per equation 5.6 (Lehmann and Romano, 2005).

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_{Ui} \tag{5.6}$$

Where U is the Mann-Whitney U test statistic, n_1 is the number of observations in distribution 1, n_2 is the number of observations in distribution 2, and R_{Ui} is the ranking of observations for both distributions.

The null hypothesis is that the distributions do not differ and so each observation in distribution 1, x_{1i}, has an equal chance (50%) of being greater than (or less than) each observation in distribution 2, x_{2i}, i.e. $H_0 : P(x_{1i} > x_{2i}) = 0.5$. Here the test is performed to determine whether headways in the off-peak are greater than headways in the peak, and so the alternative hypothesis is directional. Assuming that x_{1i} is an observation in the distribution of off-peak headways and x_{2i} is an observation in the distribution of peak headways, the alternative hypothesis can be expressed as $H_1 : P(x_{1i} > x_{2i}) > 0.5$. As shown in Table 5.1, the p-value for the test is 2.2E-16, which indicates that the null hypothesis cannot be accepted at a significance level of greater than 99%. The test therefore indicates that there is a difference between the two distributions, and that the headways in the off-peak are longer than the headways in the peak.

Through the test, it is established that the distribution of headways are quicker in the peak with a mean value of 2.50 minutes, compared to headways in the off-peak which are distributed around a mean value of 3.01 minutes. If egress times are indeed influenced by headways between trains, using the egress time distribution of unambiguous trips in the peak period may introduce a bias as the headway between trains acts as an upper bound. To reduce the likelihood of potential bias, the egress times of unambiguous trips in the off-peak are used in the assignment algorithm, as the headway between trains is longer compared to peak times. The set of unambiguous trips in the off-peak is defined as follows:

$$\mathcal{U}_{op} = \mathcal{U}(7\text{am} > t_i^{entry} > 10\text{am} \ \& \ 4\text{pm} > t_i^{entry} > 7\text{pm}) \tag{5.7}$$

From the set of unambiguous trips in off-peak times, the probability distribution functions of egress times are constructed. The general form of the equation to calculate egress times is defined in equation 5.8.

$$y_{ij}^{eg} = t_i^{exit} - DT_{dj} \tag{5.8}$$

Where t_i^{exit} is the exit timestamp of passenger trip i at the destination station, DT_{dj} is the departure time of train j at the destination station, and y_{ij}^{eg} is the egress time.

The set of ambiguous trips, $\mathcal{A}$, are allocated based on the probability of the egress time for each potential train itinerary. The probability is calculated through application of Bayes' Theorem. The prior probability of the given egress time is updated by selecting and conditioning on a randomly sampled egress time from the observed distribution of unambiguous egress times in the off-peak.

The prior probabilities for the egress time of the train itinerary under consideration and the randomly sampled egress time are computed using the distribution of egress times from the total set of unambiguous trips $\mathcal{U}$. The general form of the empirical density function is given in equation 5.9.

$$f_1(x) = \hat{f}_k(y_{ij}^{eg}), \qquad y_{ij}^{eg} \in \mathcal{U} \tag{5.9}$$

Where $f_1(x)$ is the empirical probability density function constructed from the total set of unambiguous trips $\mathcal{U}$. The empirical density function is derived via the kernel density estimator $\hat{f}_k(\cdot)$. As mentioned in the literature review chapter, this estimation method is non-parametric, and so the form of the probability density function is generated from the data points. A kernel density function corresponding to a normal distribution is applied, and a smoothing parameter is included to capture the trade-off between matching the data points and the smoothness of the function. The general form of the kernel density estimator for a given variable of interest x is expressed in equation 5.10 (Scott, 2015).

$$\hat{f}_k(x) = \frac{1}{nh_k} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h_k}\right) \tag{5.10}$$

Where $k(\cdot)$ is the kernel density function which corresponds to a normal probability density function, h_k is the smoothing parameter, and n is the number of observations of the given variable of interest x (Scott, 2015).

The updated probability for the randomly sampled egress time is computed using the distribution of egress times from the set of unambiguous trips in the off-peak $\mathcal{U}_{op}$. The mathematical definition is given in equation 5.11.

$$f_2(x) = \hat{f}_k(y_{ij}^{eg}), \qquad y_{ij}^{eg} \in \mathcal{U}_{op} \tag{5.11}$$

Where $f_2(x)$ is the empirical probability density function constructed from the total set of unambiguous trips in the off-peak $\mathcal{U}_{op}$, and $\hat{f}_k(\cdot)$ and y_{ij}^{eg} are as previously defined.

The Bayesian probabilities of the egress times associated with each feasible train itinerary for the set of ambiguous trips are therefore calculated as per equation 5.12 as follows:

$$P(y_{ij}^{eg}|r_{ij}^{eg}) = \frac{P(r_{ij}^{eg}|y_{ij}^{eg})P(y_{ij}^{eg})}{P(r_{ij}^{eg})}, \qquad y_{ij}^{eg} \in \mathcal{A}, r_{ij}^{eg} \in \mathcal{U}_{op} \tag{5.12}$$

where

y_{ij}^{eg} is the egress time for passenger i and train itinerary $j = 1, ..J_i$ in the set of ambiguous trips, i.e. $y_{ij}^{eg} \in \mathcal{A}$.

r_{ij}^{eg} is a randomly sampled egress time from the set of unambiguous trips in the off-peak period, i.e. $r_{ij}^{eg} \in \mathcal{U}_{op}$.

$P(y_{ij}^{eg})$ is calculated by applying equation 5.9, i.e. $f_1(y_{ij}^{eg})$.

$P(r_{ij}^{eg})$ is calculated by applying equation 5.9, i.e. $f_1(r_{ij}^{eg})$.

$P(r_{ij}^{eg}|y_{ij}^{eg})$ is calculated by applying equation 5.11, i.e. $f_2(r_{ij}^{eg})$.

The passenger is then assigned to the train itinerary with the highest probable egress time as per equation 5.13:

$$P(y_{ij,assigned}^{eg}) = \max_{1 \le j \le J_i}(P(y_{ij}^{eg}|r_{ij}^{eg})), \qquad y_{ij}^{eg} \in \mathcal{A}, r_{ij}^{eg} \in \mathcal{U}_{op} \tag{5.13}$$

Where

$j = 1, .., J_i$ represents the feasible train itineraries, with J_i being the total number of feasible trains for passenger trip i, and

$y^{eg}_{ij,assigned}$ is the egress time of the passenger for the assigned train.

For clarity, a general outline of the train assignment process is summarised in Figure 5.5.

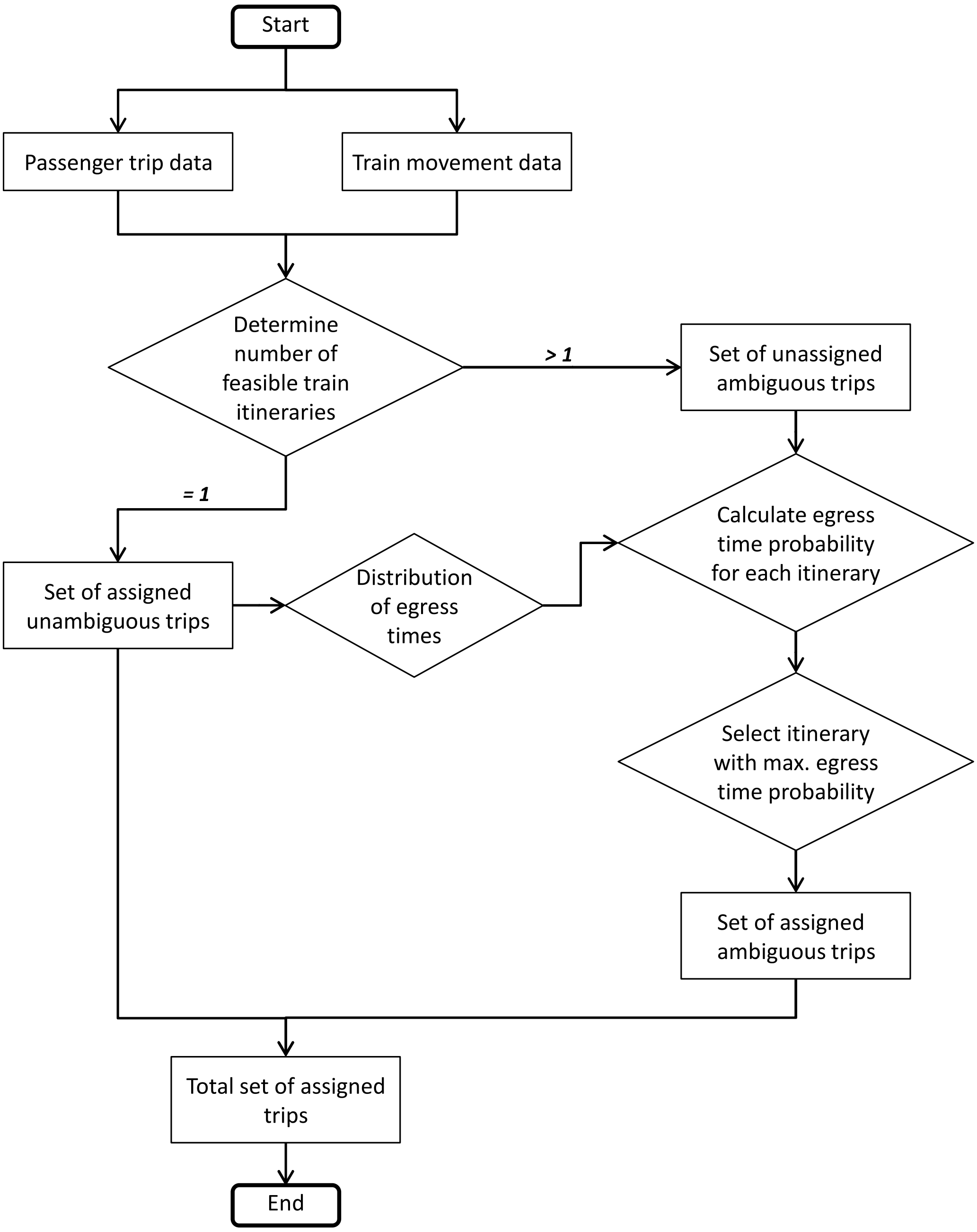

Figure 5.5: Summary of passenger to train assignment process

5.4.2 Comparison of relative performance of proposed assignment algorithm

The new assignment algorithm has been proposed on the basis that the use of egress times from unambiguous trips in the off-peak reduces the potential bias present in the method proposed by Hörcher et al. (2017), where probabilities are calculated using the distribution of the total set of unambiguous egress times. The total set of unambiguous egress times incorporates the fastest passengers moving through the system during peak periods, and so the use of this distribution would tend apply higher weightings to the allocation of faster egress times.

The assumption that egress times assigned through new method are more likely to be slower than egress times assigned through the Hörcher et al. (2017) method is tested and documented in Appendix D. As mentioned in the Appendix, the generation of appropriate simulated data is not within the scope of this thesis, and so the two methods have been tested on a sample set of data from the London Underground Oyster trip data and train movement data. The two assignment methods are performed on approximately 75,100 trips on the OD pair from Finsbury Park to Oxford Circus on the Victoria line. The results from the analysis demonstrate that egress times produced from the new assignment method are indeed slower than egress times produced from the Hörcher et al. (2017) method. It can therefore be stated that the new assignment method does reduce the relative degree of bias in favouring faster egress times present in the method proposed by Hörcher et al. (2017). As mentioned in Appendix D, future research is recommended to trial the two methods on high-quality simulated data, to compare how well each method performs relative to known "true" train itineraries.

5.5 Results

Performing the assignment according to the entry and exit time conditions given in equations 5.4 and 5.4, the following proportion of trips in the data set are unambiguously allocated: 32% of trips on the Central line, 28% of trips on the Jubilee line, and 32% of passengers on the Victoria line (see Table 5.2). In the remaining set of ambiguous trips (approximately 69%), the mean number of train options for a single passenger is 2.46 options.

Table 5.2: Summary of unambiguous trips vs ambiguous trips by line and direction

Line	Direction	Total trips	No. unambiguous	No. ambiguous	Mean no. options
Central	Eastbound	497184	175152	322032	2.42
	Westbound	460900	127091	333809	2.54
Jubilee	Eastbound	1293600	315700	977900	2.44
	Westbound	1455791	458577	997214	2.52
Victoria	Northbound	1856457	642357	1214100	2.45
	Southbound	2318831	692173	1626658	2.44
Total		7882763	2411050	5471713	2.46

As mentioned previously, the set of unambiguous trips represent the fastest passengers, as the headway between consecutive trains is the upper bound for passenger egress times. Headways are on average shorter during the peak periods compared to off-peak periods, and so it can be inferred that the egress times of the unambiguous passengers would be shorter in the peak compared to the off-peak. Consequently, in the train assignment process, the distribution of egress times in the off-peak is used.

To verify that it is indeed appropriate to use egress times in the off-peak rather than the peak, the two sets of egress times are compared. Table 5.3 provides a summary of egress times by time period. In Table 5.3, the mean values of egress times are the same in both time periods, however, the median value is higher in the off-peak at 1.30 minutes compared to the median in the peak at 1.28 minutes. To further establish whether there is a statistically significant difference in egress times between the two time periods, the Mann-Whitney U test is applied. The null hypothesis is that the distributions do not differ, i.e. $H_0 : P(x_{1i} > x_{2i}) = 0.5$ where x_{1i} is an observation in the distribution of off-peak egress times and x_{2i} is an observation in the distribution of peak egress times. The test is performed to determine whether egress times in the off-peak are greater than egress times in the peak, and so the alternative hypothesis is $H_1 : P(x_{1i} > x_{2i}) > 0.5$. As shown in Table 5.3, the p-value for the test is 2.7E-11, which indicates that the null hypothesis cannot be accepted at a significance level greater than 99%. The test therefore indicates that there is a difference between the two distributions, and that the egress times in the off-peak are longer than the egress times in the peak. The use of the distribution of unambiguous egress times in the off-peak for the assignment of ambiguous trips is therefore appropriate to reduce any level of potential bias.

Table 5.3: Egress times of unambiguous trips by time period

	Summary statistics (mins)			Mann-Whitney U test	
Time period	Median	Mean	s	Test statistic	p-value
Off-peak	1.30	1.45	1.04	5.49E11	2.69E-11
Peak	1.28	1.45	1.09		

Note: s refers to the sample standard deviation.

To further investigate whether the headway between consecutive trains does indeed influence the egress time distribution used in the assignment, the distribution of egress times of unambiguous trips in the off-peak is compared with the corresponding distribution of headways as shown in Figure 5.6 and Table 5.4.

Table 5.4: Summary statistics of egress times and headways of unambiguous trips in off-peak

	Summary statistics (mins)			Mann-Whitney U test		Pearson correlation		
Time period	Median	Mean	s	U-stat	P-value	r_{xy}	t-stat	P-value
Egress time	1.30	1.45	1.12	1.0E11	2.2E-16	-0.05	-56.5	2.2E-16
Headway	3.15	3.76	2.05					

Note: s refers to the sample standard deviation, U-stat refers to the Mann-Whitney U test statistic, r_{xy} refers to the Pearson correlation coefficient, and t-stat refers to the t-test statistic.

If egress times are indeed influenced and limited by the headway between trains, it would be expected that the distribution of egress times would mirror the distribution of headways. As shown in Figure 5.6 and Table 5.4, this is not the case, as egress times are distributed around a lower mean value of 1.45 minutes compared to the mean of headways at 3.15 minutes.

To establish whether there is a statistically significant difference between the egress times and headways in the off-peak, the Mann-Whitney U test is applied. The null hypothesis is that the distributions do not differ, i.e. $H_0 : P(x_{1i} > x_{2i}) = 0.5$ where x_{1i} is an observation in the

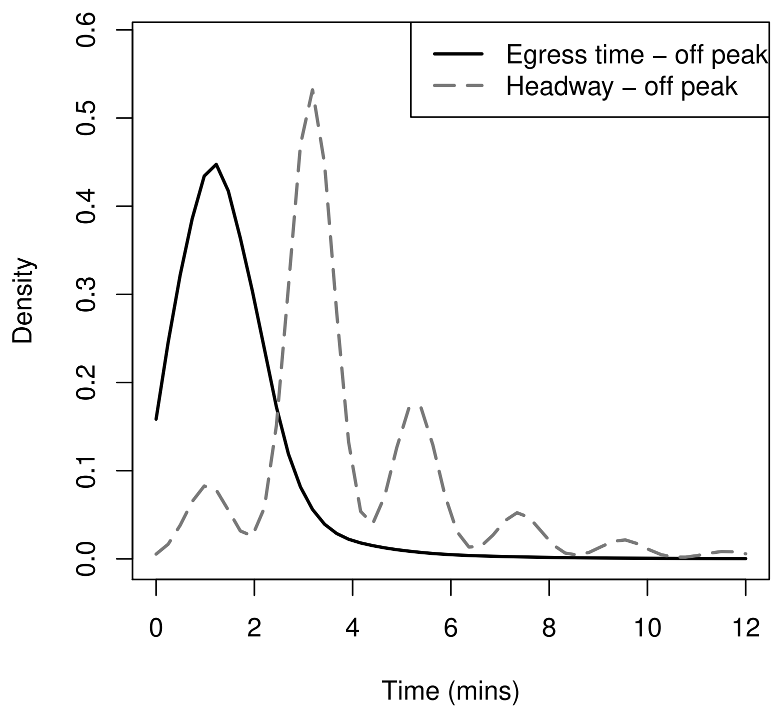

Figure 5.6: Distribution of egress times and headways for unambiguous trips in off-peak

distribution of egress times of unambiguous trips in the off-peak and x_{2i} is an observation in the distribution of the corresponding headways. The test is undertaken to determine whether the egress times are shorter than headways, and so the alternative hypothesis is $H_1 : P(x_{1i} > x_{2i}) < 0.5$. As shown in Table 5.4, the p-value for the test is 2.2E-16, which indicates that the null hypothesis cannot be accepted at a significance level greater than 99%. The test therefore indicates that there is a difference between the two distributions, and that egress times are shorter than headways.

An additional test, the Pearson correlation test, is performed to determine whether egress times are correlated with headways. If a bias is present in the assignment, this would be evidenced by a positive correlation, which would indicate that egress times increase with headways. The Pearson correlation coefficient measures the degree of linear correlation between two variables, and is calculated as per equation 5.14. A value of 1 indicates a positive correlation, a value of -1 indicates a negative correlation, and a value of 0 indicates no correlation (Wonnacott and Wonnacott, 1990). The Pearson correlation coefficient can be tested for statistical significance by performing a t-test; the test statistic is given in equation 5.15. The null hypothesis for the test is that there is no correlation between the variables, i.e. $H_0 : |r_{xy}| = 0$, while the alternative hypothesis indicates that there is a degree of correlation, i.e. $H_1 : |r_{xy}| > 0$ (Wonnacott and Wonnacott, 1990).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5.14}$$

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} \tag{5.15}$$

Where r_{xy} is the Pearson correlation coefficient, x_i and y_i are observations of the two variables being tested, $\bar{x}$ and $\bar{y}$ are the corresponding mean values of the two variables, n is the number of observations, and t is the t-test statistic distributed on Student's t-distribution with $n - 2$ degrees of freedom.

As shown in Table 5.4, the t-test indicates that the null hypothesis for the Pearson correlation coefficient cannot be accepted at a significance level greater than 99%. This suggests that there is a degree of correlation between the egress times and headways, however, the value of the correlation coefficient is -0.05. The correlation coefficient therefore implies that there is a very weak negative correlation, indicating that egress times tend to decrease as headways increase to a minimal degree. As mentioned previously, if the correlation coefficient was positive, there would be evidence that the egress times are biased and limited by the upper bound of the headways, however this is not the case. Overall the statistical tests indicate that the use of the egress times of unambiguous trips in the off-peak are not likely to be restricted or biased by headways between consecutive trains. The use of this set of egress times in the assignment of ambiguous trips is therefore valid.

5.6 Conclusions

In this chapter, a probabilistic assignment algorithm is proposed to assign individual passengers to individual trains on the Central, Jubilee, and Victoria lines on the London Underground. It should be noted that the method is readily able to be extended to include route choice and transfers on the London Underground, as well as other metros. By matching the passenger entry and exit timestamps to the train movement timestamps at station platforms, two conditions can arise. The first is when the trip and train timestamps indicate that only one train itinerary is feasible; this set of trips are termed unambiguous trips. The second condition arises when the timestamps indicate that the passenger could have feasibly boarded more than one train itinerary; this set of trips are termed ambiguous trips.

A Bayesian assignment algorithm is proposed to allocate the ambiguous trips, based on the distribution of egress times for unambiguous trips in the off-peak. The new proposed algorithm is performed on a sample of data from Finsbury Park to Oxford Circus on the Victoria line to test whether it reduces the level of bias towards favouring faster train itineraries as generated by the assignment method developed by Hörcher et al. (2017) (see Appendix D). It is found that the new assignment method does indeed produce longer egress times compared to the Hörcher et al. (2017) method, and so it can be stated that the new method does reduce the level of bias toward the selection of train itineraries associated with faster egress times.

The assignment algorithm is then used to perform an assignment for trips within the defined study area. Approximately 31% of trips on the three lines are unambiguously allocated and the remaining 69% of ambiguous trips are allocated using the Bayesian assignment algorithm. A further series of statistical tests are performed, and it is demonstrated that the use of unambiguous trips in the off-peak gives non-biased assignment results on this set of data.

The allocation of passengers to trains enables the total journey times recorded by the passenger trip data to be decomposed into their constituent components, namely: access time, on-train time, and egress time. The passenger assignment results derived here are used as an input in the analyses presented in the following Chapters 6 and 7.

Chapter 6

Decomposing journey times via semiparametric regression

6.1 Introduction

Subject to service capacity, demand levels, and individual passenger behaviour, there can be high levels of journey time variance within each component of a passenger transit trip. Operators seek to minimise this journey time variance to provide reliable service and remain attractive to customers. There are many avenues available to operators to improve journey time performance, from targeting train operations such as increasing service frequencies or train speeds, to station interventions such as providing more platform space or applying dynamic crowd control measures. The issue is that it is unclear how well these interventions perform; the magnitude of the impact on journey times is difficult to quantify, and comparing the performance of different interventions is similarly complicated. The focus of this chapter is to address these issues by: 1) disentangling the impact of supply and demand side characteristics on journey times at a disaggregate component level, and 2) using semiparametric regression modelling to provide a framework to improve the appraisal of operational policies relating to journey time performance.

As mentioned in the literature review, empirical studies have been undertaken to identify the causes of variance in journey times, however, these are based on modelling linear regression relationships. The linear regression studies mainly focus on bus transit (Ma et al., 2015; Diab and El-Geneidy, 2013; Yetiskul and Senbil, 2012; El-Geneidy et al., 2011; El-Geneidy and Vijayakumar, 2011; Surprenant-Legault and El-Geneidy, 2011; Tetreault and El-Geneidy, 2010; Mazloumi et al., 2008). Fewer studies have been undertaken to investigate the variance in passenger journey times for rail transit (Sun et al., 2012; Uniman, 2009; Krygsman et al., 2004; Taylor, 1982). The limitation of using linear models in this work is that key information about the relationship between variables can be omitted, and this can impact the accuracy of the parameter estimates.

Other empirical studies of rail journey times have been undertaken, however, these focus on analysing the components of journey time individually. The main areas of research include the analysis of train running times (Kecman and Goverde, 2015; Hansen et al., 2010; Murali et al., 2010; Gorman, 2009; Gibson et al., 2002), dwell times (Christoforou et al., 2017; Holloway et al., 2016; Oh et al., 2016; Thoreau et al., 2016; van den Heuvel, 2016; Seriani and Fernandez, 2015; Harris et al., 2014; Heinz, 2003; Wiggenraad, 2001), and platform wait times (Ingvardson et al., 2018; Nygaard and Torset, 2016; Frumin and Zhao, 2012; Fan and Machemehl, 2009; Csikos and Currie, 2008, 2007; Luethi et al., 2007). These studies are useful in providing information on the factors that influence the components of journey time considered in isolation, however, this work does not provide an holistic understanding of the relative influence of factors on both total journey times and the components of journey time for a single system.

In this chapter, a more unified approach to analyse the effect of various supply and passenger demand factors on journey times for the London Underground is proposed. From the assignment of passengers to trains presented in Chapter 5, total passenger journey times are decomposed into three constituent components: access, on-train, and egress times. A series of regression models are then developed to directly compare the relative effect of different factors on the three components of journey time, as well as total journey times. The analysis is undertaken using semiparametric regression, a method which has not been previously applied in this field. Using the semiparametric regression method, non-linear relationships between variables can be modelled to give a more robust understanding of why journey times vary and to what extent.

Journey time variance, as observed in raw smart card data, is partly caused by: 1) the fact that passengers are heterogeneous themselves, and 2) journey times may vary for a given individual, controlling for the decisions that they make in response to external travel conditions. It is important to distinguish between the two, as only the second source of variance results in disutility. This research disentangles the variance factors by modelling the effect of the following set of covariates:

1. Variables relating to service supply, including:
   i. Dynamic operational factors such as headways and train speed which vary with time, and
   ii. Static factors related to the physical elements of the system.
2. Variables relating to passenger demand, including:
   i. Passenger volumes at different points in the system, and
   ii. Passenger characteristics.

Elasticity estimates are derived for each covariate with respect to each component of journey time to quantify the relative effect of the different covariates. These estimates enable a direct comparison to be made in terms of which factors have the highest and lowest relative effect on journey times, which can in turn, be used by operators to guide where potential interventions should be made in order to improve journey time performance.

As an extension to the core analysis, the model outputs from the analysis of access times are used to investigate the effect of train headways on passenger platform wait times at the origin station. The prevailing theory in the literature is tested: that when train service frequencies are high in peak times, passengers arrive in a random manner, resulting in an average wait time equivalent to half of the headway (Kittelson & Associates Inc et al., 2013; Turnquist, 1978). As service frequency decreases in off-peak periods, it is postulated that passengers arrive in a non-random manner in order to minimise their wait times. In the most recent empirical work on rail transit covering both high frequency metro and low-frequency suburban rail modes, the transition from random to non-random passenger arrivals is reported to occur in the range from 5 to 11 minute headways (Ingvardson et al., 2018; Nygaard and Torset, 2016; Fan and Machemehl, 2009; Luethi et al., 2007). In this chapter, marginal passenger wait times are quantified by using the outputs of the semiparametric regression model of access times to compute the derivative of access times with respect to train headways. This is the first known application of advanced regression techniques coupled with the mathematical derivation of marginal platform wait times from passenger access times, while conditioning for other service supply and passenger demand factors.

The remainder of the chapter is structured as follows: the study area within the London Underground is defined in section 6.2, and the framework for the semiparametric regression models are presented in section 6.3. The analysis and discussion of the results are presented in sections 6.4 and 6.5, and the conclusions are summarised in section 6.6.

6.2 Study area and data

The results from the passenger to train assignment in Chapter 5 are used as the input data for the regression analysis. As such, the study area is as presented in Chapter 5; it covers the entire length of the Victoria line (16 stations), West Acton to Oxford Circus on the Central line (12 station), and Bond Street to North Greenwich on the Jubilee line (10 stations). Trips in both directions on the lines are analysed, and two additional conditions are imposed: first, the trips are single line trips that originate and terminate on the line of interest, and second, line sections are chosen such that there are no other probable routes between OD pairs under normal operating conditions. A map of the geographical analysis boundaries can be found in Figure 3.1, Chapter 3.

The data cover weekday travel (Monday to Friday) over seven weeks from October to December 2013. The total passenger journey times from entry at the origin station to exit at the destination station are recorded by the Transport for London (TfL) Oyster smart card system. The TfL NetMIS system records the train departure timestamps at each station in the network.

Additional data on the physical and operational characteristics of the lines and stations are used to model the supply and demand covariates in the models. For further detailed information on the properties and limitations of the data sets, refer to Chapter 3.

Using the assignment algorithm in Chapter 5, individual passenger trips are allocated to individual trains. This enables the total journey times for each trip to be decomposed into the constituent components of access, on-train, and egress times. Figure 6.1 illustrates the kernel density distributions of each journey time component. Egress times are on average the shortest component, followed by access times, and on-train times. The journey time decomposition is further discussed in section 6.3.

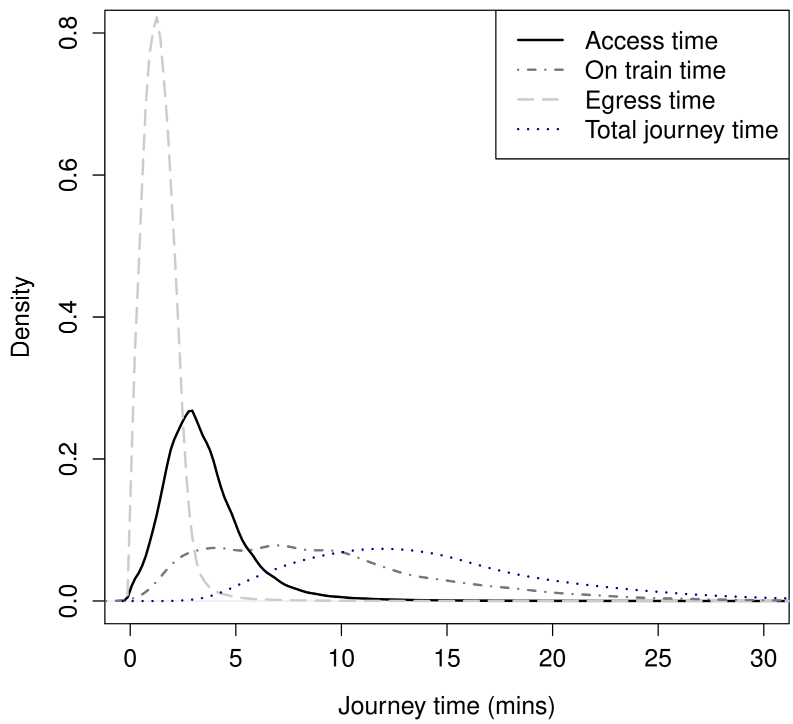

Figure 6.1: Distribution of journey times by journey time component

6.3 Methods

Semiparametric regression models are developed for each component of journey time. From the models, elasticities of journey time are calculated for different service supply and demand factors to determine to what degree these factors influence journey times. The following section presents the general framework for the regression models, including detailed descriptions of the dependent and independent model variables, and model evaluation criteria.

6.3.1 General regression model framework

The advantage of the semiparametric regression method is that no a priori assumptions are held regarding the nature of the relationship between the independent and dependent variables in the model. The relationship is instead able to be defined via basis functions generated from the data, therefore enabling non-linearities and potential discontinuities to be modelled (Wood, 2006). The basis functions are penalised thin plate regression splines, which are interpolation functions that are assigned to match the observed data points with a penalty applied to capture the trade-off between the degree to which the spline functions match the data and the degree of smoothness. Estimates of the smooth functions are obtained by minimising the general equation 6.1 (Wood, 2006).

$$||y - f_k||^2 + \lambda J(f_k) \tag{6.1}$$

where y is a vector of data points, f_k are the estimated smooth functions, $J(f_k)$ is the penalty function which measures the degree of 'wiggliness' of f_k, and λ_s is the smoothing parameter which controls the balance between the degree to which f_k fit the data and the smoothness of f_k.

Further information on the underlying mathematical theory of thin plate regression splines can be found in Wood (2006). The modelling is undertaken using the R statistical software package 'mgcv'. The 'bam' function designed to handle large data sets is used to estimate the models. The models are fitted using penalised iteratively reweighted least squares (PIRLS) and the restricted maximum likelihood (REML) technique is used to estimate the parameters of the models (Wood et al., 2015; Wood, 2006). Further details of the REML optimisation algorithm for semiparametric regression models are included in Appendix E.

The regression models adhere to a generalised additive mixed models (GAMMs) structure; the general form is given in equation 6.2.

$$g(y_i^c) = \alpha + X_i^T \beta + \sum_{k=1}^{K} f_k X_i^* + Z_i^{fT} c + \epsilon_i \tag{6.2}$$

where

y_i^c is component c of journey time evaluated for trip i with link function $g(\cdot)$,

α is the model constant,

β are the parameter coefficient estimates for the covariates X_i modelled parametrically,

$f_k, k = 1..K$ are the smooth basis functions based on penalised thin plate regression splines of the covariates X_i^* modelled non-parametrically,

c is a vector of estimated group-specific fixed effects for the fixed categorical factors Z_i^f, and

ϵ_i is the random error term, assumed to be independently and identically distributed with mean 0 and given variance σ_ϵ^2, such that $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

It should be noted that the regression modelling is undertaken on the aggregate data set of all trips within the defined study area, with the resulting error term capturing the residual random variance across all trips, conditional on covariates and fixed effects in the models.

As discussed in the literature review and further supported by the analysis in Chapter 4, journey times are typically distributed in a right-skewed form. As a result, log transformations of journey times and all covariates in the models are trialled. The link function $g(\cdot)$ for the dependent variable of journey times y_i^c is defined as the Gaussian identity function. The properties of the dependent variables are further discussed in section 6.3.2.

The independent variables, herein referred to as the model covariates, can be modelled in three ways. Variables which are measurable on a continuous scale are modelled either parametrically or non-parametrically, and these capture different service supply and demand attributes of the network. The first are the parametric variables, which are those variables that follow known parametric forms (for example, linear, quadratic, exponential). The second are the non-parametric variables which are modelled via thin plate regression splines generated from the data points. The continuous variables are specified further in sections 6.3.4 and 6.3.5.

Variables that can be considered as index variables for which longitudinal observations exist are modelled as categorical factors that have group-specific effects. These factors represent the unobserved variation between the given categories of the variable (Greene, 2018). In the analysis, group effects are assigned for the following: stations, OD routes, line/direction, and passenger characteristics.

The group effects are modelled using a fixed effects structure; this form is considered appropriate if the variable under consideration has been drawn from a finite population, and inferences regarding the effect of the variable are confined to the categories of the variable included the model (Greene, 2018; Searle et al., 1992). The fixed effects are assumed to be correlated with the other covariates in the model (Greene, 2018; Searle et al., 1992). In general mathematical terms, the fixed effects Z_i^f are correlated with the parametric and non-parametric covariates in the model X_i and X_i^*. The expected mean of the vector of the estimated fixed parameters c can be expressed as $E[c|X_i, X_i^*]$, with variance $Var[c|X_i, X_i^*]$ (Greene, 2018; Wood, 2006). As there are multiple covariates that concurrently capture characteristics of the network and passengers, correlation between the covariates is likely. Moreover, the stations, lines, and OD pairs are drawn from a finite population specific to the London Underground network, and so the fixed effects form is appropriate in this analysis.

6.3.2 Dependent variables

The components c of journey time for an individual passenger trip i assigned to train j are defined as follows:

1. Access time – time from passenger tap-in t_i^{entry} to departure of assigned train at origin

station DT_{oj}

$$y_{ij}^{ac} = DT_{oj} - t_i^{entry} \tag{6.3}$$

2. On-train time – time between train departure at the origin station DT_{oj} to departure at the destination station DT_{dj}

$$y_{ij}^{ot} = DT_{dj} - DT_{oj} \tag{6.4}$$

3. Egress time – time from train departing at destination station DT_{dj} to passenger tap-out time t_i^{exit}

$$y_{ij}^{eg} = t_i^{exit} - DT_{dj} \tag{6.5}$$

4. Total journey time – total time from passenger tap-in at the origin station t_i^{entry} to passenger tap-out at the destination station t_i^{exit}

$$y_i^{jt} = t_i^{exit} - t_i^{entry} \tag{6.6}$$

A summary of the descriptive statistics for each component and total journey time is provided in Table 6.1.

Table 6.1: Summary statistics of dependent variables

	Summary statistics (mins)			
Dependent variable	Min.	Max.	Mean	s
Access time	1.00E-6	85.87	3.55	2.19
On-train time	1.66E-2	114.17	9.23	5.60
Egress time	1.00E-6	88.25	1.42	0.93
Total journey time	2.00	121.00	14.20	6.09

Note: s refers to the sample standard deviation.

6.3.3 Model covariates

Through the review of the literature on journey time regression studies and from knowledge of the London Underground system, the model covariates are selected from the available data sets to capture sources of variance in journey times. All covariates are listed in Table 6.2 and are discussed in more detail in the following sections.

Table 6.2: Summary of regression covariates by model

Service supply covariates		Demand covariates	
Dynamic	Static	Passenger volumes	Passenger characteristics
Headway (All)	Platform length (A,E)	Plat. loading (All)	Passenger age (All)
cv headway (All)	Walk distance (A,E)	Line loading (All)	Card type (All)
Norm. headway (All)	Station age (A,E)		Discount (All)
Train speed (T,J)	Station zone (A,E)		Travel frequency (All)
Gates (A,E)	No. lines (A,E)		
Escalators (A,E)	Terminal station (A,E)		
Lifts (A,E)	Connection to NR (A,E)		
Stairs (A,E)	Platform screen doors (A,E)		
Platform points (A,E)	Platform open/closed (A,E)		
Time of day (All)	Stops (T,J)		
	Group effects for stations (A,E,J),		
	OD route (T,J), and		
	line/direction (All)		

Note: A – Access time model; T – On-train time model; E – Egress time model;
J – Total journey time model; All – All models, *cv* is the coefficient of variation,
NR refers to National Rail, Norm. refers to Normalised, and Plat. refers to platform.

6.3.4 Supply side covariates

Dynamic service supply covariates

Headway - The headway associated with each trip is obtained through the passenger to train assignment process. Different service frequencies are in operation at different times of the day on different lines. This leads to fluctuations in platform wait times, which correspond to fluctuations in journey times. In the regression models, the headway for the train that the passenger has boarded is used.

Coefficient of variation (cv) of headway - The coefficient of variation of headway is calculated over a 15 minute period for each OD pair. This covariate is included to capture the variation in train frequencies which would otherwise skew the model results, particularly at the beginning and end of service early and late in the day, and at the transition periods between peak and off-peak times. It should be noted that headway is the only covariate in the models where a measure of its associated variance is included. The variance of other covariates do not affect the model results to the same degree, and so have not been included in order to maintain model parsimony.

Headway normalised by mean headway - The train headway associated with each passenger trip is normalised by the mean value of headway over the corresponding period of 15 minutes for each OD pair. This covariate is included to specifically capture any potential effects of train bunching. This operating condition can arise at high levels of passenger demand, where train dwell times at stations can become longer than scheduled to accommodate higher levels of passenger boardings and/or alightings. As a result, headways between trains become shorter, and this is termed as train bunching (Kittelson & Associates Inc et al., 2013). As highlighted in Chapter 3, in the train movement data set, trains are observed to operate at headways shorter than the minimum scheduled operating headway of approximately 1.67 minutes. Approximately 15% of all trips in the data set are undertaken when trains operate at headways less than 1.67 minutes.

Train speed - Speed is included as a covariate in the on-train time and total journey time models, and is calculated by taking the train route distance of each trip and dividing by the on-train time of the trip.

Dynamic station elements - The number of gates are dynamically controlled at some stations to accommodate fluctuations in passenger demand. As precise information on gate control is not available, the number of available gates are multiplied by a usage factor as specified in equation 6.7. The same method is applied to the dynamic vertical transport elements of the stations including escalators, lifts, and stairs, as well as the number of platform entry and exit points.

$$e_{qt} = e\left(\frac{n_q}{n_{entries} + n_{exits}}\right)_t \tag{6.7}$$

where t is the time period in intervals of 15 minutes, $n_{entries}$ is the number of entries, n_{exits} is the number of exits, q denotes the quantity of interest (either entries or exits), e refers to the total number of station elements available (i.e. gates, escalators, lifts, stairs, or platform entry/exit points), and e_{qt} is the number of station elements dynamically available for the quantity of interest in the entry or exit direction.

Time of day - An additional covariate for the time of day is included to represent any residual time-varying conditions that are not captured by the other dynamic covariates. This covariate is measured in 15 minute intervals in 24 hour time.

A summary of the descriptive statistics for each dynamic service supply covariate is given in Table 6.3.

Table 6.3: Summary statistics of dynamic service supply covariates

Covariate	Units	Min.	Max.	Mean	s
Headway	min	1.00E-6	352.40	2.88	2.16
cv headway	Numeric	1E-6	3.22	0.28	0.19
Norm. headway	Numeric	3.6E-7	21.31	1.00	0.32
Speed	km/min	2.1E-2	896.40	0.60	1.06
Gates	Numeric	0.13	43.52	14.35	11.65
Escalators	Numeric	4.6E-8	17.76	3.09	3.49
Lifts	Numeric	1.9E-8	4.74	0.79	0.91
Stairs	Numeric	8.0E-8	99.20	20.13	21.59
Platform points	Numeric	0.04	10.79	2.56	1.74
Time of day	Categorical	72 categories - Four 15 min intervals for each hour from 6am to 12am			

Note: s refers to the sample standard deviation.

Static service supply covariates

The static covariates for the access and egress time models capture the fixed physical attributes of the stations, and include:

- The length of the station platforms, and a value of the average access/egress walk distance within the stations, which is measured as the distance between the midpoints of the station gatelines and platforms.
- The maximum age of the station, the presence of platform screen doors, and whether the station is open-air or underground.
- The connectivity of the station including (i) the TfL zone in which the station is located, (ii) the number of lines that the station services, (iii) whether the station is a terminal station at the end/start of the line, and (iv) whether the station is connected to the overground National Rail service.

The number of intermediate stops that a train makes along its route is also included in the on-train time and total journey time models.

A final set of static effects capture any residual group-specific effects of lines, OD routes, and stations. As mentioned previously, these group effects are modelled as fixed effects, thus accommodating potential correlation with the other station, line, and route-based covariates in the models.

A summary of the descriptive statistics for each static service supply covariate is given in Table 6.4.

Table 6.4: Summary statistics of static service supply covariates

Covariate	Units	Min.	Max.	Mean	s
Platform length	Metres	122.00	153.00	136.90	4.96
Walk distance	Metres	19.30	400.60	143.76	56.87
Station age	Years	15.00	174.00	104.40	46.14
Station zone	Numeric	1.00	3.00	1.51	0.66
Connecting lines	Numeric	1.00	6.00	1.93	1.09
Terminal station	Categorical	Dummy, value 1 or 0			
Connection to National Rail	Categorical	Dummy, value 1 or 0			
Platform screen doors	Categorical	Dummy, value 1 or 0			
Open/closed platform	Categorical	Dummy, value 1 or 0			
Intermediate stops	Numeric	1.00	15.00	4.30	2.45
Stations	Categorical	35 stations			
OD route	Categorical	451 OD pairs			
Line/direction	Categorical	6 line/direction combinations - 3 lines, 2 directions each			

Note: s refers to the sample standard deviation.

6.3.5 Demand covariates

Passenger volumes

Passenger demand surveys of the London Underground indicate that passenger levels fluctuate across the system spatially and temporally. Varying levels of passenger congestion are expected to influence journey times. Indicators of passenger density are evaluated at three points for each trip: platform loading at the origin station, inter-station line loading, and platform loading at the destination station. These indicators are included in the regression models to capture the effect of passenger volume on journey time. Paper ticketing was also available in 2013 alongside the Oyster card, and so to obtain an estimate of total passenger volumes including those from alternative ticketing modes, data from the TfL Rolling Origin and Destination survey (RODs) are used in the calculation of the passenger volume indicators.

The indicator for passenger volume ρ_{qt} for the quantity of interest q (platform loading at origin or destination, or line loading) is the number of trips n_{qt} during a 15 minute time period t, normalised by the average number of trips $\bar{n}_{qt,day}$ per 15 minute period t over the day:

$$\rho_{qt} = \frac{n_{qt}}{\bar{n}_{qt,day}} \tag{6.8}$$

Passenger characteristics

Passenger-level attributes capturing different passenger characteristics as recorded by the AFC data are also included in the models. The total number of trips a passenger makes on a given OD route over the seven week analysis period is modelled. For regular commuters that complete many trips, it is expected that their access and egress walk times are shorter as a result of route familiarity.

Information on passenger age, ticket type, and discount type are recorded by the Oyster system for each trip, and these attributes are included in the regression models as categorical variables. There are four age categories, thirteen ticket types, and four discount types, which are listed in Table 6.5. As mentioned previously, these passenger categories are modelled as fixed effects to accommodate potential correlation between the multiple covariates representing passenger characteristics in the models.

A summary of the descriptive statistics for the passenger demand covariates is given in Table 6.6.

Table 6.5: Definition of passenger categories

Passenger age Category	% of trips	Card type Category	% of trips	Discount type Category	% of trips
Adult	98.78	Oyster Card	82.31	No discount	98.78
16-17	0.71	Oyster Photocard	9.45	NR Railcard	1.06
Child	0.52	Elderly Freedom Passes	4.71	Privilege	0.15
Non-Conc Child	2.79E-4	Staff Pass	1.04	Under 11	0.02
		Disabled Freedom Passes	0.91		
		Visitor Oyster Card	0.49		
		Staff Nominee Pass	0.32		
		Barclaycard OnePulse	0.29		
		Bus Nominee Pass	0.22		
		Bus Operator Pass	0.20		
		Retired Pass	0.08		
		Not specified	5.07E-5		
		Barclays Debit	1.27E-5		

Table 6.6: Summary statistics of passenger demand covariates

Covariate	Units	Min.	Max.	Mean	s
Passenger volumes					
Origin platform loading	Numeric	0.01	72.00	1.78	1.53
Dest. platform loading	Numeric	0.01	72.00	1.69	1.09
Line loading	Numeric	0.06	4.16	1.40	0.77
Passenger characteristics					
Passenger age	Categorical	4 categories			
Card type	Categorical	15 categories			
Discount type	Categorical	6 categories			
Passenger travel frequency	Numeric	1.00	793.00	9.47	10.41

Note: s refers to the sample standard deviation.

6.3.6 Model evaluation

To assess the performance of the models, conventional goodness-of-fit and significance testing methods are used. To test the significance of each covariate, the t-statistics and p-values are examined. A high t-statistic and low p-value indicate that the covariate is influential in capturing variation in the data (Dougherty, 2016).

In terms of the overall goodness-of-fit of the models, the semiparametric regression package "mgcv" in the R statistical analysis software outputs a number of indicators, including: the adjusted coefficient of determination $R^2_{adj.}$, the deviance indicator $D_{explained}$, the restricted maximum likelihood (REML) score, and the Akaike Information Criterion (AIC). The models are fitted using restricted maximum likelihood methods with penalised iteratively reweighted least squares. The deviance, REML score, and AIC are based on model likelihood. The adjusted coefficient of determination is more suited to models fitted through the ordinary least squares method, however, it is also used here to assess the goodness-of-fit of the semiparametric models alongside the likelihood-based indicators.

The adjusted coefficient of determination measures the degree of variance that is explained by the model. The indicator includes a penalty for the number of predictors in the model; a higher value of the adjusted coefficient of determination indicates that a greater degree of variance in the data is captured (Dougherty, 2016). The adjusted coefficient of determination $R^2_{adj.}$ is calculated as per equation 6.9 (Dougherty, 2016).

$$R^2_{adj.} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{6.9}$$

where

p is number of predictors or estimated parameters in the model,

n is the number of observations, and

R^2 is the coefficient of determination calculated as per equation 6.10 (Dougherty, 2016).

$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{y_i})^2}{\sum_i (y_i - \bar{y})^2} \tag{6.10}$$

where

y_i is the observed value i of the dependent variable y,

$\widehat{y_i}$ is the predicted value of the dependent variable as estimated by the model under consideration, and

$\bar{y}$ is the mean value of the dependent variable.

The deviance indicator is similar to the adjusted coefficient of determination in concept; it assesses the degree to which the model explains variation in the data. The indicator is based on measuring the maximised likelihood of the model under consideration, relative to the maximised likelihood of the model with no explanatory covariates (Wood, 2006). A higher value of the proportion of deviance explained implies a that a higher degree of variance is captured by the model. The formula for the proportion of deviance explained $D_{explained}$ is given in equation 6.11, and the formula for model deviance D is given in equation 6.12 (Wood, 2006).

$$D_{explained} = \frac{D_{null} - D_{residual}}{D_{null}} \tag{6.11}$$

where

$D_{explained}$ is the proportion of deviance explained by the model under consideration,

D_{null} is the deviance of the model with only the constant term present, and

$D_{residual}$ is the deviance of the model under consideration.

$$D = 2[L(\widehat{\beta}_{max}) - L(\widehat{\beta})]\phi \tag{6.12}$$

where

$L(\widehat{\beta}_{max})$ is the maximised likelihood score of the saturated model. This is the highest possible likelihood score for the data, and it is derived from the model that assigns a parameter for each data point.

$L(\widehat{\beta})$ is the maximised likelihood of the model under consideration, and

ϕ is the scale parameter of the model.

The restricted maximum likelihood (REML) score of the model is computed as part of the iterative optimisation process to obtain the best estimates of the model parameters. The score can also be used as an indicator of model fit, with lower scores corresponding to a better fitted model (Wood et al., 2015; Wood, 2006). The REML scores are calculated as per equation 6.13 (Wood et al., 2015).

$$\nu_r(\lambda) = \frac{||f_k - \mathbf{R}\widehat{\beta_\lambda}||^2 + ||r||^2 + \widehat{\beta}_\lambda^T \mathbf{S}_\lambda \widehat{\beta}_\lambda}{2\phi} + \frac{n - M_p}{2} log(2\pi\phi) + \frac{log|\mathbf{R}^T\mathbf{R} + \mathbf{S}_\lambda| - log|\mathbf{S}_\lambda|_+}{2} \tag{6.13}$$

where

ν_r is the restricted maximum likelihood (REML) of the model,

λ is the smoothness parameter,

n is the number of rows in the model matrix which represents the number of observations of the dependent variable, and p is the number of columns in the model matrix (or the number of model covariates),

ϕ is the scale parameter,

f_k are the smooth functions based on penalised thin plate regression splines,

$\mathbf{R}$ is the upper triangular $p \times p$ factor when the total model matrix $\mathbf{X}$ is decomposed into $\mathbf{X} = \mathbf{QR}$ to enable faster computation, where $\mathbf{Q}$ is the column orthogonal $n \times p$ factor,

$||r||^2$ is evaluated as $||r||^2 = ||y||^2 - ||f_k||^2$ where $f_k = \mathbf{Q}^T y$,

$\widehat{\beta_\lambda}$ is the matrix of unknown model parameters,

$\mathbf{S}_\lambda$ is the matrix of known coefficients and $|\mathbf{S}_\lambda|_+$ is the product of the strictly positive eigenvalues of $\mathbf{S}_\lambda$, and

M_p is the formal degree of rank deficiency of $\mathbf{S}_\lambda$.

The Akaike Information Criterion (AIC) is another indicator based on the likelihood score of the model. The AIC includes a penalty for the number of model parameters, and smaller values of the AIC indicate a better fitted model (Hastie et al., 2009). The formula for the AIC is given in equation 6.14 (Hastie et al., 2009).

$$AIC = -2\log(L) + 2p \tag{6.14}$$

where

p is the number of estimated parameters in the model, and

$\log(L)$ is the REML of the model as defined in equation 6.13.

6.4 Results

Four models are generated in the analysis: three for the components of journey time including access time, on-train time, and egress time, and one for the total journey time. The proceeding sub-sections summarise the modelling process beginning with a discussion of the data cleaning process, followed by a summary of the process to establish the final model form, a discussion of covariate significance, model residuals, and finally, the results of the covariate elasticity estimates are presented.

6.4.1 Data preparation

Initially, the raw data set including outliers was used and all variables that are measured on a continuous scale were modelled as non-parametric smooth terms so that the form of the relationship between the independent and dependent variables could be visually inspected. Through inspection of the non-parametric curves, outlying values of covariates were identified as indicated by divergence in the confidence intervals at the extremities of the covariate ranges. As a result of this process, outlying values at the minimum and maximum ranges for the following covariates have been discarded: covariates capturing headway variance, platform passenger volumes, line loading, and passenger trip frequency. Headways were previously filtered in the train assignment process to exclude very small and very large erroneous records. A summary of the resulting range of the covariates and dependent variables is given in Table 6.7.

As there are numerous covariates in the models, correlation between covariates is assessed for the presence of linear relationships via Pearson correlation testing and for the presence of non-parametric monotonic associations through Spearman correlation testing. For both tests, a value of 1 indicates a positive correlation, a value of -1 indicates a negative correlation, and a value of 0 indicates no correlation (Wonnacott and Wonnacott, 1990). Correlation matrices are generated for the filtered data set. The results show that the highest degree of correlation occurs between the covariates representing headways and normalised headways. The Spearman correlation coefficient has a maximum value of 0.76 and the Pearson correlation coefficient is 0.71. As the maximum values of the correlation coefficients do not indicate strong correlations between the covariates, it is appropriate to initially include all covariates in the models and conduct further model refinement based on covariate significance values (discussed further in section 6.4.3).

Table 6.7: Discarded model outliers

		Raw		Filtered				
Covariate	Unit	Min.	Max.	Min.	Max.	Mean	SD	COV
Time period	15 min intervals	06:00	00:00	06:00	00:00	07:40	4.52	30%
Headway	mins	0.00	352.40	1.00	19.98	2.78	1.47	53%
COV headway	mins	0.00	3.22	0.05	1.10	0.29	0.16	56%
Norm. headway	mins	0.00	21.31	0.05	6.86	1.00	0.32	32%
Origin platform loading	15 min value/daily avg.	0.01	72.00	0.30	3.00	1.40	0.69	49%
Dest. platform loading	15 min value/daily avg.	0.01	72.00	0.30	3.00	1.40	0.69	49%
Line loading	15 min value/daily avg.	0.06	4.16	0.30	3.00	1.19	0.61	51%
Pass trips	mins	1.00	793.00	1.00	35.00	9.39	10.10	108%
Access time	mins	0.00	85.87	0.10	81.47	3.52	2.07	59%
Egress time	mins	0.00	88.25	0.10	45.37	1.40	0.79	56%
On-train time	mins	0.02	114.17	0.10	104.98	8.92	5.33	60%
Total journey time	mins	2.00	121.00	2.00	109.00	13.77	5.69	41%

6.4.2 Model form

The models accommodate covariates in three forms: as parametric terms, as non-parametric smooth terms, and as group-specific fixed effects. As mentioned previously, all continuous variables were initially modelled non-parametrically to inspect the relationship with the dependent variable. After trialling a number of transformations, a log-log form was found to provide the best fit to the data. It should be noted that the log-normal transformation applied to the response variables of journey times corresponds to findings in the literature on transit journey time distributional form. Under regular operating conditions, transit journey times are reported to follow right-skewed distributional forms (refer to literature review in Chapter 2, and analysis on distributional form in Chapter 4).

Through inspection of the non-parametric functions of each covariate, it is found that for some covariates, the functions approximate known parametric forms. Therefore, for analytical tractability and to improve interpretation of the model results, parametric forms are assigned where possible. As mentioned in the presentation of the regression modelling framework, the index variables that capture group-specific effects are modelled with fixed effects. An alternative form for the group-specific effects is to assign random effects. The main distinction between the two forms is that the fixed effects allow the group-effects to be correlated with other covariates in the model, while the random effects do not allow for correlation between the group-effects and other covariates (Greene, 2018; Searle et al., 1992). As mentioned previously, a fixed effects structure is more appropriate if the variable has been drawn from a finite population, while a random effects structure is more appropriate in cases where the variable has been drawn from a large or infinite population, and the observations used in the model are considered a random sample of the population (Greene, 2018; Searle et al., 1992). Although theoretically a fixed effects structure is more appropriate for the journey time models, a random effects structure is also trialled for verification. For simple linear panel models, the Hausman test is conventionally undertaken to determine the suitability of the fixed vs random effects structures (Greene, 2018), however, no such testing method is available for the semiparametric models developed here. As such, evaluation of goodness-of-fit criteria and inspection of model residuals is undertaken to verify the most appropriate group-effects structure.

For demonstrative purposes, the model goodness-of-fit statistics for four model forms are presented in Table 6.8: 1) Models with all continuous covariates modelled with non-parametric smooths and fixed group effects, 2) Models with all continuous covariates modelled with non-parametric smooths and random group effects, 3) Models with all continuous covariates modelled with linear relationships and fixed effects, and 4) Final model form where the most parsimonious parametric form has been applied to the continuous variables while maintaining model goodness-of-fit, and fixed group effects. Four goodness-of-fit criteria as recommended by Wood (2006) are used to assess the model performance: the adjusted coefficient of determination ($R^2_{adj.}$), the percentage of explained model deviation ($D_{explained}$), the Akaike Information Criterion (AIC), and the restricted maximum likelihood (REML) score. The residuals of the models are also inspected for further diagnostics.

As shown in Table 6.8, the results for the fixed and random effects structures (model forms 1 and 2) show that the performance of the two models is almost identical, with the exception of the REML scores in the on-train time model. In the on-train time model, the REML score for the fixed effects model (model form 1) is lower, indicating that the fixed effects model performs better. For the remaining models however, the difference between the two model structures is not discernible, and so the theoretical basis for applying fixed effects for all models is adopted in the analysis.

In terms of the form of the continuous model covariates, the simplest linear models perform worst (model form 3), while the fully non-parametric models (model form 1) perform best across the criteria that represent the degree to which the models capture variation in the data, namely the $R^2_{adj.}$, $D_{explained}$, and residuals dispersion. The AIC and REML scores additionally penalise for model complexity, and so the linear model forms perform best across all models relative to the REML scores and best for the access time and total journey time models according to the AIC. As mentioned previously, a parametric form is more advantageous in terms of model tractability and interpretation, and so the final model forms (model form 4) have been determined to strike a balance between maintaining model parsimony while maximising goodness-of-fit performance.

The models for on-train time and total journey time perform best in terms of the $R^2_{adj.}$ value. The on-train time model reports an $R^2_{adj.}$ value of 0.99, which arises as a result of the inclusion of fixed effects to capture the effect of each of the 444 individual OD routes. The majority of the residual variance in on-train times is captured by these fixed OD effects; indeed, if the fixed

Table 6.8: Model goodness-of-fit

Indicator	Model form	Access time	Egress time	On-train time	Total journey time
n		4.85E6	4.85E6	4.85E6	4.85E6
$R^2_{adj.}$	1	0.192	0.255	0.99	0.878
	2	0.192	0.255	0.99	0.878
	3	0.114	0.249	0.99	0.852
	4	0.191	0.255	0.99	0.878
$D_{explained}$	1	0.192	0.255	0.99	0.878
	2	0.192	0.255	0.99	0.878
	3	0.106	0.246	0.99	0.809
	4	0.191	0.255	0.99	0.878
AIC	1	7.781E6	8.450E6	-1.648E8	-5.036E6
	2	7.781E6	8.450E6	-1.648E8	-5.036E6
	3	-8.407E7	8.511E6	-1.648E8	-4.046E7
	4	7.788E6	8.452E6	-1.648E8	-5.032E6
REML	1	3.891E6	4.226E6	-7.425E7	-2.516E6
	2	3.891E6	4.226E6	-7.381E7	-2.516E6
	3	3.820E6	4.218E6	-7.506E7	-2.842E6
	4	3.895E6	4.227E6	-7.425E7	-2.514E6
Residuals - Range	1	-4.06 – 3.42	-3.33 – 3.63	-4.5E-7 – 4.9E-7	-0.95 – 2.14
	2	-4.06 – 3.42	-3.33 – 3.63	-4.5E-7 – 4.9E-7	-0.95 – 2.14
	3	-4.00 – 3.46	-3.35 – 3.62	-4.6E-7 – 4.9E-7	-0.87 – 2.88
	4	-4.09 – 3.41	-3.33 – 3.62	-4.5E-7 – 4.9E-7	-0.95 – 2.14
Residuals - Standard deviation	1	0.528	0.566	5.070E-08	0.144
	2	0.528	0.566	5.070E-08	0.144
	3	0.553	0.568	5.070E-08	0.159
	4	0.529	0.566	5.070E-08	0.144

Note: n is the number of observations, $R^2_{adj.}$ is the adjusted coefficient of determination, $D_{explained}$ is the proportion of deviance explained, AIC is the Akaike Information Criterion, and REML is the restricted maximum likelihood score.
Model forms as follows: 1) Non-parametric smooths and fixed effects, 2) Non-parametric smooths and random effects, 3) Linear and fixed effects, 4) Parametric and fixed effects.

effects for routes are removed, the $R^2_{adj.}$ of the model drops to 0.86.

In the access and egress time models, individual OD variables are not quantities of interest, as the focus in these models are the station characteristics rather than route characteristics. Fixed effects are included to capture the station level effects in the access and egress models, and as there are 35 stations, the explanatory power of the access and egress models is lower in comparison to the on-train time and total journey time models. There may also be missing covariates in these models that have not been able to be included due to a lack of data availability, and this may be another reason for the lower degree of explanatory power in the access and egress time models.

6.4.3 Covariate significance

The final forms and significance of the model covariates are given in Table 6.9. As shown in the table, all statistically significant terms are significant at a level of $\alpha = 0.001$ (99.9%).

Table 6.9: Form and significance of model covariates

	Access time		Egress time		On-train time		Total journey time	
Covariate	Form	Sig.	Form	Sig.	Form	Sig.	Form	Sig.
Service capacity covariates - dynamic								
Headway	poly deg 5	***	poly deg 3	***	splines		poly deg 5	***
COV headway	poly deg 3	***	poly deg 3	***	splines		poly deg 3	***
Headway/mean headway	poly deg 3	***	poly deg 3	***	splines		poly deg 3	***
Speed					linear	***	splines	***
No. gates (U)	splines	***	splines	***				
No. escalators (U)	splines	***	splines	***				
No. lifts (U)	splines	***	splines	***				
No. stairs (U)	splines	***	splines	***				
No. platform entry/exits	poly deg 2	***	linear	***				
Service capacity covariates - static								
Platform length (U)	splines	***	splines	***				
Walk distance (U)	splines	***	splines	***				
Stops					linear	***	linear	***
Station O/C	dummy	***	dummy	***				
Station PSD	dummy		dummy	***				
Station terminal	dummy		dummy	***				
Station NR connection	dummy	***	dummy	***				
Station age	linear		linear					
Station zone	linear		linear					
Station no. lines	linear		linear					
Station	fixed	***	fixed	***			fixed	***
OD					fixed	***	fixed	***
Line/direction	fixed	***	fixed	***	fixed	***	fixed	***
Passenger demand covariates - demand volumes								
Origin plat. loading	poly deg 2	***			splines		poly deg 2	***
Destination plat. loading			poly deg 2	***	splines		poly deg 2	***
Line loading	poly deg 3	***	poly deg 2	***	splines		poly deg 2	***
Passenger demand covariates - passenger characteristics								
Pass. trips	poly deg 3	***	poly deg 3	***	splines		poly deg 2	***
Pass. age	fixed	***	fixed	***	fixed		fixed	***
Pass. discount	fixed	***	fixed	***	fixed		fixed	***
Pass. type	fixed		fixed		fixed		fixed	

Significance notation: 0 ‘*** 0.001 ‘** 0.01 ‘* 0.05 ‘. 0.1 ‘ 1
Note: For covariates marked with (U), functions are unstable

For the on-train time model, the covariates capturing headway, headway variance, and all passenger demand related covariates are insignificant at the lowest significance bound of 90%. Train speed and static route characteristics are the primary determinants of variance in train running times between stations. For the access, egress, and total journey time models, headways and the passenger demand covariates are significant.

In both the access and egress models, station age, zone, and the number of lines are not significant, and in the access time model, platform screen doors and the connection to the National Rail system are additionally not significant at the lowest significance bound of 90%. The insignificance of these covariates is likely a correction for potential over-specification in the models arising from the inclusion of numerous covariates capturing station characteristics. Passenger card type is also not significant across all models; the passenger age and discount categories most likely capture the variance across card types. The fixed effects for time are also not significant across all models at the lowest significance bound of 90%. This suggests that the variation of journey times over the day are captured by variation in the dynamic service supply and demand covariates, without any additional latent time-specific effects.

In Table 6.9 a number of station covariates in the access and egress time models are marked with (U), which indicates that although these covariates are significant at a level of 99.9% in the models, the resulting functions are unstable, as indicated by high levels of functional oscillation and high levels of divergence in the confidence intervals. The covariates capturing the number of available gates, escalators, lifts, and stairs produce unstable elasticity estimates in the access and egress time models. Moreover, during the modelling process, it was observed that the direction of the functional relationships of these covariates switched when parametric forms were assigned to the other covariates in the models. Further inspection of the covariate plots showed that the data points were dispersed in a highly irregular manner, and as a result, a stable form (parametric or non-parametric) was not able to be established. It is therefore concluded that these covariates do not influence access and egress times in a systematic way.

For the covariates of station platform length and average walk distance, it was initially hypothesised that longer platform lengths and longer average walking distances through the stations would lead to a monotonic increase in access and egress times. Although both covariates were significant in the models, it was found that the two covariates also produced functional forms with a high degree of oscillation and highly divergent confidence intervals. It was initially thought that the forms of these covariates were reflecting interaction and correlation with the station fixed effects. However, after trialling a removal of the station fixed effects, the functional forms for platform lengths and walk distances remained the same. It was therefore concluded that the covariates did not demonstrate any degree of systematic variation with platform length and walking distance, and that the oscillations in the covariate forms were instead representing the inherent properties of each individual station, which were also represented by the station fixed effects.

The covariate plots of the unstable covariates, namely the number of available gates, escalators, lifts, and stairs, and platform length and average walk distance are included in Appendix F for reference.

6.4.4 Model residuals

The error structure for the models is specified such that the errors ϵ_i are identically and independently distributed with mean 0, i.e. $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The errors do approximate a normal distribution around an approximate mean of 0 for each model. For reference, the summary statistics and kernel density plots of the model residuals are given in Appendix F.

As mentioned previously, the Oyster card timestamps are accurate to one minute only. For each individual timestamp, the magnitude of the inaccuracy is unknown. It can therefore be assumed that the inaccuracy has a uniform random distribution and this is accommodated within the random error term of the regression models. It is worth highlighting that the inaccuracy in the timestamps does not impact the magnitude of the parameter estimates, as the effect of the inaccuracy is contained within the random error component of the models. In the access and egress time models, the random error due to the inaccuracy is in the order of 0-59 seconds, and in the total journey time models the random error is in the range of 0-118 seconds. Since the AVL data are accurate to one second, the random error due to timestamp inaccuracy for the on-train time model is in the order of less than one second. In terms of the proportional magnitude of the errors, the mean access and egress times are 3.52 and 1.40 minutes, respectively. Therefore, the average timestamp errors are 14% and 35% of the mean access and egress times. Mean total journey times are 13.77 minutes, and so the average timestamp error equates to 7% of the mean journey time.

6.4.5 Results for model covariates

Tables 6.10 and 6.11 summarise the results for the service supply and demand covariates under the additional subcategories of dynamic and static factors, respectively. The model results for the continuous variables are reported in terms of covariate elasticities. Elasticity values are frequently reported in econometric applications. An elasticity represents the marginal effect on the dependent variable as a result of a unit increase in the covariate value (Cameron and Trivedi, 2005). In general terms, the elasticity of journey time y with respect to the kth model covariate X_k, is given by e_{yk} as per equation 6.15 (Cameron and Trivedi, 2005):

$$e_{yk} = \frac{\partial y}{\partial X_k}\frac{X_k}{y} \tag{6.15}$$

It is important to note that as elasticities represent the rate of change in the dependent variable relative to a unit change in the covariate value, even small values of elasticity can be influential. For example, for a given covariate with an elasticity of 0.05, a 100 unit increase in the value of the covariate (which could arise in extreme disrupted operating conditions), would result in a five-fold increase in journey times. In the following discussion of the covariate elasticity results, as well as reporting the elasticity values, the magnitude of the change in journey times relative to an arbitrary 10% increase in the covariate values is also reported for demonstrative purposes.

The majority of the covariates in the models have a monotonic relationship with journey times, with relatively stable elasticities of journey time reported over a wide range of covariate values. For the dynamic covariates, the majority have non-linear forms and so the elasticities of these covariates vary with the value of the covariate. As such, minimum, maximum, and mean values of the elasticity of each journey time component with respect to the covariates is reported in Table 6.10. The static covariates are either modelled as dummy variables or as fixed effects, or follow linear relationships with journey time. For these covariates, the constant parameter value or constant elasticity is given in Table 6.11.

Overall, the static factors that capture the inherent physical characteristics of the stations and OD routes have the highest magnitudes of covariate elasticities and parameter values (see Tables 6.10 and 6.11). Across all models, the station fixed effects in the access and egress models have the greatest degree of influence, ranging from -320 to 310 log-mins in the egress time model and -120 to 110 log-mins in the access time model. Dummy variables for additional station characteristics, i.e. whether the station is open/closed, the presence of PSD, connection to National Rail, and whether the station is a terminus, have the same order of magnitude. For scale, the average access and egress times correspond to 1.3 and 0.3 log-mins (3.52 and 1.40 minutes), respectively. Therefore, this result indicates that the initial network planning to determine station locations and features has the greatest degree of influence on journey times, and this should be an important consideration in the development of new networks. For existing systems, since the physical characteristics are constant and not easily able to be altered once established, the fixed station and route effects can be interpreted as being conditioning variables that capture the base performance of the network.

In terms of the dynamic operational variables, train speed and headway are the most influential covariates across all models. Train speed is the covariate with the greatest degree of influence with an average elasticity of -0.54 in the total journey time model. This corresponds to an average reduction of 44 seconds in total journey times for every 10% increase in train speed. Headway is the second most influential covariate, with an average elasticity of 0.05 in the total journey time model, corresponding to an average increase of 4 seconds for every 10% increase in headway. These findings show that an increase in train speed and reduction in operating train headways lead to journey time savings.

In the following sections, a detailed discussion of the results of the models is presented, beginning with the service supply covariates in section 6.4.6, followed by the results for the passenger demand covariates in section 6.4.7. Using the model results on access times, a further analysis of the impact of train headways on marginal passenger wait times is undertaken in section 6.5.1.

Table 6.10: Covariate elasticities - dynamic covariates

	Access time			Egress time			On-train time			Total journey time		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Service capacity covariates												
Headway	2.3E-1	-2.0E0	1.3E0	5.4E-2	-7.2E-1	1.3E-1	-	-	-	5.3E-2	-5.6E-1	3.7E-1
cv headway	3.1E-2	1.5E-3	1.7E-1	7.9E-3	2.6E-3	3.8E-2	-	-	-	1.3E-2	5.2E-4	6.8E-2
Hdway/mean hdway	-6.7E-2	-2.1E0	-3.4E-2	-6.6E-2	-6.5E-1	-1.6E-2	-	-	-	-4.5E-2	-6.2E-1	-3.7E-2
Speed	-	-	-	-	-	-	-1.0E0	-1.0E0	-1.0E0	-5.4E-1	-7.2E-1	1.2E-1
No. entry/exits	1.0E-2	-2.0E-2	7.5E-2	-1.6E-1	-1.6E-1	-1.6E-1	-	-	-	-	-	-
Passenger demand covariates												
Origin plat.	2.9E-2	2.4E-2	3.7E-2	-	-	-	-	-	-	1.1E-2	6.3E-3	1.7E-2
Destination plat.	-	-	-	4.6E-2	1.7E-2	6.5E-2	-	-	-	5.4E-3	-9.8E-4	9.9E-3
Line loading	5.1E-2	-7.9E-3	2.9E-1	4.6E-4	-3.2E-2	2.6E-2	-	-	-	1.9E-2	-1.1E-2	4.9E-2
Pass. trips	-7.3E-2	-1.1E-1	-1.5E-2	-5.4E-2	-6.6E-2	-2.0E-2	-	-	-	-2.2E-2	-3.4E-2	-1.4E-3
Mean times (mins)		3.52			1.40			8.92			13.77	

Note: All values are unit-less elasticities, unless otherwise stated.

Table 6.11: Covariate elasticities and predicted values - static covariates

	Form	Units	Access time	Egress time	On-train time	Total journey time
Service capacity covariates						
Route stops	Linear	elasticity	-	-	1.2E0	6.5E-1
Station O/C	Dummy	log-mins	-9.9E1	-2.8E2	-	-
Station PSD	Dummy	log-mins	-	-4.8E0	-	-
Station terminal	Dummy	log-mins	-	2.5E2	-	-
Station NR connection	Dummy	log-mins	-1.0E2	-3.0E2	-	-
Station and route fixed effects						
Entry station	Fixed	log-mins	-1.2E2 – 1.1E2	-	-	-6.7E-1 –1.7E-1
Exit station	Fixed	log-mins	-	-3.2E2 – 3.1 E2	-	-6.1E-1 – 4.3E-1
OD	Fixed	log-mins	-	-	-8.8E-1 –1.7E0	-1.2E0 – 7.3E-1
Line/direction	Fixed	log-mins	-1.1E1 – 8.7E2	-2.4E1 – 6.0E0	-4.5E-1 – 5.2E-1	-8.1E-2 – 7.0E-1
Model intercept	Constant	log-mins	1.2E2	3.7E1	1.7E0	2.6E0

Note: elasticity is unit-less

6.4.6 Results for service supply covariates

Dynamic service supply covariates

Headway covariates

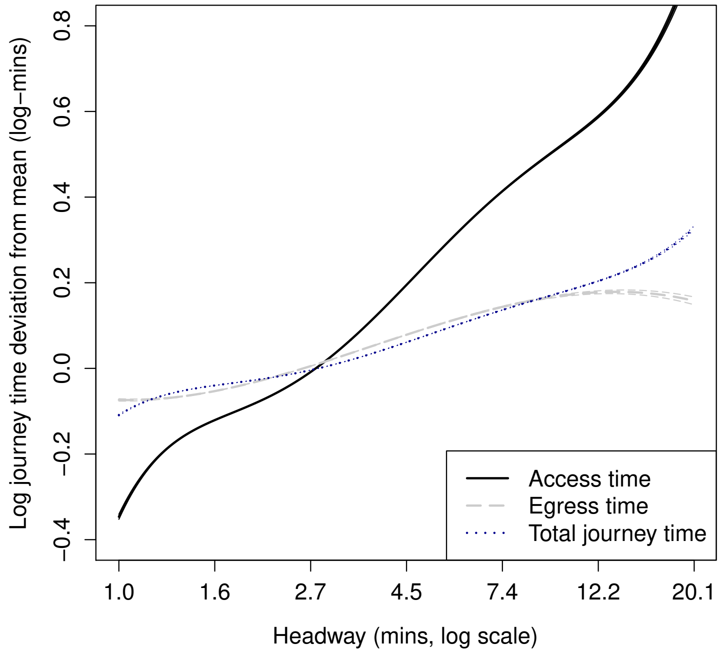

Figure 6.2: Regression plot of headway

The headway between consecutive trains is significant in the access, egress, and total journey time models. The effect of headways is greatest for access times, followed by total journey times, and egress times. Referring to Figure 6.2, access and total journey times tend to increase with longer headways. The average effect of headway is most pronounced for access times where the average elasticity is 0.23; this corresponds to a 4.9 second increase in access time for every 10% increase in headway. Headway is on average the most influential factor out of all dynamic supply and demand covariates in the access time model. Headway is the second most influential dynamic covariate in the total journey time model behind train speed, with an average elasticity of 0.05 (4.4 seconds increase for every 10% increase in headway). As with the access time model, total journey times tend to increase with longer headways between trains, and this can be linked to the effect of longer platform wait times at longer headways. The relationship between access

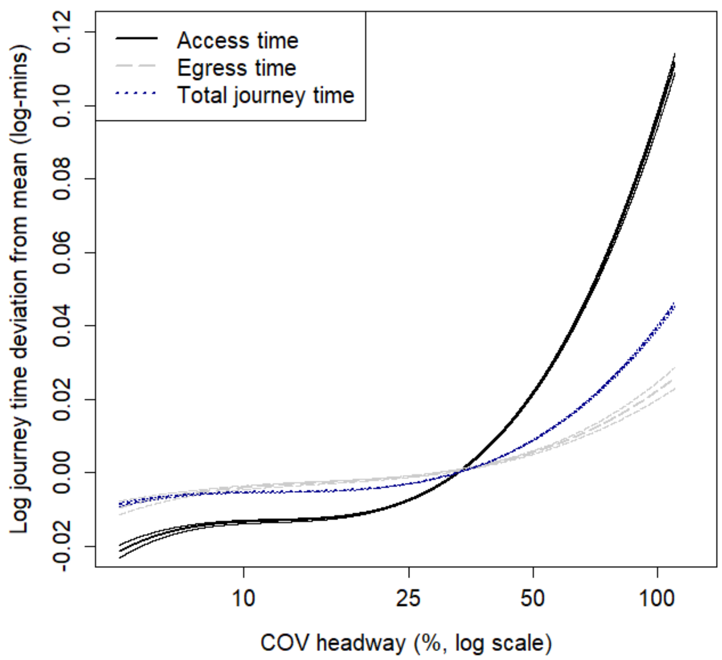

Figure 6.3: Regression plot of *cv* headway

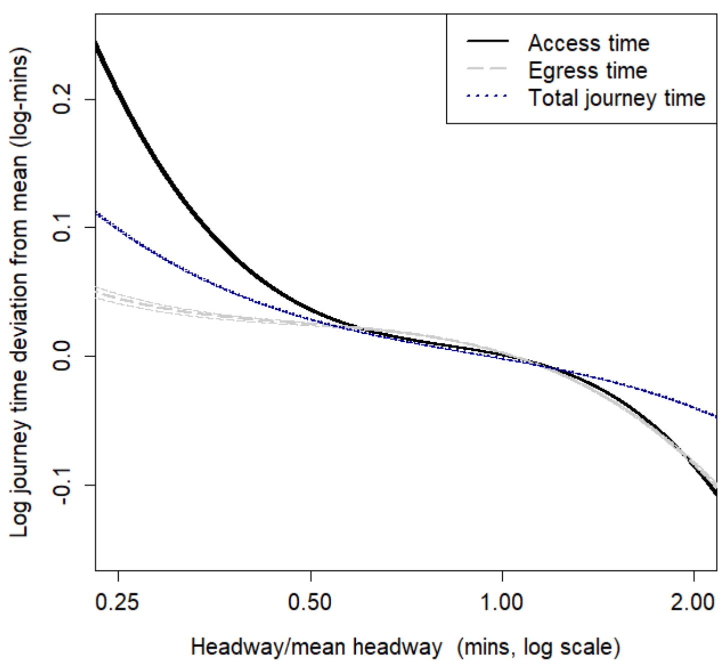

Figure 6.4: Regression plot of headway normalised by mean headway

time and headway is further explored to infer platform wait times, and this is presented in more detail in section 6.5.1.

It should be noted that the headway covariate is assigned one of the most complex parametric forms in the access and total journey time models, with the form specified as a polynomial of 5th degree. This form has been assigned in order to capture the intermediate plateau that occurs at the 1.6 - 2.0 minute headway range. The highest scheduled operating frequency of trains within the study area is 35 trains per hour on the Central line, and this corresponds to a minimum headway of approximately 1.67 minutes between trains. Approximately 15% of all trips in the data set are undertaken when trains operate at headways less than 1.67 minutes, and so the plateaus in the access time and total journey time curves are most likely representing the transition from bunched train operations in highly congested conditions at headways shorter than the published minimum, to regular train operations in line with the published schedules. The plateau is not observed in the egress time model, and so a simpler quadratic form is assigned for that model.

Headway is among the least influential covariates in the egress time model, with an average elasticity of 0.05 which corresponds to an average 0.5 second increase for every 10% increase in headway (see Table 6.10). The relatively minimal effect of headways on egress times demonstrated in the regression model reinforces the assumptions for the passenger assignment algorithm in Chapter 5, where hypothesis testing indicates that egress times are independent of headways.

The coefficient of variance (cv) of headway is included in the model as a measure of the regularity of headways. The cv of headways for 90% of trips in the data set ranges from approximately 10% to 83%. As shown in Figure 6.3, in this range of values, all models show that journey times tend to increase with higher levels of headway irregularity as expected, however the magnitude of the elasticities indicate that the cv of headway does not influence journey times to a high degree compared to other covariates in the models. In terms of average elasticities, the elasticity of access time with respect to cv headway is 0.03, 0.008 for egress time, and 0.01 for total journey times. This corresponds to an average 0.7, 0.07, and 1.1 second increase in access, egress, and total journey times for every 10% increase in cv headway, respectively. It should be noted that in highly inconsistent operating conditions when the cv of headway rises beyond approximately 50%, journey times increase at a faster rate. At the highest observed values of COV headway, average journey times increase by 35, 0.3, and 56 seconds in the access, egress, and total journey time models for every 10% increase in COV headway, respectively.

The covariate measuring the train headway normalised by the 15 minute mean of headways represents the degree of train bunching, which can be used to infer levels of network congestion. Values close to or equivalent to 1 indicate normal operations. Values less than 1 indicate that trains are operating in congested conditions, while values higher than 1 indicate less congested conditions. The data show that for 90% of trips, the value of the normalised headway indicator lies in the range from approximately 0.55 to 1.56. Over this range, Figure 6.4 shows that journey times tend to decrease as the value of the indicator increases. In terms of the interpretation of the indicator, this result suggests that journey times tend to decrease as network congestion decreases. In terms of average elasticities, the elasticity of access time with respect to the normalised headway indicator is -0.07, -0.07 for egress time, and -0.05 for total journey times. This corresponds to an average 1.4, 0.6, and 3.7 second decrease in access, egress, and total journey times for every 10% increase in the value of the normalised headway indicator, respectively.

Dynamic station characteristics

The availability of the fare gates, vertical transport, and platform entry and exit points are influenced by usage. For these covariates, the number of available components are multiplied by a usage factor. As previously discussed in section 6.4.3, although the models show that these covariates are significant, stable parameter estimates are not able to be established for the number of available gates, the number of escalators, the number of lifts, and the number of stair steps. Stable elasticity estimates are able to be derived, however, for the number of entry and exit points along the platforms (see Figure 6.5). In terms of elasticity magnitudes, the number of exit points are more influential in the egress time model compared to the number of entry points in the access time model. Egress times decrease with a linear elasticity of -0.16, or a 1.4 second reduction for every 10% increase in the number of available exit points. Though significant in the model, the number of platform entry points have a very minimal effect on access times. The average elasticity is the lowest of all covariates in the access time model at

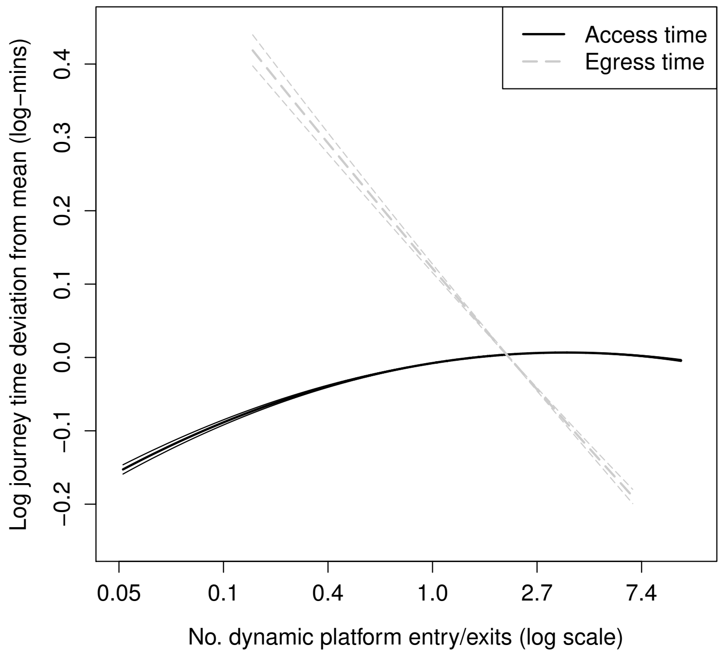

Figure 6.5: Regression plot of number of platform entries/exits interaction term

0.01, or a 0.22 second increase in access times for every 10% increase in the number of available entry points.

Train speed

Train speed is modelled as a covariate in the on-train time and total journey time models, and is significant at a level of 99% in both models. In both models, train speed is the most influential dynamic covariate. As expected, the elasticities are negative, implying that train times and total journey times decrease as train speed increases (see Table 6.3 and Figure 6.6). Train speed has similar levels of influence in both models; the average elasticity of train speed with respect to on-train time is -1.0 and is linear, which corresponds to a reduction of 54 seconds in on-train time for every 10% increase in speed. For total journey times, the average elasticity is -0.5, corresponding to an average 44 second reduction in total journey time for every 10% increase in train speed.

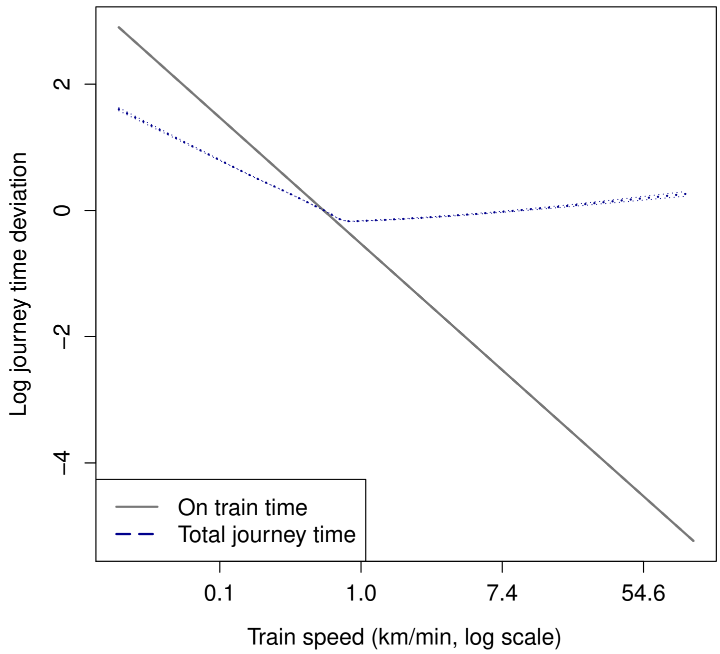

Figure 6.6: Regression plot of train speed

Static service supply covariates

The time-invariant covariates capture the base physical station and route-specific characteristics that may have an effect on journey times. Categorical factors are also included in each model to capture any residual group-specific effects of each station and route which have not been captured by the other covariates. As a consequence of the numerous variables capturing different aspects of the same entities, there is a strong possibility of correlation between the variables and to accommodate for this, the effects are modelled using a fixed effects framework. As such, the results for any given fixed factor should be interpreted relative to the other fixed factors. The results are given in Table 6.11.

In the access and egress time models, the factors that capture static station characteristics are the most influential. The magnitude of the effects of the station dummy variables range in the order from 1E-1 to 1E2 log-mins. In both the access and egress time models, the sign of the elasticities of each significant covariate are the same, implying that the effects of the covariates in terms of resulting in an increase or decrease in journey times are the same.

The covariate that represents whether a station is open-air or underground has a negative elasticity in both models. Open air platforms are typically at or closer to ground level and so there is a shorter distance to travel to reach the exit, compared to the stations with underground platforms. As a result of the shorter walking distance, access and egress times are correspondingly lower. Access and egress times for stations that are connected to the National Rail system also decrease. The layout of stations with these attributes are likely designed to accommodate large volumes of transfer passengers and as a result may be better suited to allow quick passenger movements; this is the most likely cause of the shorter access and egress times at these stations. The presence of platform screen doors (PSD) results in a decrease in egress times. The literature on the effect of PSD on train dwell times is mixed. A recent review of door opening times across 33 metro systems worldwide imply that the presence of PSD leads to longer train dwell times at stations (Barron et al., 2018), while recent laboratory experiments on the London Underground platform-rolling stock interface indicate that PSD do not have a detrimental effect on dwell times (de Ana Rodrguez et al., 2016). Within the analysis, stations on the Jubilee line are the only stations fitted with PSD, and so the result indicating that PSD reduce egress times may also be representing residual characteristics specific to the Jubilee line. Finally, egress times are shown to increase at the three terminus stations in the analysis, namely, Walthamstow Central and Brixton on the Victoria line and Waterloo station on the Waterloo and City line.

In terms of the fixed effects capturing residual station-specific characteristics, the order of magnitude is in line with the previously discussed static station covariates. For the access time model, the station effects range from -120 to 110 log-mins, while for the egress time model, the station effects range from -320 to 310 log-mins. The line/direction effects are relatively smaller with an order of magnitude in the range of 1E-2 to 1E1 log-mins, implying that most of the variance in journey times for the access and egress time models arise from station-specific characteristics rather than line and direction-specific factors. The magnitude of the access and egress model intercepts are in line with the covariates that capture station characteristics.

In the on-train time model, the fixed effects for OD routes are the most influential in the model ranging from -0.9 to 1.7 log-mins, followed by the line/direction effects which range from

-0.45 to 0.52 log-mins. In the total journey time model, the fixed effects capturing OD routes have the greatest influence ranging from -1.2 to 0.73 log-mins, followed by the fixed effects for stations, and line/direction which range from -0.08 to 0.7 log-mins. A penalty is applied for the number of intermediate stops in both the on-train time and total journey time models, with linear elasticities of 1.2 for on-train times and 0.65 for total journey times. This corresponds to a 66 second and 54 second increase in on-train and total journey times for every 10% increase in the number of intermediate stops, respectively.

6.4.7 Results for demand covariates

Passenger volumes

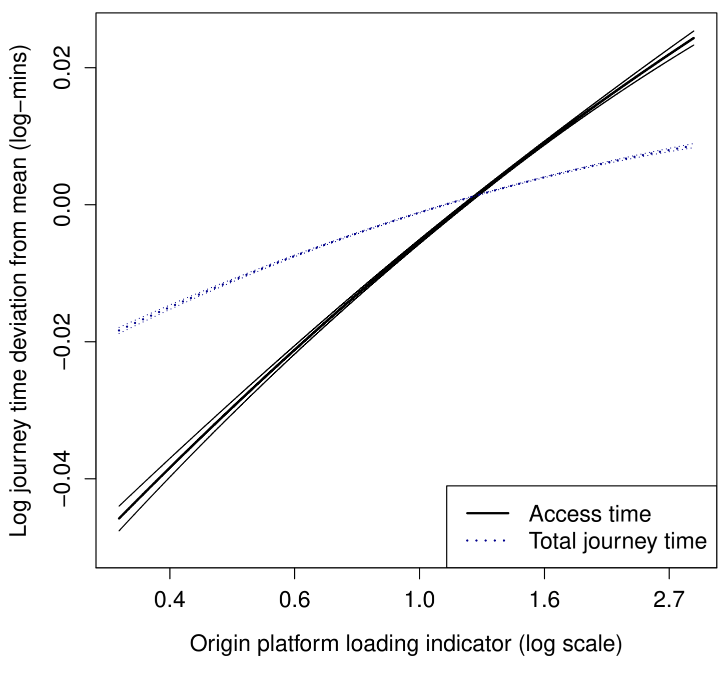

Figure 6.7: Regression plot of origin platform passenger volume indicator

Three covariates are included in the journey time models to represent passenger demand: entry platform loading at the origin station, exit platform loading at the destination station, and on-train line loading. From Table 6.10, the average effect of all passenger demand covariates results in an increase in the access, egress, and total journey time models.

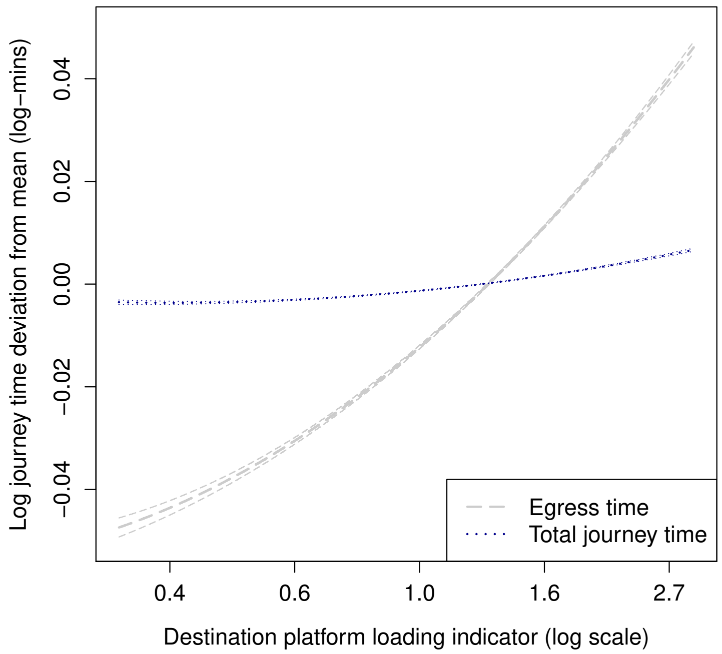

Figure 6.8: Regression plot of destination platform passenger volume indicator

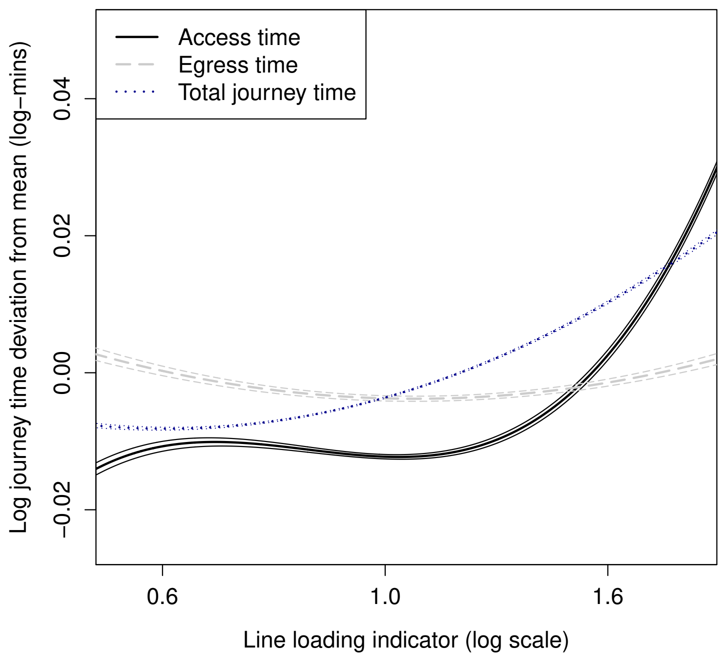

Figure 6.9: Regression plot of passenger line loading indicator

For the access time model, the most influential component of passenger demand is line loading followed by platform loading (see Table 6.10). Both covariates have positive average elasticity values. Referring to Figures 6.7, 6.8, and 6.9, access times generally monotonically increase with line and platform loading, with some minimal localised fluctuation in the line loading plot. The functions of egress times also show increasing trends with platform loading, implying that egress times become longer as passenger congestion at the destination platform increases. Line loading has a minimal influence on egress times, with an average covariate elasticity of 5E-4, the lowest of all significant covariates in the model. For the total journey times model, all passenger demand covariates again have positive average elasticity values. The plots further show that total journey times tend to monotonically increase as passenger demand increases.

If the elasticity effects of platform and line loading are combined, the degree to which each component of journey time is affected by passenger demand as a whole can be compared. The combined average elasticity is 0.08 for access times, 0.05 for egress times, and 0.04 for total journey times. This corresponds to an increase of 1.7, 0.4, and 2.9 seconds in access, egress, and total journey times as a result of a 10% increase in passenger demand, respectively. Although the effect of passenger demand is not as great as other covariates in the models, overall journey times tend to increase as passenger demand increases, and the effect is greatest for total journey times and access times.

Passenger characteristics

Table 6.12: Passenger effects (log-mins)

Category	% Obs.	Access time	Egress time	Total journey time
Passenger age				
Non-concession child	3E-4	4.1E-1	5.7E-2	1.2E-1
Child	0.5	1.2E-1	4.7E-2	3.5E-2
16-17	0.7	3.0E-2	1.9E-2	9.8E-3
Adult	98.8	0.0E0	0.0E0	0.0E0
Passenger discount				
Under 11	1E-2	6.0E-2	5.9E-2	1.5E-2
None	98.6	3.9E-2	5.3E-2	7.3E-3
NR railcard	1.2	3.0E-2	5.1E-2	4.2E-3
Privilege	0.2	0.0E0	0.0E0	0.0E0

Note: all values in log-mins unless otherwise stated.

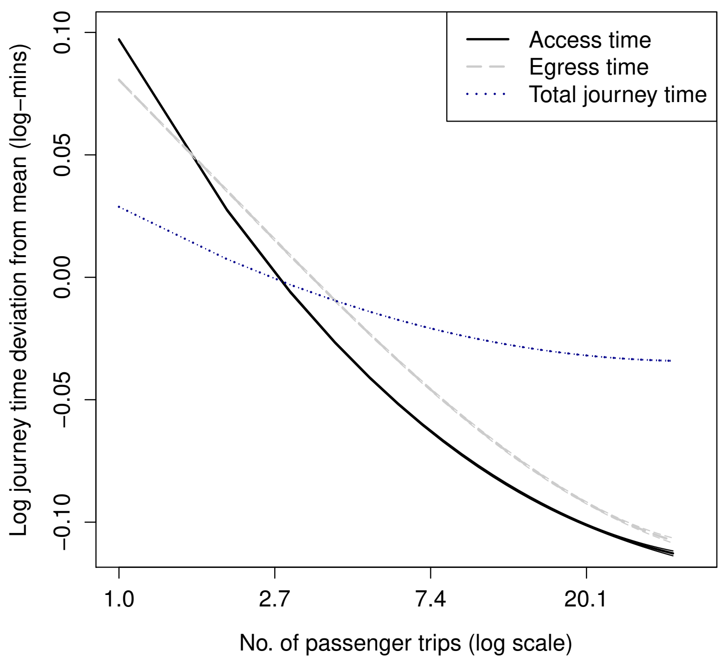

Figure 6.10: Regression plot of number of passenger trips

Covariates are included to capture the effects of passenger-specific characteristics on journey times. The significant variables across all models are passenger age and discount type, which are modelled as categorical factors with fixed effects, and the number of passenger trips which is modelled as a continuous numerical variable. The results for the significant categorical passenger attributes are given in Table 6.12, while the elasticities for the number of passenger trips are given in Table 6.10 and illustrated in Figure 6.10.

In terms of the passenger demographics, the majority of trips in the modelled data sets (97%) are made by passengers holding adult Oyster cards with no additional discounts. Table 6.12 shows that across all models, the passengers with the lowest age effects are those holding adult cards. Conversely, those holding non-concession 11-15 years of age child cards (3E-4% of trips) have the highest effect across all models, while those holding under 16 years of age concession child cards (0.5% of trips) have the second highest effect across all models, indicating that these child groups move slowest through the system. For the discount types, TfL staff holding cards with "privilege" discounts (0.2% of trips) travel fastest through the stations in all models. Children under 11 (0.01% of trips) have the slowest journey times across all models (see Table

6.12).

The results capturing the passenger effects show that in general, adults and staff holding "privilege" cards travel fastest through the stations. It is likely that these groups comprise of regular commuters who are familiar with the system. In further support of this, the covariate capturing the number of trips a passenger completes on the same route shows that access, egress, and total journey times monotonically decrease as trip frequency increases (see Figure 6.10). The results therefore suggest that system familiarity does have an effect on journey times, and that regular commuters are more likely to travel faster through stations as a result. The elasticity of access times with respect to passenger trips is -0.07, -0.05 for egress times, and -0.02 for total journey times. This corresponds to a decrease of 1.5, 0.5, and 1.8 seconds in access, egress, and total journey times for every 10% increase in trip frequency, respectively. The results indicate that the effect of route familiarity is greatest for total journey times and access times.

6.5 Applications

6.5.1 Headways and waiting times

Access time is defined as the time taken from passenger tap-in at the origin station (t_i^{entry}) to the point where the train departs from the origin station platform (DT_{oj}). There are two distinct phases of passenger movements during this time: 1) the time taken to walk from the ticket gates to the platform, t_{oi}^{walk} and 2) the wait time at the platform before boarding, w_{ij}. Based on this, the access time y_{ij}^{ac} for passenger i boarding train j, can therefore be expressed as per equation 6.16. In the access time model, walk times are quantified through fixed effects that capture the station- and line/direction-specific characteristics, and the passenger demand covariates.

$$\begin{aligned} y_{ij}^{ac} &= DT_{oj} - t_i^{entry} \\ &= w_{ij} + t_{oi}^{walk} \end{aligned} \tag{6.16}$$

Using mean predicted values of access times for calculation purposes, the expected value of access times can be expressed as:

$$E[y_{ij}^{ac}] = E[w_{ij}] + E[t_{oi}^{walk}] \tag{6.17}$$

After having taken into account all other variables in the access time model, if the sole influence of headways h_{ij} on access times is analysed, the derivative of access time with respect to headway is:

$$\begin{aligned} \frac{\partial E[y_{ij}^{ac}]}{\partial h_{ij}} &= \frac{\partial E[w_{ij}]}{\partial h_{ij}} + \frac{\partial E[t_{oi}^{walk}]}{\partial h_{ij}} \\ &= \frac{\partial E[w_{ij}]}{\partial h_{ij}} \end{aligned} \tag{6.18}$$

Passenger walk times are independent of the headway between trains, and so in equation 6.18, the rate of change in passenger walk times with respect to headways $\frac{\partial E[t_{oi}^{walk}]}{\partial h_{ij}}$ is omitted. The derivative of access time with respect to headways therefore primarily represents the marginal wait time at the platform $\frac{\partial E[w_{ij}]}{\partial h_{ij}}$.

The relationship between access times and headways is introduced in section 6.4.6. As shown in Figure 6.2, access times tend to increase as headways become longer, though the rate at which access times increases differs at different values of headway. As demonstrated in equation 6.18, the rate of change of the access time-headway curve represents marginal passenger wait times, and so the derivative of access times with respect to headways is computed.

Two steps are involved to compute the derivative of access times with respect to headways. First, predicted values are required for the value of access times with respect to the headway covariate. The predicted values are able to be extracted from the regression model for access times with respect to the headway covariate. Finite differences between each pair of sequential data points are then required to be taken in order to compute the derivative of access times with respect to headways. Due to the large number of observations and high levels of dispersion in the predicted values generated from the access time model, the finite differences between a set of two consecutive individual points contain extreme outlying values which tend to artificially skew the results. In order to smooth out the outliers, the mean of the predicted values is calculated for a fixed value of headway at 1 second intervals. This corresponds to the equivalent level of accuracy in the train movement data timestamps. Repeating this calculation for the range of headways observed allows the generation of a dose-response function, where headway is the dose and the expected value of access time as a function of headway is the response. The calculation of the dose-response function of expected access times with respect to headway is given in equation 6.19.

$$
\begin{aligned}
E[y^{ac}(h)] &= E[[E[y^{ac}_{ij}|h_{ij}, X_{ij}]] \\
&= \frac{1}{n}\sum_{h=1,ij}^{h=20,ij} E[y^{ac}_{ij}|h_{ij}, X_{ij}]
\end{aligned}
\tag{6.19}
$$

where $E[y^{ac}(h)]$ is the expected value of access times with respect to headway h, $E[y^{ac}_{ij}|h_{ij}, X_{ij}]$ is the expected value of the access time regression function estimated previously (refer to section 6.4) conditional on headway and the remaining covariates in the access time model denoted collectively by X_{ij}. The term y^{ac}_{ij} represents individual access times for passenger i assigned to train itinerary j, and so the summation refers to the summation of access times across all n number of passenger and train itinerary combinations ij when the headway associated with the train itinerary is equivalent to h. The bounds for headway are defined from $h = 1$ to $h = 20$ minutes, as this corresponds to the filtered values of the data set used in the regression modelling in section 6.4.

The second step is the estimation of the derivative of access times with respect to headways. The derivative is estimated through a discrete approximation by computing finite differences between the average predicted values calculated in equation 6.19. This process is undertaken for the total data set, and separately for peak and off-peak periods. Four models are estimated, segmented by time period: 1) all time periods, 2) peak of the AM peak, 3) inter-peak, and 4) peak of the PM peak. The equation for calculating the derivative from the average predicted values is given in equation 6.20.

$$\frac{\partial E[y^{ac}(h)]}{\partial h} = \frac{E[y^{ac}(h+a)] - E[y^{ac}(h)]}{(h+a) - h} \tag{6.20}$$

where $E[y^{ac}(h)]$ is the expected value of access time with respect to headway h as calculated per equation 6.19, and a is a small finite increment.

In order to understand the underlying trends in marginal passenger wait times, non-parametric regression splines are fitted to the derivative data points. It was initially found that the non-parametric splines produced a high degree of oscillation in the curves, and so parsimonious parametric forms have been assigned to capture the general trend. Exponential decay parametric forms are found to provide the best fit to all marginal wait time curves. For demonstrative purposes, the goodness of fit statistics for three model forms are summarised in Table 6.13 as follows: 1) Headway modelled with non-parametric smooths, 2) Headway modelled with linear relationship, and 3) Headway modelled with exponential decay form.

Table 6.13: Derivative of access times vs headway model goodness-of-fit statistics

Indicator	Model form	All	AM peak of peak	Inter-peak	PM peak of peak
n		892	458	729	676
$R^2_{adj.}$	1	0.56	0.62	0.57	0.79
	2	0.02	0.06	0.06	0.08
	3	0.45	0.47	0.47	0.68
$D_{explained}$	1	0.56	0.63	0.58	0.80
	2	0.02	0.06	0.06	0.08
	3	0.45	0.47	0.47	0.68
AIC	1	-390.80	10.45	-453.04	-351.09
	2	304.05	422.50	119.24	648.36
	3	-208.19	160.61	-303.92	-68.67
REML	1	-170.07	28.50	-201.63	-147.54
	2	158.13	215.84	65.62	329.23
	3	-101.84	81.43	-149.71	-32.40

Note: n is the number of observations, $R^2_{adj.}$ is the adjusted coefficient of determination, $D_{explained}$ is the proportion of deviance explained, AIC is the Akaike Information Criterion, and REML is the restricted maximum likelihood score
Model forms as follows: 1) Non-parametric smooths, 2) Linear, and 3) Exponential decay.

As shown in the table, model form 1 with the non-parametric smooths performs best and the simplest linear model form 2 performs worst across all criteria for all models. As a parametric form is more advantageous in terms of model tractability and interpretation, the final exponential decay model form (model form 3) has been chosen to achieve a balance between maximising the goodness-of-fit performance and maintaining model parsimony and tractability. The general exponential decay form of the derivative of access times with respect to headway is given in equation 6.21.

$$\frac{\partial E[y^{ac}(h)]}{\partial h} = \alpha + e^{-kh} + \epsilon \tag{6.21}$$

where

h refers to headway

$\frac{\partial E[y^{ac}_{ij}(h)]}{\partial h}$ is the derivative of the expected value of access time with respect to headway as calculated per equation 6.20

k is a parametric coefficient, α is the model constant, and ϵ is the random error term

The plot of the derivative of access times with respect to headway is shown in Figure 6.11. Note that plot is displayed with an upper headway bound of 8 minutes (as opposed to the upper bound of 20 minutes used in the generation of the curves) to highlight the interception point of the curves with the line that denotes wait times as being equivalent to half the headway. Key attributes of the curves are summarised in Table 6.14.

Table 6.14: Marginal platform wait times by time period

Time period		Max.	Headway at 0.5 (mins)	Mean headway (mins)	% trips < 0.5 (%)	Mean < 0.5
All	6am-12am	1.35	2.42	2.82	0.52	0.24
Peak of AM peak	8:30-9:30am	1.58	2.48	2.45	0.31	0.21
Inter-peak	10:30am-4pm	1.25	2.25	3.00	0.73	0.22
Peak of PM peak	5:30-6:30pm	2.00	2.80	2.44	0.23	0.23

Note: h refers to observed headway.

In the literature, the prevailing consensus is that for high frequency services, passengers arrive randomly with a uniform random distribution. Under perfectly random conditions, the average passenger wait time (w) is equivalent to half of the headway between trains, i.e. $w = 0.5h = 0.5\psi^{-1}_{freq.}$, where h is train headway and $\psi_{freq.}$ is the frequency of services. Under such conditions the marginal impact of headway on waiting time should remain 0.5, no matter what the initial

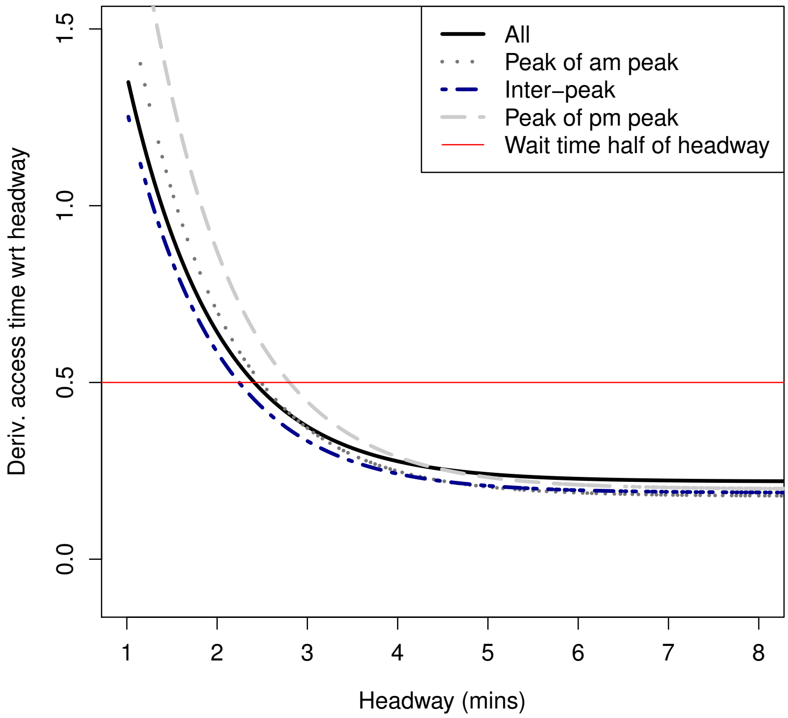

Figure 6.11: Derivative of access time with respect to headway

frequency is. As train frequency decreases, passengers are reported to arrive non-randomly after consulting timetable and/or live arrivals information in order to minimise wait times, and so the average passenger wait time is expected to be less than half of the headway between trains.

As shown in Figure 6.11 and Table 6.14, under high frequency, the impact of a change in headways can be much greater than half of the headway itself. This outcome can be attributed to station congestion and the possibility of denied boarding. By contrast, in low frequency periods, the model predicts that the average passenger saves less time than half of the headway if the operator raises train frequency. This is likely to be caused by the fact that passengers do not arrive to the station randomly; to some extent they consult the timetables instead. This particular finding is therefore in line with the literature for wait times during long headways.

The maximum values of the functions represent the maximum marginal wait time of passengers, and for all functions, this occurs when trains operate at the shortest observed headways. In this region, travel conditions are most likely to be congested with high levels of passenger demand

and/or train bunching. Maximum marginal wait times are higher than a value of 1 across all time periods, which implies that passengers are likely to miss at least one train when trains are operating at minimum headway. The maximum value is lowest during the inter-peak, with a value of 1.25. The maximum value of marginal wait time during the peak of the AM peak is 1.58, and this increases to 2.00 during the peak of the PM peak. The results therefore show that passengers may miss approximately one and a half trains during the morning peak, increasing to two missed trains in the afternoon peak. Across all time periods, the maximum marginal wait time is 1.35, which indicates that considering all trips on average, passenger wait times are slightly longer than one headway during the busiest times of day, when trains operate at the shortest headways.

The transition between random and non-random passenger arrivals, i.e. the threshold level of headway where the derivative falls below 0.5, is in the range of 2-3 minutes, depending on the time of day. The inter-peak function is equivalent to 0.5 at a headway of 2.25 minutes. The mean headway of trips undertaken during the inter-peak is 3.00 minutes, and 73% of trips have marginal wait times less than half of the headway. In the AM and PM peak times, the functions cross 0.5 at longer headways, which reflects a higher degree of congestion during the peaks compared to the inter-peak. In the peak of the AM peak, the function is equivalent to 0.5 at a headway of 2.48 minutes, and approximately 31% of trips have marginal wait times less than half of the headway. In the peak of the PM peak, the function is equivalent to 0.5 at a headway of 2.80 minutes, the longest out of all time periods, indicating that passenger congestion has the highest impact on marginal waiting times during the PM peak. A much lower percentage of trips, approximately 23%, made during this time period have marginal wait times less than half of the headway. Considering all trips together, passengers arrivals tend to become non-random at a headway of 2.42 minutes, and approximately 52% of passengers experience marginal wait times less than half of the headway. Across all time periods, for those passengers that experience marginal wait times less than 0.5, the average marginal wait time is approximately 0.24, or almost one quarter of the operating headway.

6.5.2 Case study: Oxford Circus to Brixton

Table 6.15: Oxford Circus to Brixton mean predicted values and elasticities

	Access time		Egress time		On-train time		Total journey time	
	Value	e	Value	e	Value	e	Value	e
Service capacity covariates - dynamic								
Headway	-7.8E-2	1.8E-1	-3.4E-2	3.5E-2			-2.9E-2	3.5E-2
cv headway	-1.0E-3	2.9E-2	-1.0E-4	6.7E-3			-7.6E-4	1.1E-2
Hdway/mean hdway	-1.9E-3	-6.3E-2	1.6E-4	-6.5E-2			-8.4E-4	-4.3E-2
Speed					8.8E-2	-1.0E0	5.3E-2	-5.4E-1
No. entry/exits	6.1E-3	4.8E-3	-4.1E-2	-1.6E-1				
Service capacity covariates - static								
Stops					6.4E-1	1.2E0	3.3E-1	6.5E-1
Station O/C	0.0E0	-9.9E1	0.0E0	-2.8E2				
Station PSD	0.0E0	-2.1E-1	0.0E0	-4.8E0				
Station terminal			2.5E2	2.5E2				
Station NR connection	0.0E0	-1.0E2	-3.0E2	-3.0E2				
Passenger demand covariates								
Entry plat. loading	2.0E-2	2.5E-2					7.2E-3	7.3E-3
Exit plat. loading			3.8E-2	6.2E-2			5.0E-3	9.0E-3
Line loading	1.2E-2	1.0E-1	3.2E-3	1.8E-2			1.6E-2	4.3E-2
Pass. trips	-3.9E-2	-5.6E-2	-3.4E-2	-4.8E-2			-1.3E-2	-1.6E-2
Fixed effects								
Entry station	-1.2E2						3.4E-1	
Exit station			0.0E0				2.9E-1	
OD					-3.6E-1		-1.6E-1	
Line/direction	-2.4E-1		6.0E0		5.2E-1		-8.1E-2	
Pass age	7.3E-4		3.2E-4				2.5E-4	
Pass discount	3.9E-2		5.3E-2				7.3E-3	
Intercept	1.2E2		3.7E1		1.7E0		2.6E0	
Residuals	-6.3E-2		1.4E-1		-4.1E-9		2.0E-2	
Mean journey times	1.2E0		1.0E-1		2.6E0		2.9E0	
(Mins)	(3.49)		(1.11)		(13.87)		(18.36)	

Note: All values are in log-mins and elasticities (e) are unit-less, unless otherwise stated.

A representative OD pair is chosen to further illustrate the application of the results from the regression models. The OD route from Oxford Circus station to Brixton station on the Victoria line southbound is selected during one of the busiest times on the route, the PM peak period from 5.30pm to 6pm. A total of 6,105 trips are undertaken during this period, and the average headway associated with the trips is approximately 2.0 minutes. Table 6.15 gives mean predicted covariate values in log-mins and average elasticities with respect to each covariate for each journey time component. The mean predicted values of the covariates are included to show the degree to which each covariate contributes to each journey time model, and the elasticity is given to show the rate of change in journey time with respect to a unit increase in each covariate value. A second table, Table 6.16, explicitly shows the station and OD-route specific effects to separately demonstrate the total effect of the static physical characteristics of the network.

As shown in Table 6.15, the static station and OD route characteristics are the most influential across all models. The results for this OD pair therefore align with the aggregate model results presented previously. As shown in the table, the predicted values and the fixed effects of the static station and OD route characteristics are high in magnitude compared to the remainder of the covariates in the models. Since these characteristics represent immutable physical aspects of the system and are specific to the station and OD route of interest, the static covariates and fixed effects can be summed to understand the aggregate effect of the station and OD route effects. Further combining this aggregate effect with the model intercept, a station or OD route-specific intercept for each model can be generated. This is illustrated in Table 6.16. As shown in the table, the aggregation of all the static effects produces station and route-specific intercepts with magnitudes in line with the order of the mean values of journey times for each journey time component. The same process can applied to generate station and route-specific model intercepts for all of the stations and OD routes in the data set to enable comparisons.

Table 6.16: Oxford Circus to Brixton station and OD route-specific intercepts

	Access time	Egress time	On-train time	Total journey time
Static service capacity covariates				
Stops			0.64	0.33
Station O/C	0.00	0.00		
Station PSD	0.00	0.00		
Station terminal		253.57		
Station NR connection	0.00	-297.24		
Fixed effects				
Entry station	-115.14			0.34
Exit station		0.00		0.29
OD			-0.36	-0.16
Line/direction	-0.24	5.99	0.52	-0.08
Total station/OD route-specific effect	-115.38	-37.68	0.80	0.72
Model intercept	116.82	37.22	1.74	2.56
Station/OD route-specific intercept	1.44	-0.45	2.54	3.28
(Mins)	(4.22)	(0.64)	(12.64)	(26.60)

Note: All values are in log-mins, unless otherwise stated.

The static station and route characteristics aside, rankings of the remainder of the dynamic characteristics and passenger fixed effects from most influential to least can be produced. Beginning with the covariate values in the total journey times model, train speed is the most influential of the remaining covariates, followed by the covariate that captures the effect of headways. The passenger demand covariates and the covariates that capture passenger characteristics have a relatively lower amount of influence on total journey times. Using the ranking of covariates for total journey times as a guide, the influence of the covariates on the subcomponents of access, egress, and on-train times can be interpreted.

In terms of the on-train time model, train speed is the only remaining significant covariate in the model aside from the fixed OD route and line/direction effects. The elasticity value is -1.0, which corresponds to a 83 second reduction in on-train times as a result of a 10% increase in train speed.

For the access time model, behind the base physical station characteristics for Oxford Circus, the second most influential factor is the headway covariate. An increase in headway is associated with an increase in access times at a rate of 0.18 (4 seconds for every 10% increase in headway). For egress times, headway is one of the least influential covariates with an elasticity of 0.04 (0.2 seconds for every 10% increase in headway). Overall, the effect of headway is most prominent in the access time component and this trend is consistent with the findings presented in section 6.4.6. In terms of the *cv* of headway, access times increase at an elasticity of 0.03 (0.6 seconds for every 10% increase in *cv* headway) and egress times increase at an elasticity of 0.007 (0.04 seconds for every 10% increase in *cv* headway). For the normalised headway indicator, access times decrease at an elasticity of -0.06 (-1 second for every 10% increase in indicator value) and egress time decrease at an elasticity of -0.07 (-0.4 seconds for every 10% increase in indicator value). As shown in the results, the overall effect of the variance of headways is relatively small in magnitude.

The passenger demand covariates are the next most influential set of covariates after the station and route characteristics. For access times at Oxford Circus station, the combined effect of platform and line loading results in an elasticity of 0.13 (3 second increase for every 10% increase in passenger loading). For egress times at Brixton, the combined effect of the passenger demand covariates is a 0.08 elasticity (0.5 second increase for every 10% increase in passenger loading). The result illustrates that overall, higher passenger demand in the PM peak corresponds to longer journey times. The effect of passenger demand is greater for access times compared to egress times.

The passenger age and discount fixed effects are the next most influential factors, and these effects have an order of magnitude of 1E-2 to 1E-4 log-mins across all models. In terms of the passenger demographics on the route, 98% of passengers are adults who hold regular Oyster cards with no additional discounts. In terms of the number of trips an individual passenger makes, as the number of passenger trips increases, journey times decrease as a likely consequence of route familiarity. This effect is more influential for access times at a -0.06 elasticity compared to egress times at a -0.05 elasticity, which correspond to reductions of 1 second and 0.3 seconds, respectively, as a result of a 10% increase in trip frequency.

Finally, the random sources of variance that cannot be measured by the model covariates are captured by the model residuals. For access times the average value of the model residuals is -0.06 log-mins, indicating that the average access times are slightly overestimated. Egress times are underestimated, with an average residual value of 0.14 log-mins. The variance of on-train times is mostly captured by the modelled covariates, with a model residual of -4E-9 log-mins.

The case study illustrates how the results from the regression analysis can be interpreted. The models enable an understanding of which covariates affect metro journey times at subcomponent level, and the elasticity values for each covariate quantify to which degree journey times are affected. As demonstrated, these results can be used to provide guidance for operators to determine where interventions should be targeted in order to improve journey time performance and reliability within their systems.

6.6 Conclusions

In this chapter, the extent to which different service supply and passenger demand factors influence journey times on urban metro systems is quantified via semiparametric regression. Non-linear elasticity estimates of variables are generated, allowing a more accurate measure of the underlying influential factors of journey time variance compared to the linear estimates used in previous research. The analysis is undertaken on a subset of data representing single-line trips, excluding transfer and route choice options, on three lines on the London Underground. It should be noted that the method is readily able to be extended in future to include route choice and transfers on the London Underground, as well as other metros.

From the passenger to train assignment analysis presented in Chapter 5, journey times are decomposed into three constituent parts: access time, on-train time, and egress time. A semi-parametric regression analysis is then performed, with the journey time for each component set as the response variable. Elasticities of access, on-train, egress, and total journey times are derived with respect to two types of explanatory variables: service supply covariates, and passenger demand covariates. Across all models, it is found that the static covariates capturing the physical characteristics of stations and OD routes have the greatest degree of influence on journey times. For existing systems, the physical components of the network are by and large immutable, and these variables can be interpreted as representing the base characteristics of the system. The dynamic operational factors capture the majority of the remaining variance in journey times.

In terms of total journey times, train speed, followed by train headway are the most influential dynamic covariates. The elasticity of journey time is -0.54 with respect to train speed and 0.05 with respect to headway, which corresponds to a 44 second reduction and a 4 second increase in average journey times as a result of a 10% increase in train speeds and headways, respectively. In terms of the individual components of journey times: for access times, the most influential dynamic covariate is headway with an average elasticity of 0.23; for on-train times, speed is the most influential dynamic covariate with an average elasticity of -1.00; and for egress times, the number of available exit points at the destination station platform is the most influential dynamic covariate with an average elasticity of -0.16.

As an extension of the access time regression model, the effect of headways on passenger wait times at the origin station is quantified. The results indicate that passenger arrival patterns transition from random to non-random in the range of 2-3 minute headways during peak times. In off-peak times where train frequencies are on average lower, the majority of passengers tend to arrive in a non-random way in order to minimise their wait times below a value of half of the headway. Lastly, a case study of trips between Oxford Circus and Brixton on the Victoria line is included to further illustrate how the results of the analysis can be interpreted and applied to infer the degree to which different service supply and demand factors affect journey time variance for a typical passenger route.

The results generated in this analysis can be directly applied by operators to identify the common causes of journey time variance in an urban metro system. By extending this method to analyse other metro systems, benchmarking can also be undertaken to compare the journey time performance across different metro networks.

Chapter 7

Individual passenger-level heterogeneity via semiparametric regression

7.1 Introduction

The models of journey time variance in the previous chapter predominately focus on quantifying the effects of service supply and demand characteristics at a station and line level. Fixed effects are included to capture passenger-level characteristics, however, these are based on classifying passengers into categories related to Oyster card usage; namely, age, card type, discount type, and passenger trip frequency. The passenger trip frequency is derived by observing the number of times an individual passenger makes a trip. The individual passengers are represented by unique Oyster card identification numbers. The identification numbers can be used to not only to provide an indication of trip frequency; the regression models can be structured to enable individual passenger-specific travel characteristics to be quantified.

As with the regression models presented in Chapter 6, past work on quantifying passenger-level effects on journey times has been limited to capturing effects that are descriptive of passenger attributes, rather than individual-specific characteristics. For example, in their linear regression model of transit journey times in the Netherlands, Krygsman et al. (2004) include the passenger characteristics of gender, age, and the number of children per household. Empirical work on train dwell times by Wiggenraad (2001) and Heinz (2003) categorise passengers based on their walking behaviours. Wiggenraad (2001) defines different types of passenger movement categories, distinguishing between clustered and non-clustered passengers and their movements. Heinz (2003) uses the underlying theory on the physiological mechanisms of walking to evaluate the impact of door width and step height, and in turn, dwell times. In their analysis of bus wait times in Austin, Texas, Fan and Machemehl (2009) perform a linear regression analysis and include passenger categories of gender and ethnicity in the model.

Quantifying passenger effects at an individual level using semiparametric regression methods has not been previously undertaken, and this is the key contribution of this chapter. Specifically, the aim is to quantify to what extent variance in journey times can be explained by individual passenger level heterogeneity, across the different components of journey time. The effects represent the unobserved individual differences in travel behaviour, for example, personal choices regarding the route taken through the origin and destination stations, individual walking speeds, train boarding/alighting location choices, amongst other individual-specific factors.

The semiparametric regression models presented in this chapter are an extension of the models presented in Chapter 6. The network supply and demand characteristics are retained in the models, but are aggregated to act as conditioning variables. To enable the modelling of individual passenger effects, multiple observations of the same passenger travelling on the same origin-destination (OD) route are required. As such, the analysis focuses on the set of regular commuters who complete 30 or more repeat trips on the same route over the seven week analysis period. The data are arranged in a longitudinal panel data format; the passenger card identification number is set as the index, with multiple observations of each passenger travelling on the same OD route at different times. The effects of individual passenger heterogeneity are then quantified by assigning a random effects structure at the level of the individual card numbers in each journey time component model.

The remainder of the chapter is organised as follows: the study area within the London Underground is defined in section 7.2, and the framework for the semiparametric regression models are presented in section 7.3. The analysis and discussion of the results are given in section 7.4, and the conclusions are summarised in section 7.5.

7.2 Study area and data

The semiparametric regression models in this chapter are an extension of the semiparametric regression analysis presented in Chapter 6. As such, the study area covers the same extents as follows: the entire length of the Victoria line (16 stations), West Acton to Oxford Circus on the Central line (12 station), and Bond Street to North Greenwich on the Jubilee line (10 stations). Trips in both directions on the lines are analysed, and two additional conditions are imposed: first, the trips are single line trips that originate and terminate on the line of interest, and second, line sections are chosen such that there are no other probable routes between OD pairs under normal operating conditions. A map of the geographical analysis boundaries is shown in Figure 3.1, Chapter 3.

Weekday trips (Monday to Friday) over seven weeks from October to December 2013 are analysed. The Transport for London (TfL) Oyster smart card system records total passenger journey times from entry at the origin station to exit at the destination station for each trip. The TfL NetMIS system records the train departure timestamps at each station in the network. Further detailed information on the properties and limitations of these data sets is given in Chapter 3. In Chapter 5, a passenger to train assignment algorithm is presented to allocate individual passenger trips to individual trains. In Chapter 6, the results from the passenger assignment are used to decompose the total journey times for each trip into the constituent components of access, on-train, and egress times. In this chapter, the data set from Chapter 6 including the journey time components is used.

For the analysis of individual characteristics, multiple observations of an individual passenger completing trips on the same origin-destination route are required. Figure 7.1 shows the non-parametric kernel density distribution of the number of trips an individual passenger undertakes on the same OD pair for the entire data set. Table 7.1 also provides a summary of the number of trips per passenger. The data set covers 35 weekdays; the table and plot show that the majority of passengers complete fewer than 10 repeat trips on the same OD pair.

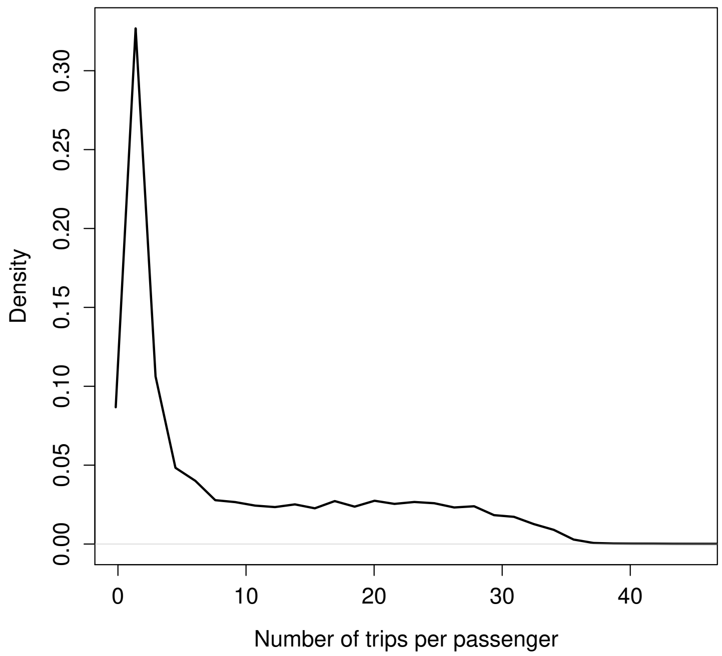

Figure 7.1: Distribution of the number of trips per individual passenger

Table 7.1: Number of trips per passenger

No. trips per passenger	No. trips	% of trips
< 10	4935978	67%
≥ 10 & < 20	1269241	17%
≥ 20 & < 30	1034225	14%
≥ 30	135608	2%
Total	7375052	100%

Selecting an appropriate range of trips is in part dependent on the computational time required to successfully generate the regression models. The time taken for the semiparametric regression models to converge to a solution is in the order of $O(np)$ where n is the number of observations in the data set and p is the number of model parameters to be estimated (Wood et al., 2015). In this analysis, each individual passenger is treated as a unique parameter in the model. Therefore, a greater number of individual passengers in the data set leads to longer computational times. Through a number of trials, the most appropriate trip range for the analysis is found to be passengers who complete ≥ 30 trips on the same route over the 35 day data time frame, which corresponds to a total of 6112 individual passengers. This subset of passengers therefore represent regular commuters who undertake almost one trip per day on the same OD route.

Analysing the properties of the regular commuter data set, the majority (84%) of trips are made by adults holding Oyster cards with no additional discounts applied. The kernel density distribution of trips by time of day is illustrated in Figure 7.2. The plot shows that the regular commuters are more likely to travel during the morning and afternoon peak periods, with more passengers completing trips in the morning peak. Regression models are developed for the components of journey time for each trip; the kernel density distributions of the journey times by component are illustrated in Figure 7.3. Egress times are, on average, the shortest component, followed by access times, and on-train times. Summary statistics of the journey time components are further presented in section 7.3.

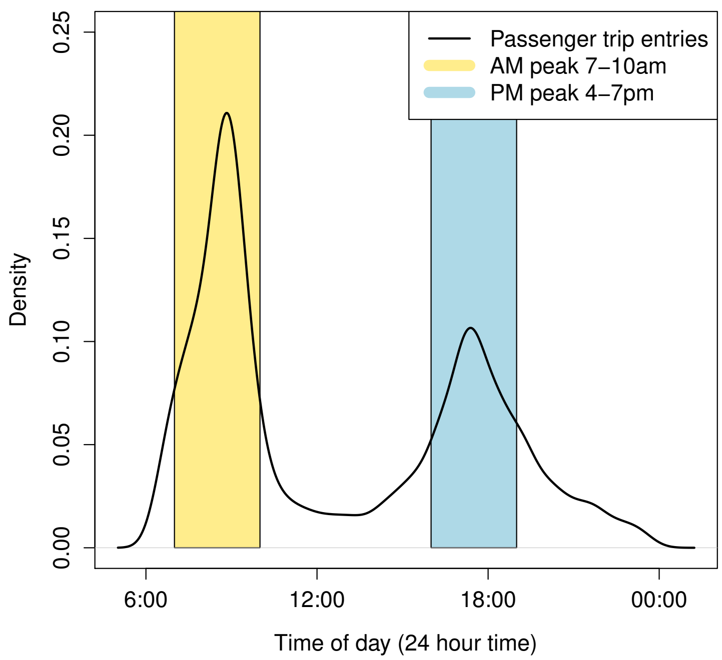

Figure 7.2: Distribution of regular commuter trips by time of day

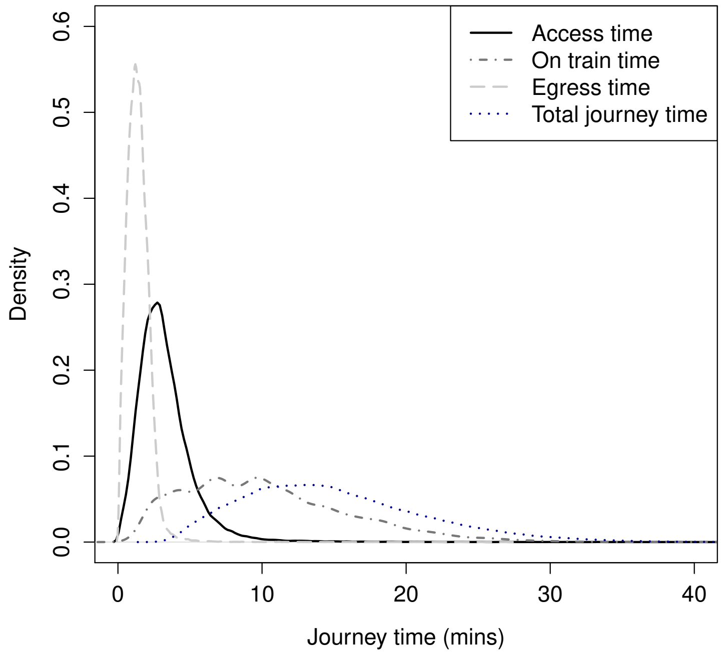

Figure 7.3: Distribution of journey times by journey time component for regular commuters

7.3 Methods

7.3.1 General regression model framework

The primary aim of the analysis is to quantify the effect of individual passenger heterogeneity on journey times. Along with the passenger effects, covariates that capture the service supply and passenger demand levels are included. These covariates are considered as conditional variables to represent the base operating environment, and are modelled at a relatively more aggregate level, compared to the more detailed models that capture station and route-specific factors in Chapter 6. As with the analysis in Chapter 6, a semiparametric regression framework is adopted to exploit the large volume of data. The general form of the regression model is shown in equation 7.1. To avoid repetition, detailed discussions of the semiparametric regression framework are omitted here, but can be found in section 6.3.1, Chapter 6 and in Wood (2006) and Wood et al. (2015).

$$g(y_i^{ct}) = \alpha + X_i^T\beta + \sum_{k=1}^{K} f_k X_i^* + Z_i^{fT} c + Z_i^{rT} u + \epsilon_i \tag{7.1}$$

where

y_i^{ct} is component c of journey time evaluated for trip i undertaken during time period t, with link function $g(\cdot)$,

α is the model constant,

β are the parameter coefficient estimates for the covariates X_i modelled parametrically,

$f_k, k = 1..K$ are the smooth basis functions based on penalised thin plate regression splines of the covariates X_i^* modelled non-parametrically,

c is a vector of estimated group-specific fixed effects for the fixed categorical factors Z_i^f,

u is a vector of estimated group-specific random effects for the random categorical factors Z_i^r, and

ϵ_i is the random error term, which is independently and identically distributed with mean 0, such that $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

As discussed in Chapter 6, a log transformation is applied to the journey times and all covariates in the model, and the link function $g(\cdot)$ is defined as the identity function. The dependent variables in the models are further detailed in section 7.3.2.

Four types of variables are modelled. The first are the parametric variables, which are those variables that follow known parametric forms (for example, linear, quadratic, exponential). The second are the non-parametric variables which are modelled via thin plate regression splines generated from the data points. The parametric and non-parametric variables are continuous variables that capture different service supply and demand characteristics which are discussed further in sections 7.3.4 and 7.3.5.

The third and fourth type of variables are the categorical factors that capture group-specific effects. These factors can be considered as index variables for which longitudinal observations exist, and they represent the unobserved variation between the different categories of the variable (Greene, 2018). In the analysis, group effects are assigned for the following: individual passengers, stations, OD routes, and line/direction. As defined in equation 7.1, the group effects can be modelled in two forms: as either fixed or random. The main distinction between the two forms is that the fixed effects assume that the group-effects are correlated with the other covariates in the model, while the random effects assume that there is no correlation between the group-effects and the other covariates (Greene, 2018; Searle et al., 1992).

To define the effects in general mathematical terms, the fixed effects variables Z_i^f are correlated with the parametric and non-parametric covariates in the model X_i and X_i^*. The expected mean of the vector of fixed parameters c is conditional on the other variables in the model such that $E[c|X_i, X_i^*]$, with variance $Var[c|X_i, X_i^*]$ (Greene, 2018; Wood, 2006). A fixed effects structure is more appropriate if the variable has been drawn from a finite population, where inferences regarding the effect of the variable are confined to the categories of the variable included the model (Greene, 2018; Searle et al., 1992). The fixed effects structure is therefore applied for the group effects for stations, OD routes, and line/direction. This is consistent with the models in Chapter 6.

Conversely, a random effects structure is more appropriate in cases where the variable has been drawn from a large or infinite population, and the observations used in the model are considered a random sample of the population (Greene, 2018; Searle et al., 1992). Under a random effects structure, the vector of random effects coefficients u is considered to be an independently and identically distributed random variable with mean 0 and variance σ_u^2, i.e. $u \sim \mathcal{N}(0, \sigma_u^2)$ (Greene, 2018; Wood, 2006). In the data set, individuals are identified from their unique Oyster card serial numbers. These unique identifiers are defined as categorical factors in the models, and are modelled as random effects. A random effects structure is the most appropriate form as the individual passengers in the analysis represent a sample of a large population of commuters in London. Moreover, the additional passenger characteristics of age, card, discount types, and trip frequency included in the previous analysis are not included in the models here. The individual identifiers solely represent passenger characteristics in the models, and so correlation with other covariates is not anticipated, further supporting the use of the random effects structure.

7.3.2 Dependent variables

The dependent variable in equation 7.1 is the journey time of each trip in the data set, indexed by time of day t and journey time component c. There are a total of four journey time components and three time periods, and so a total of twelve models are estimated.

The components c of journey time for an individual passenger trip i assigned to train j are defined as follows:

1. Access time – time from passenger tap-in t_i^{entry} to departure of assigned train at origin station DT_{oj}

$$y_{ij}^{ac} = DT_{oj} - t_i^{entry} \tag{7.2}$$

2. On-train time – time between train departure at the origin station DT_{oj} to departure at the destination station DT_{dj}

$$y_{ij}^{ot} = DT_{dj} - DT_{oj} \tag{7.3}$$

3. Egress time – time from train departing at destination station DT_{dj} to passenger tap-out time t_i^{exit}

$$y_{ij}^{eg} = t_i^{exit} - DT_{dj} \tag{7.4}$$

4. Total journey time – total time from passenger tap-in at the origin station t_i^{entry} to passenger tap-out at the destination station t_i^{exit}

$$y_i^{jt} = t_i^{exit} - t_i^{entry} \tag{7.5}$$

The journey times are further indexed by time of day t, which is arranged into three sub-categories as follows:

1. Morning (AM) peak – 7am to 10am

2. Midday inter-peak – 10am to 4pm

3. Afternoon (PM) peak – 4pm to 7pm

The descriptive statistics for the dependent variables corresponding to the twelve regression models are summarised in Table 7.2. Generally, the journey times are longer in the AM and PM peak periods, and shorter during the inter-peak. This is as expected, and is most likely a consequence of higher levels of passenger congestion during peak times compared to off-peak times.

Table 7.2: Summary statistics of dependent variables

		Summary statistics (mins)			
Component	Time period	Min.	Max.	Mean	s
Access time	AM peak	0.10	37.80	3.22	2.00
	Inter-peak	0.10	73.88	3.00	1.89
	PM peak	0.10	44.50	3.37	2.17
On-train time	AM peak	0.77	49.00	10.42	5.69
	Inter-peak	1.02	46.75	9.50	5.72
	PM peak	0.18	43.43	10.87	6.70
Egress time	AM peak	0.10	23.73	1.41	0.80
	Inter-peak	0.10	15.78	1.25	0.81
	PM peak	0.10	29.33	1.28	0.87
Total journey time	AM peak	3.00	57.00	15.06	6.30
	Inter-peak	3.00	91.00	13.75	6.16
	PM peak	3.00	60.00	15.52	7.35

Note: s is the sample standard deviation.

7.3.3 Model covariates

In the regression models, variables must be included to control for variance in the physical and operational characteristics of the network. As mentioned previously, the network covariates are modelled at a more aggregate level compared to the detailed analysis of the supply and demand characteristics presented in Chapter 6. The covariates in each of the access, on-train, egress, and total journey time models are summarised in Table 7.3, and are discussed in further detail in the next sections.

Table 7.3: Summary of regression covariates by model

Demand covariates		Service supply covariates	
Individual passenger characteristics	Passenger demand levels	Dynamic	Static
Passenger effect (All)	Platform loading (All)	Headway (All)	Train route stops (T,J)
		cv headway (All)	Group effects for stations (A,E,J),
	Line loading (All)	Norm. headway (All)	OD route (T,J), and
		Speed (T,J)	line/direction (All)
		Time of day (All)	

Note: A – Access time model; T – On-train time model; E – Egress time model;
J – Total journey time model; All – All models, and *cv* is the coefficient of variation.

Table 7.4: Number of passengers per time period

Time period	No. passengers	Mean no. trips
AM peak	2326	31.9
Inter-peak	1956	34.5
PM peak	1830	32.3

7.3.4 Passenger demand covariates

The effects of the passenger characteristics are captured at an individual level, identified via their unique Oyster card serial numbers. As mentioned in the discussion of the regression model framework, the unique identifiers are defined as categorical factors in the models, and are modelled as random effects.

The number of individual passengers by time period are summarised in Table 7.4. In the AM and PM peak periods, the passengers typically make 32 trips on average on the same OD route over a period of seven weeks. In the midday inter-peak period, on average the individual passengers make approximately 35 trips along the same OD route, which equates to approximately one one-way trip per working day for seven weeks. It can therefore be inferred that the passengers in the data set are daily commuters who are likely familiar with the routes and station layouts.

The individual characteristics aside, a set of covariates are included to represent passenger demand levels at three different locations for each trip. Indicators of passenger volumes are evaluated as follows: platform loading at the origin station, line loading, and platform loading at the destination station. As described in Chapter 6, the values of passenger loading are determined via the TfL Rolling Origin and Destination survey (RODs). The indicator for passenger volumes ρ_{qt} for the quantity of interest q (origin platform loading, line loading, destination platform loading) is defined as the number of trips n_{qt} during time period t, normalised by the average number of trips $\bar{n}_{qt,day}$ per 15 minute period t over the day as shown in equation 7.6. The summary statistics of the passenger volume indicators are given in Table 7.5 by time period.

$$\rho_{qt} = \frac{n_{qt}}{\bar{n}_{qt,day}} \tag{7.6}$$

Table 7.5: Summary statistics of demand covariates

Covariate	Units	Time period	Min.	Max.	Mean	s
Origin platform loading	Numeric	AM peak	0.30	3.00	1.88	0.96
		Inter-peak	0.30	2.59	0.80	0.26
		PM peak	0.43	3.00	2.26	0.89
Dest. platform loading	Numeric	AM peak	0.30	3.00	2.23	0.93
		Inter-peak	0.30	2.02	0.82	0.26
		PM peak	0.44	3.00	1.98	0.73
Line loading	Numeric	AM peak	0.30	3.00	2.12	0.72
		Inter-peak	0.30	1.59	0.73	0.19
		PM peak	0.48	3.00	1.87	0.59

Note: s is the sample standard deviation.

7.3.5 Supply side covariates

A series of covariates are included to control for the network operational and physical characteristics. The dynamic operating covariates are those that vary with time, namely: headways, measures of headway variance, and train speeds, and these are included in the models. As the focus of the analysis is on quantifying individual passenger characteristics, the time-invariant physical characteristics of the stations, OD routes, and line and directions are modelled at an aggregate level. Rather than explicitly including different measures for different physical characteristics, each station, route, and line/direction is defined as a categorical variable which represents all physical characteristics and any other residual time-invariant properties associated with the entity. It should be noted that unlike the models in Chapter 6, time of day is not included as an additional variable, as it is already conditioned for through the partitioning of the dependent variables by time period. Each covariate is discussed in further detail in the following sections, and the corresponding summary statistics of each covariate by time of day are given in Table 7.6.

Headway - The headway associated with each trip is obtained through the passenger to train assignment process. Different service frequencies are in operation at different times of the day on different lines. This leads to fluctuations in platform wait times, which correspond to fluctuations in journey times. In the regression model, the headway for the train that the passenger has boarded is used.

Coefficient of variation (cv) of headway - The coefficient of variation of headway is calculated over a 15 minute period for each OD pair. This covariate is included to capture the variation in train frequencies which would otherwise skew the model results, particularly at the beginning and end of service early and late in the day, and at the transition periods between peak and off-peak times. It should be noted that headway is the only covariate in the models where a measure of its associated variance is included. The variance of other covariates do not affect the model results to the same degree, and so have not been included in order to maintain model parsimony.

Headway normalised by mean headway - The train headway associated with each passenger trip is normalised by the mean value of headway over the corresponding period of 15 minutes for each OD pair. This covariate is included to specifically capture any potential effects of train bunching. As highlighted in Chapter 3, in the train movement data set, trains are observed to operate at headways shorter than the minimum scheduled operating headway of approximately 1.67 minutes. Approximately 15% of all trips in the data set are undertaken when trains operate at headways less than 1.67 minutes.

Train speed - Speed is included as a covariate in the on-train time and total journey time models, and is calculated by taking the distance of each trip and dividing by the on-train time of the trip. To maintain consistency in units with the other covariates in the models, speed is measured here in kilometres per minute.

In terms of static covariates, the number of intermediate stops that a train makes along its route is included in the on-train time and total journey time models. The final set of static covariates represent the physical characteristics and residual group-specific effects of stations, OD routes, and lines by direction. As mentioned in the discussion of the model framework, these covariates are modelled with fixed effects to accommodate potential correlation with the other multiple variables capturing network properties.

Table 7.6: Summary statistics of service supply covariates

Covariate	Units	Time period	Min.	Max.	Mean	s
Dynamic service supply covariates						
Headway	Mins	AM peak	1.00	19.98	2.50	1.23
		Inter-peak	1.00	18.93	3.15	1.51
		PM peak	1.00	19.78	2.69	1.48
cv headway	Numeric	AM peak	0.05	1.10	0.28	0.14
		Inter-peak	0.05	1.10	0.30	0.17
		PM peak	0.05	1.10	0.30	0.16
Normalised headway	Numeric	AM peak	0.05	4.48	1.00	0.30
		Inter-peak	0.12	4.51	1.01	0.32
		PM peak	0.08	4.61	1.01	0.33
Speed	Km/min	AM peak	0.12	4.93	0.62	0.12
		Inter-peak	0.17	6.03	0.66	0.12
		PM peak	0.08	62.45	0.63	0.36
Static service supply covariates						
Stops	Numeric	AM peak	1.00	14.00	4.58	2.43
		Inter-peak	1.00	14.00	4.42	2.56
		PM peak	1.00	14.00	4.72	2.78
Entry station	Categorical	AM peak	33 stations			
		Inter-peak	33 stations			
		PM peak	32 stations			
Exit station	Categorical	AM peak	35 stations			
		Inter-peak	35 stations			
		PM peak	34 stations			
OD route	Categorical	AM peak	282 routes			
		Inter-peak	326 routes			
		PM peak	296 routes			
Line/direction	Categorical	AM peak	6 combinations of 3 lines and 2 directions			
		Inter-peak	6 combinations of 3 lines and 2 directions			
		PM peak	6 combinations of 3 lines and 2 directions			

Note: s is the sample standard deviation.

7.3.6 Model evaluation criteria

There are a number of criteria used to assess the goodness-of-fit and parameter significance of the regression models. As the criteria used here are the same as those detailed in Chapter 6, only a brief summary of the relevant indicators is provided here. A more thorough presentation of the mathematical formulae associated with each indicator can be found in section 7.3.6 in Chapter 6.

To determine whether the individual passenger effects are indeed influential on journey times, indicators of significance are used. T-statistics and p-values associated with each covariate are reported, with a lower p-value (and higher t-statistic) indicating that the covariate parameters are more significant in expressing variation in the data (Dougherty, 2016).

In terms of the goodness-of-fit of the models, the semiparametric regression 'bam' function in R outputs a number of indicators, including: the adjusted coefficient of determination $R^2_{adj.}$, the proportion of deviance explained by the model $D_{explained}$, the restricted maximum likelihood (REML) score, and the Akaike Information Criterion (AIC). As the model is fitted using restricted maximum likelihood methods with penalised iteratively reweighted least squares, three indicators are based on model likelihood, with the exception of the adjusted coefficient of determination indicator.

The adjusted coefficient of determination indicator is more applicable for models fitted using the ordinary least squares method, however, it is used here in conjunction with the other likelihood-based indicators to verify model fit. The adjusted coefficient of determination provides an indication of the degree of variance that is explained by the model while penalising for the number of predictors in the model. A higher value of the adjusted coefficient of determination indicates that a greater degree of variance in the data is explained by the model (Dougherty, 2016).

An analogous indicator based on model likelihood is an indicator that captures the proportion of deviance explained by the model. This indicator assesses the degree to which the model explains variation in the data based on the maximised likelihood of the model, relative to the maximised likelihood of the model with no explanatory covariates (Wood, 2006). As with the adjusted coefficient of determination, a higher value of the model deviance explained implies that a higher degree of variance is captured by the model. The restricted maximum likelihood score of the model is also reported as a stand-alone indicator, where lower values indicate a better fitted model (Wood, 2006). This value is also used as an input into the AIC. Smaller values of the AIC indicate a better fitted model (Hastie et al., 2009).

7.4 Results

Twelve models are generated in the analysis: four measures of journey time including access time, on-train time, egress time, and the total journey time are modelled at three time periods covering the AM peak, the midday inter-peak, and the PM peak. The following sub-sections begin with a summary of the modelling process and include a discussion of the following: the data cleaning process, the development of the final model form, covariate significance, and model residuals. These sub-sections are followed by a summary of the elasticity results for the network supply and demand covariates, a detailed discussion of the individual passenger effects, a discussion of the station and line fixed effects, and a case study of the individual passenger effects for the Oxford Circus to Brixton OD route.

7.4.1 Data preparation

As a subset of the filtered data set used in Chapter 6 is used in the models in this chapter, additional data filtering is not required as outliers have already been removed. The upper and lower bounds for all covariates are tabulated in the discussion of model covariates in sections 7.3.4 and 7.3.5.

Correlation between covariates is tested to determine the degree of linear association via Pearson correlation testing and for the presence of non-parametric monotonic associations through Spearman correlation testing. A value of 1 indicates a positive correlation, a value of -1 indicates a negative correlation, and a value of 0 indicates no correlation for both tests (Wonnacott and Wonnacott, 1990). The correlation matrices show that the highest degree of correlation occurs between the covariates representing headways and normalised headways. The Spearman correlation coefficient has a maximum value of 0.72 and the Pearson correlation coefficient is 0.66. As with the models in the previous chapter, the maximum values of the correlation coefficients do not indicate strong correlations between the covariates, and so it is appropriate to initially include all covariates in the models and conduct further model refinement based on covariate significance values as required (discussed further in section 7.4.3).

7.4.2 Model form

The main covariates of interest are the covariates that capture individual passengers measured through their unique card number identifiers, and these are modelled as random effects. As discussed in the section 7.3.1, the random effects form is chosen as the passengers within the data represent a sample of a wider population of passengers and there are no other covariates in the models that capture passenger characteristics, and so correlation with other covariates is not anticipated. An alternative form for the passenger effects is a fixed effects structure, which is used when the data are drawn from a finite sample confined to the values present in the model. The fixed effects are able to accommodate correlation with other covariates in the model. To verify the use of the random effects structure, an alternate model form is trialled with the passenger effects modelled as fixed effects. Furthermore, to establish whether the inclusion of individual passenger effects is warranted, the models are also generated excluding the individual passenger effects.

The covariates capturing the base station, OD route and line/direction effects are modelled as fixed effects as established through trials of model form in Chapter 6. The remaining variables that capture operational characteristics of the system are measured on a continuous scale, and are modelled as smooth terms using non-parametric regression splines. In contrast to the regression models in Chapter 6, complex parametric forms of these covariates are not trialled, as the focus of this analysis is quantifying the passenger effect.

The model goodness-of-fit statistics for three model forms are presented in Tables 7.7, 7.8, and 7.9 for the following: 1) Models with all continuous covariates modelled with non-parametric smooths, fixed network effects, and random passenger effects (Table 7.7) 2) Models with all continuous covariates modelled with non-parametric smooths, fixed network effects, and fixed passenger effects (Table 7.8), 3) Models with all continuous covariates modelled with non-parametric smooths, fixed network effects, and no passenger effects (Table 7.9). Four goodness-of-fit criteria as recommended by Wood (2006) are used to assess the model performance: the adjusted coefficient of determination ($R^2_{adj.}$), the percentage of explained model deviation ($D_{explained}$), the Akaike Information Criterion (AIC), and the restricted maximum likelihood (REML) score.

Following on from the regression models in Chapter 6, a log-log form is applied across all models as this form performs best in terms of the model evaluation criteria. The covariates capturing individual passenger effects are found to be significant at a significance level of 99.9% in the access, egress, and total journey times models, however, the passenger effects are not significant in the on-train time models at a lower bound of 90% significance. As such, the results for the on-train time models are not further presented (further discussion is included in section 7.4.3 on covariate significance).

In terms of the comparison between the random and fixed effects structures for the passenger effects, Table 7.7 and 7.8 show that the model performance is similar. The $R^2_{adj.}$ and $D_{explained}$ indicators remain unchanged for all total journey time models, however, the AIC and REML scores are generally lower for the model forms with random passenger effects. This indicates that the total journey time models perform better with a random effects structure. For the access and egress models in the AM period, the $R^2_{adj.}$ and $D_{explained}$ indicators remain unchanged and the AIC and REML scores are lower for the models with random passenger effects, indicating that the random effects structure is better for these models. For the remaining time periods, however, the criteria results are more mixed. For each model, across some criteria, the random effects structure is preferred and across other criteria the fixed effects structure is preferred. Overall, the majority of models perform better with a random effects structure, and so to maintain consistency and adherence to the theoretical basis for the application of random effects, a random effects structure for the passenger effects is adopted for the final model forms.

In terms of investigating the inclusion of the passenger effects, for the regression models that do include the passenger effect, the $R^2_{adj.}$ and $D_{explained}$ indicators are higher, and the AIC and REML scores are lower across all models. The inclusion of heterogeneity across individual passengers therefore improves the explanatory power of the regression models, which indicates that individual passenger characteristics do contribute to journey time variance. Comparing across models, the inclusion of passenger effects has a greater impact on the goodness-of-fit statistics for the access and egress time models. With the inclusion of passenger effects, the $R^2_{adj.}$ and $D_{explained}$ indicators increase by approximately 52-102% for the access time models and 42-72% for the egress time models, depending on the time of day when comparing model form 1 (Table 7.7) with model form 3 (Table 7.9). The same indicators increase by approximately 2% across all time periods with the inclusion of passenger effects in the total journey time model when comparing model form 1 with model form 3. The higher impact of the passenger effect in the access and egress time models is as expected; these components of journey time capture passenger walking and waiting behaviours at the stations. For the total journey time, individual passenger effects are not as influential on overall journey times, however, the inclusion of these effects does improve the performance of the models.

For the final model forms with random passenger effects (model form 1, see Table 7.7), the $R^2_{adj.}$ and $D_{explained}$ indicators of the total journey time models including passenger effects are in the order of 93-95%, while the same indicators lie in the order of 26-35% for the access and egress time models. The total journey time models include individual effects for each OD route in addition to the individual effects for stations and operational variables that are included in the access and egress time models. The inclusion of these additional covariates therefore increases the explanatory power of the total journey time models. The lower explanatory power of the access and egress time models indicates that there are likely missing covariates which not been able to be included in the models due to a lack of information.

Table 7.7: Model form 1 goodness-of-fit statistics - final model form

Time period	Indicator	Access time	Egress time	Total journey time
AM peak	n	58308	58308	58308
	$R^2_{adj.}$	0.28	0.28	0.93
	$D_{explained}$	30%	30%	93%
	AIC	9.2E4	8.7E4	-8.9E4
	REML	4.7E4	4.4E4	-4.2E4
Inter-peak	n	18355	18355	18355
	$R^2_{adj.}$	0.26	0.31	0.93
	$D_{explained}$	30%	35%	94%
	AIC	2.9E4	3.1E4	-5.9E4
	REML	1.5E4	1.6E4	-1.2E4
PM peak	n	34217	34217	34217
	$R^2_{adj.}$	0.30	0.26	0.94
	$D_{explained}$	32%	29%	95%
	AIC	5.0E4	5.6E4	-5.5E4
	REML	2.6E4	2.9E4	-2.5E4

Note: n is the number of observations, $R^2_{adj.}$ is the adjusted coefficient of determination, $D_{explained}$ is deviance explained, AIC is Akaike Information Criterion, and REML is restricted maximum likelihood.

Table 7.8: Model form 2 goodness-of-fit statistics - fixed passenger effects

Time period	Indicator	Access time	Egress time	Total journey time
AM peak	n	58308	58308	58308
	$R^2_{adj.}$	0.28	0.28	0.93
	$D_{explained}$	31%	31%	93%
	AIC	9.3E4	8.8E4	-4.8E4
	REML	4.7E4	4.5E4	-4.0E4
Inter-peak	n	18355	18355	18355
	$R^2_{adj.}$	0.27	0.32	0.93
	$D_{explained}$	35%	39%	94%
	AIC	2.9E4	3.2E4	-2.8E4
	REML	1.4E4	1.5E4	-1.0E4
PM peak	n	34217	34217	34217
	$R^2_{adj.}$	0.30	0.26	0.94
	$D_{explained}$	34%	30%	95%
	AIC	5.0E4	5.7E4	-5.5E4
	REML	2.5E4	2.8E4	-2.3E4

Note: n is the number of observations, $R^2_{adj.}$ is the adjusted coefficient of determination, $D_{explained}$ is deviance explained, AIC is Akaike Information Criterion, and REML is restricted maximum likelihood.

7.4.3 Covariate significance

The final form and significance of the model covariates are presented by time period in Tables 7.10, 7.11, and 7.12. As shown in the tables, the individual passenger effects are significant at a level of 99.9% in all access, egress, and total journey time models. The group-specific

Table 7.9: Model form 3 goodness-of-fit statistics - without passenger effects

Time period	Indicator	Access time	Egress time	Total journey time
AM peak	n	58308	58308	58308
	$R^2_{adj.}$	0.15	0.20	0.91
	$D_{explained}$	15%	20%	91%
	AIC	1.0E5	9.2E4	-8.0E4
	REML	5.0E4	4.6E4	-3.8E4
Inter-peak	n	18355	18355	18355
	$R^2_{adj.}$	0.17	0.22	0.92
	$D_{explained}$	18%	22%	92%
	AIC	3.0E4	3.2E4	-3.2E4
	REML	1.5E4	1.6E4	-1.1E4
PM peak	n	34217	34217	34217
	$R^2_{adj.}$	0.16	0.17	0.92
	$D_{explained}$	16%	17%	92%
	AIC	5.5E4	5.9E4	-5.3E4
	REML	2.7E4	3.0E4	-2.1E4

Note: n is the number of observations, $R^2_{adj.}$ is the adjusted coefficient of determination, $D_{explained}$ is deviance explained, AIC is Akaike Information Criterion, and REML is restricted maximum likelihood.

fixed effects for stations, OD routes, and line/direction are also significant in all models. The operational variables capturing headways and the passenger demand covariates are significant in the AM peak and PM peak models, however, these covariates are less significant in the inter-peak models.

The passenger effect is not significant in any of the on-train time models. This result is as expected; in the regression models in Chapter 6, the covariates capturing general passenger characteristics are also not significant in the on-train time models. The network operational characteristics and physical route characteristics are the primary determinants of the variance in on-train times. Consequently, the results of the on-train time models are not further presented.

7.4.4 Model residuals

The error structure for the models is specified such that the errors ϵ_i are identically and independently distributed with mean 0, i.e. $\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon)$. The errors approximate a normal distribution around an approximate mean of 0 for each model. For reference, the summary statistics and kernel density plots of the model residuals are given in Appendix G.

Table 7.10: Covariate significance - AM peak

Covariate	Access time Form	Access time Sig.	Egress time Form	Egress time Sig.	Total journey time Form	Total journey time Sig.
Service capacity covariates - dynamic						
Headway	splines	***	splines	**	splines	***
cv headway	splines	***	splines	**	splines	***
Hdway/mean hdway	splines	***	splines	***	splines	***
Speed					splines	***
Service capacity covariates - static						
Stops					splines	
Station	fixed	***	fixed	***	fixed	***
OD					fixed	***
Line/direction	fixed	***	fixed	***	fixed	***
Passenger demand covariates - demand volumes						
Origin plat. loading	splines	***			splines	***
Dest. plat. loading			splines	**	splines	***
Line loading	splines	***	splines	***	splines	***
Passenger demand covariates - individual characteristics						
Indiv. pass.	random	***	random	***	random	***

Significance notation: 0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' 1

Table 7.11: Covariate significance - inter-peak

Covariate	Access time Form	Access time Sig.	Egress time Form	Egress time Sig.	Total journey time Form	Total journey time Sig.
Service capacity covariates - dynamic						
Headway	splines	***	splines	.	splines	***
cv headway	splines	.	splines	*	splines	***
Hdway/mean hdway	splines	***	splines		splines	***
Speed					splines	***
Service capacity covariates - static						
Stops					splines	.
Station	fixed	***	fixed	***	fixed	***
OD					fixed	***
Line/direction	fixed	***	fixed	***	fixed	***
Passenger demand covariates - demand volumes						
Origin plat. loading	splines	.			splines	**
Dest. plat. loading			splines		splines	**
Line loading	splines		splines	*	splines	
Passenger demand covariates - individual characteristics						
Indiv. pass.	random	***	random	***	random	***

Significance notation: 0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' 1

As discussed in Chapters 3 and 6, each individual Oyster card timestamp is accurate to one minute only, and the magnitude of the inaccuracy is unknown. The timestamp inaccuracy is

Table 7.12: Covariate significance - PM peak

Covariate	Access time Form	Sig.	Egress time Form	Sig.	Total journey time Form	Sig.
Service capacity covariates - dynamic						
Headway	splines	***	splines	***	splines	***
cv headway	splines	***	splines	**	splines	***
Hdway/mean hdway	splines	***	splines	***	splines	***
Speed					splines	***
Service capacity covariates - static						
Stops					splines	***
Station	fixed	***	fixed	***	fixed	***
OD					fixed	***
Line/direction	fixed	***	fixed	***	fixed	***
Passenger demand covariates - demand volumes						
Origin plat. loading	splines	***			splines	***
Dest. plat. loading			splines	***	splines	***
Line loading	splines	***	splines	***	splines	***
Passenger demand covariates - individual characteristics						
Indiv. pass.	random	***	random	***	random	***

Significance notation: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

assumed to follow a uniform random distribution which is captured in the random error term of the regression models. It should be noted that the inaccuracy in the timestamps does not impact the magnitude of the parameter estimates. In the access and egress time models, the random error due to the inaccuracy is in the order of 0-59 seconds, and in the total journey time models the random error is in the range of 0-118 seconds. The AVL data are accurate to one second and so the random error due to timestamp inaccuracy for the on-train time model is in the order of less than one second. The mean access and egress times are 3.30 and 1.41 minutes across all time periods, respectively, and so the average timestamp errors are proportionally 15% and 35% of the mean access and egress times. Mean total journey times are 15.02 minutes across all time periods, and so the average timestamp error equates to 7% of the mean journey time.

7.4.5 Network supply and demand covariates

The mean predicted values and elasticities for the significant network supply and demand covariates are given for each journey time model arranged by time period in Tables 7.13, 7.14, and 7.15. The mean predicted values represent the average effect that each covariate contributes to the variance in journey times in each model. The mean elasticities for the continuous variables are also provided for completeness, and represent the marginal effect on journey times associated with a unit increase in the covariate value (refer to section 7.4, Chapter 6 for the general mathematical definition of covariate elasticities).

It should be noted that excluding the passenger effects, the results of the models are consistent with those for the regression models in the previous chapter, which focus on the effects of service supply and passenger demand factors. Across all time periods and journey time measures, the static station, OD route, and line direction effects have the greatest effect, followed by the operational factors of speed (in the total journey time model), headways, and passenger demand levels.

Table 7.13: Mean predicted values and covariate elasticities - AM peak

	Access time		Egress time		Total journey time	
	Value	e	Value	e	Value	e
Service capacity covariates - dynamic						
Headway	-1.3E-3	3.6E-1	3.8E-5	-1.6E-2	-9.6E-4	7.3E-2
cv headway	3.3E-5	2.9E-2	4.6E-5	1.2E-2	-7.5E-5	1.1E-2
Headway/mean headway	2.1E-4	-9.1E-2	2.3E-4	-5.7E-2	2.2E-4	-6.2E-2
Speed		-		-	2.1E-4	-6.4E-1
Passenger demand covariates - dynamic						
Origin plat. loading	1.0E-3	1.4E-1			4.5E-4	4.0E-2
Dest. plat. loading			-9.6E-4	-4.1E-2	4.0E-4	2.0E-2
Line loading	2.2E-3	3.1E-1	6.2E-4	6.1E-2	8.8E-4	6.9E-2
Station and route fixed effects						
Entry station	-1.2E-1		NA	-	-1.5E5	
Exit station	NA		3.3E-1		2.1E4	
OD	NA		NA		5.0E4	
Line/direction	4.4E-1		1.3E-2		1.0E4	
Model intercept	6.7E-1		-7.4E-2		6.6E4	
Mean journey times	1.2E0		4.1E-1		2.7E0	
(Mins)	(3.2E0)		(1.5E0)		(1.5E1)	

Note: All values are in log-mins and elasticities (denoted e) are unit-less, unless otherwise stated.

Table 7.14: Mean predicted values and covariate elasticities - inter-peak

	Access time Value	e	Egress time Value	e	Total journey time Value	e
Service capacity covariates - dynamic						
Headway	4.9E-3	4.3E-1	2.1E-5	-5.1E-2	-1.6E-4	9.5E-2
cv headway	2.1E-4	1.8E-2	-1.4E-6	2.1E-2	2.1E-6	9.3E-3
Headway/mean headway		5.1E-2			4.3E-5	-1.1E-2
Speed		-		-	-1.4E-3	-5.5E-1
Passenger demand covariates - dynamic						
Origin plat. loading	-1.9E-4	3.2E-2			-9.0E-5	5.8E-3
Dest. plat. loading					-1.4E-4	2.0E-2
Line loading			-1.6E-4	6.9E-2		
Service capacity covariates - static						
Stops		-		-	-9.2E2	7.5E4
Station and route fixed effects						
Entry station	-1.8E-1		NA	-	3.8E5	
Exit station	NA	-	6.1E-2		3.7E5	
OD	NA	-	NA		-1.6E5	
Line/direction	2.0E-2		2.3E-1		-4.6E5	
Model intercept	1.1E0		-2.2E-1		-1.2E5	
Mean journey times	1.1E0		2.6E-1		2.6E0	
(Mins)	(3.1E0)		(1.3E0)		(1.4E1)	

Note: All values are in log-mins and elasticities (denoted e) are unit-less, unless otherwise stated.

7.4.6 Individual passenger effects

The passenger effects are treated as random intercepts in the models, and are used to predict journey times at an individual passenger level. This can be expressed in mathematical terms for a general regression model equation with random effects u_i:

$$Y_i = \alpha + \beta X_i + u_i + \epsilon_i \tag{7.7}$$

The model is interpreted in relation to the prediction of the realised value of the random effect for each individual, which can be expressed as:

$$E[Y_i|X_i, u_i], \text{where } Y_i|X_i, u_i \sim \mathcal{N}(\alpha + \beta X_i + u_i, \sigma^2_\epsilon) \tag{7.8}$$

Table 7.15: Mean predicted values and covariate elasticities - PM peak

	Access time		Egress time		Total journey time	
	Value	e	Value	e	Value	e
Service capacity covariates - dynamic						
Headway	-1.6E-3	2.8E-1	-7.3E-4	3.4E-2	-7.3E-4	6.7E-2
cv headway	-1.9E-5	3.1E-2	-1.4E-4	2.1E-2	-1.2E-4	1.6E-2
Headway/mean headway	6.4E-5	-3.8E-2	2.1E-4	-8.6E-2	8.4E-5	-5.0E-2
Speed		-		-	-1.5E-4	-5.8E-1
Passenger demand covariates - dynamic						
Origin plat. loading	9.0E-4	1.2E-1			2.5E-4	2.8E-2
Dest. plat. loading			5.8E-4	8.1E-2	2.7E-4	2.7E-2
Line loading	8.4E-4	1.3E-1	7.3E-4	9.1E-2	5.3E-4	5.5E-2
Service capacity covariates - static						
Stops		-		-	-3.0E2	1.6E5
Station and route fixed effects						
Entry station	-1.6E-1		NA	-	3.6E5	
Exit station	NA		3.0E-2		-9.3E5	
OD	NA		NA		-8.4E4	
Line/direction	2.1E-2		3.6E-1		-1.8E5	
Model intercept	1.2E0		-2.6E-1		8.4E5	
Mean journey times	1.2E0		3.0E-1		2.7E0	
(Mins)	(3.4E0)		(1.3E0)		(1.5E1)	

Note: All values are in log-mins and elasticities (denoted e) are unit-less, unless otherwise stated.

Where

Y_i represents journey time for each individual passenger i,

α is the regression model intercept,

X_i represents all other covariates in the regression models (excluding passenger effects), and β are the associated coefficients,

u_i is the value of the random effect associated with passenger i,

ϵ_i is the model random error term, and σ^2_ϵ is the variance of the model error term.

The summary statistics for the distribution of the individual passenger effects are provided in Table 7.16. Figures 7.4, 7.5, and 7.6 illustrate the probability density distributions of the individual effects using kernel density estimation.

Table 7.16: Individual passenger effects summary statistics

	Time period	Summary statistics (log-mins) Min.	Max.	Mean	s	Range
Access time	AM peak	-0.87	1.29	1.2E-14	0.20	2.16
	Inter-peak	-0.60	1.12	-2.6E-15	0.12	1.72
	PM peak	-0.74	1.55	-3.5E-15	0.18	2.29
Egress time	AM peak	-0.63	0.62	-5.9E-15	0.15	1.24
	Inter-peak	-0.78	0.56	5.0E-15	0.14	1.34
	PM peak	-0.98	0.50	-2.0E-15	0.15	1.47
Total journey time	AM peak	-0.22	1.05	-1.1E-12	0.06	1.27
	Inter-peak	-0.20	0.82	-8.1E-12	0.05	1.01
	PM peak	-0.23	1.11	4.2E-11	0.06	1.34

Note: s is the sample standard deviation.

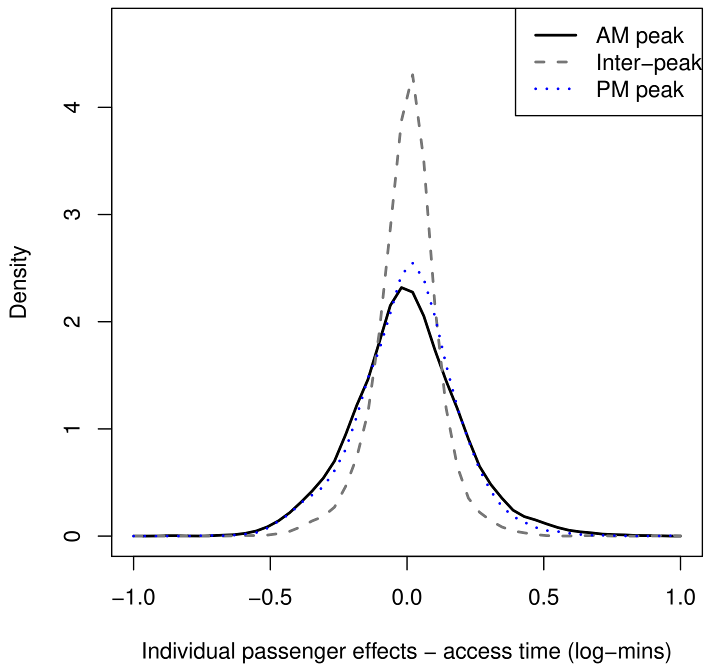

Figure 7.4: Distribution of individual passenger effects by time period - access times

As shown in Figures 7.4 to 7.6 and Table 7.16, the passenger effects for all models are approximately symmetrically distributed around an approximate mean value of 0 log-mins. This is consistent with the random effects structure applied, which specifies that the passenger effects u should be normally distributed with mean 0 such that $u \sim \mathcal{N}(0, \sigma_u^2)$, where is σ_u^2 the variance of the individual passenger effects.

Comparing across the components of journey time, the range and standard deviation of the

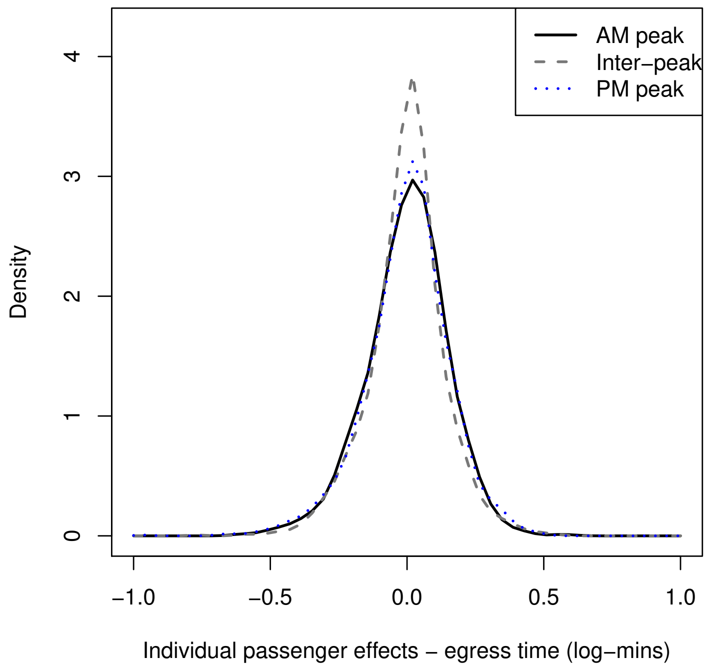

Figure 7.5: Distribution of individual passenger effects by time period - egress times

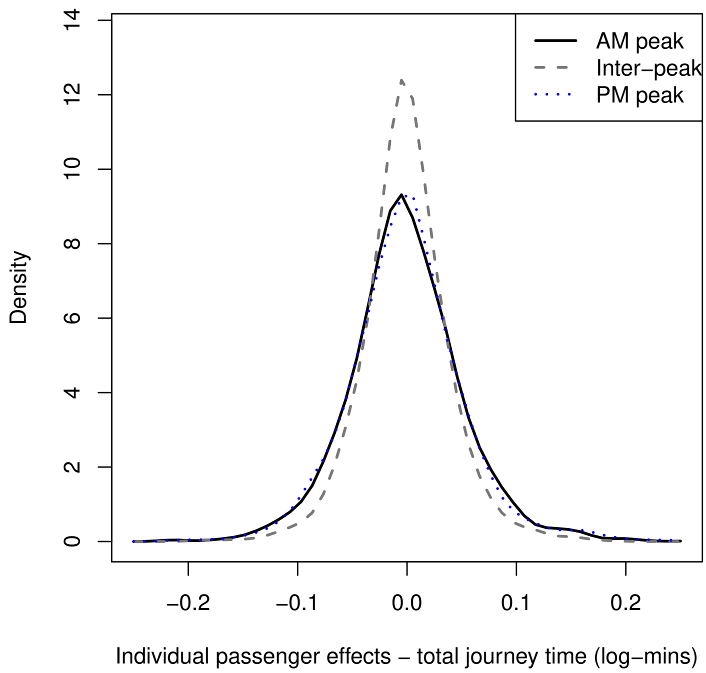

Figure 7.6: Distribution of individual passenger effects by time period - total journey times

passenger effects are generally greatest for the access time models, followed by the egress time models, and the total journey time models. This indicates that heterogeneity between individual passengers has the greatest degree of influence for the access time models. The access time component captures the walking as well as waiting time of the passenger from entry at the origin station to train boarding. Individual-based characteristics relating to walking and waiting times are represented by the passenger effects. For the walking component, individual characteristics include the walking route taken through the station to the platform, and walking speeds. For the waiting component, individual characteristics include passenger arrival time, and passenger boarding decisions such as choosing the location on the platform to board, and choosing whether or not to board, particularly in congested conditions. The egress time component represents the walking time of the passenger from train alighting to exit at the destination station. All characteristics related to individual walking behaviours are represented by the passenger effects. Comparing between the two components of journey time, there are more individual-based passenger choices to be made in the access time component, and so it is therefore expected that heterogeneity in individual passenger effects is more influential in the access time model. In the total journey time model, on-train times represent a large proportion of the journey time variance, and so individual characteristics at the origin and destination stations are relatively less influential.

Considering the influence of the time of day, the distribution of the passenger effects are more variable during the AM and PM periods compared to the inter-peak across all journey time models. The difference between the time periods is most pronounced for the access time models. In the access time models, the passenger effects during the AM and PM peak lie within the range from -0.87 log-mins (25.1 seconds) to 1.55 log-mins (282.2 seconds), while the passenger effects during the inter-peak range from -0.6 log-mins (32.9 seconds) to 1.12 log-mins (183.7 seconds). The narrower range of effects during the inter-peak indicates that there is less heterogeneity among passengers in terms of their influence on midday inter-peak access times, compared to peak times. This result is most likely attributed to the fact that passenger demand levels are less variable during off-peak periods. As demonstrated in section 6.11 of Chapter 6, high levels of passenger demand lead to a higher degree of variance in passenger wait times during peak periods, while in the off-peak, passenger wait times are both lower and less variable as passenger demand levels are lower, and passengers are more likely to board the first available train.

For egress and total journey times, the difference between the three time periods is not as marked (see Figure 7.5 and 7.6). The distribution of passenger effects in the inter-peak are slightly less variable compared to the AM peak and PM peak periods. For the egress time models, the standard deviation of the passenger effects during the AM and PM peak is jointly 0.15 log-mins, while the standard deviation during the inter-peak is marginally less at 0.14 log-mins. Across all time periods, the passenger effect ranges from -0.98 log-mins (22.6 seconds) to 0.62 log-mins (111.0 seconds). For the total journey time models, the standard deviation during the AM and PM peaks is jointly 0.06 log-mins, and is again slightly lower at 0.05 log-mins during the inter-peak. The passenger effect across all time periods ranges from -0.23 log-mins (47.5 seconds) to 1.11 log-mins (182.3 seconds).

As with the access time models, for the egress time and total journey time models, the slightly lower levels of variance in passenger effects during the inter-peak are most likely due to the effect of passenger demand and/or network congestion levels. During the peak periods, passenger volumes are higher and this likely disrupts the travel behaviour of passengers, resulting in higher levels of journey time variance. Passenger volumes are lower in the inter-peak, and as a result, passenger travel behaviours are likely to be more consistent with a comparatively lower degree of variance in journey times.

7.4.7 Station and line effects

The access and egress time models presented in this analysis do not include additional covariates capturing the physical characteristics of the stations. Instead, all static physical qualities relating to the stations are captured by categorical variables with group-specific effects that represent each station. Therefore, as a valuable by-product of the analysis, the group-specific effects of stations in the access and egress time models can be used to provide a ranking of the stations in terms of their journey time performance for regular commuters. The rankings represent for which stations passenger access and egress times are fastest and slowest in reference to the physical station layouts, locations, and any other unobserved static characteristics, after taking into account the effect of all other covariates in the models.

Table 7.17 provides a summary of the station effects for the access and egress models. As shown in the table, passenger access times are fastest at North Greenwich station (Jubilee line), Canary Wharf (Jubilee line), and White City (Central line) in the morning, inter-peak, and afternoon peak periods, respectively. The effect of these stations leads to a -0.57 log-mins (34 seconds) reduction in access times at North Greenwich in the AM peak, a -0.61 log-mins (33 seconds) reduction at Canary Wharf in the inter-peak, and a -0.84 log-mins (26 seconds) reduction at White City in the PM peak. Conversely, access times are slowest at Euston station (Victoria line) in the morning and inter-peak and at Warren Street station (Victoria line) in the afternoon peak. Access times increase by 0.71 log-mins (122 seconds) in the AM peak and by 0.18 log-mins (72 seconds) in the inter-peak at Euston station, and by 0.12 log-mins (67 seconds) in the PM peak at Warren Street.

Egress times are fastest at North Greenwich (Jubilee line) in the morning peak and Canary Wharf (Jubilee line) in the inter-peak and afternoon peaks. Egress times are quicker at these stations by -0.51 log-mins (36 seconds) at North Greenwich in the AM peak, and by -0.70 log-mins (30 seconds) and -0.75 log-mins (28 seconds) at Canary Wharf in the inter-peak and PM peak respectively. On the other hand, egress times are slowest at Notting Hill Gate station (Central line) across all time periods. Passenger egress times are slower by 0.86 log-mins (141 seconds) in the AM and PM peaks, and 0.79 log-mins (132 seconds) slower in the inter-peak.

Table 7.17: Station effects summary statistics

		Summary statistics (log-mins)					
	Time period	Min.	Min. station	Max.	Max. station	Mean	s
Access time	AM peak	-5.7E-1	North Greenwich	7.1E-1	Euston	8.0E-2	3.5E-1
	Inter-peak	-6.1E-1	Canary Wharf	1.8E-1	Euston	-1.6E-1	2.0E-1
	PM peak	-8.4E-1	White City	1.2E-1	Warren Street	-2.3E-1	2.1E-1
Egress time	AM peak	-5.1E-1	North Greenwich	8.6E-1	Notting Hill Gate	2.4E-1	3.5E-1
	Inter-peak	-7.0E-1	Canary Wharf	7.9E-1	Notting Hill Gate	1.2E-1	3.9E-1
	PM peak	-7.5E-1	Canary Wharf	8.6E-1	Notting Hill Gate	1.6E-1	4.2E-1

Note: s is the sample standard deviation.

The line/direction effects variables capture the residual unobserved physical characteristics specific to each line by direction of travel, after taking into account all other covariates in the models. In practical terms, the main differences between the lines and directions in the analysis are more related to operational characteristics; for example, different lines have different train operating frequencies, signalling systems, and rollingstock. Unlike the station effects, it is therefore difficult to ascribe the effects of the lines/directions to distinct tangible characteristics that may have an effect on passenger access and egress times. Nevertheless, for completeness, a summary of the line/direction effects in the access and egress models is given in Table 7.18.

Table 7.18: Line/direction effects summary statistics

		Summary statistics (log-mins)					
	Time period	Min.	Min. l-d	Max.	Max. l-d	Mean	s
Access time	AM peak	-6.9E-2	C-E	8.3E-1	J-W	3.7E-1	4.0E-1
	Inter-peak	-1.0E-1	V-S	1.9E-1	J-W	2.9E-2	1.4E-1
	PM peak	-9.4E-2	V-S	1.6E-1	J-E	4.3E-2	1.2E-1
Egress time	AM peak	-2.2E-1	V-S	2.7E-1	J-E	-5.1E-3	2.4E-1
	Inter-peak	1.7E-2	V-S	5.6E-1	J-E	2.4E-1	2.7E-1
	PM peak	7.3E-2	C-E	7.3E-1	J-E	3.4E-1	3.4E-1

Note: s is the sample standard deviation, l-d denotes line and direction,
C-E refers to Central line eastbound, J-E refers to Jubilee line eastbound,
J-W refers to Jubilee line westbound, and V-S refers to Victoria line southbound.

Generally, passenger access and egress times are quickest for stations on the Victoria and Central lines, and slowest for stations on the Jubilee line. For access times, the effects range from a minimum of -0.10 log-mins on the Victoria line southbound in the inter-peak (54 seconds) to a maximum of 0.83 log-mins (138 seconds) on the Jubilee line westbound in the AM peak. For egress times, the line/direction effects range from a minimum of -0.22 log-mins (48 seconds) on the Victoria line southbound in the AM peak to a maximum value of 0.73 log-mins (124 seconds) on the Jubilee line eastbound in the PM peak.

It should be noted that a discussion of the station, OD route, and line/direction effects is not included for the total journey time models, as these covariates capture only the residual effects after taking into account all other covariates in the model. Covariates capturing the speed (and distance by proxy) and number of intermediate stops along the OD routes are included in the models, and so the group-specific effects for OD routes capture left over characteristics of the routes only. In terms of the station effects, producing rankings of the origin and destination stations in terms of total journey times are not as valuable for operators, compared to quantifying rankings of the stations with respect to passenger access and egress times.

7.4.8 Case study: Oxford Circus to Brixton

Table 7.19: Oxford Circus to Brixton mean predicted values and elasticities

	Access time		Egress time		Total journey time	
	Value	e	Value	e	Value	e
Service capacity covariates - dynamic						
Headway	-7.3E-2	2.7E-1	-2.9E-2	2.1E-2	-2.7E-2	6.1E-2
cv headway	-1.2E-3	3.1E-2	-2.5E-3	2.0E-2	-9.4E-4	1.7E-2
Headway/mean headway	-4.7E-4	-4.2E-2	2.8E-3	-7.8E-2	9.0E-4	-5.1E-2
Speed	-	-	-	-	6.2E-2	-5.9E-1
Service capacity covariates - static						
Stops	-	-	-	-	7.2E4	1.6E5
Passenger demand covariates						
Origin plat. loading	9.2E-3	1.8E-1			1.9E-4	4.5E-2
Dest. plat. loading			2.3E-2	1.5E-1	6.5E-3	3.8E-2
Line loading	2.2E-3	1.7E-1	1.1E-2	8.3E-2	5.5E-3	6.8E-2
Fixed effects						
Entry station	5.4E-2		-		-2.9E5	
Exit station	-		5.8E-2		4.3E5	
OD	-		-		-7.3E5	
Line/direction	-9.4E-2		8.3E-2		-4.0E5	
Random effects						
Indiv. pass.	-3.6E-1 to 5.4E-1		-1.3E-1 to 2.4E-1		-4.2E-2 to 1.2E-1	
Intercept	1.2E0		-2.6E-1		8.4E5	
Residuals	2.8E-2		-8.5E-3		-1.2E-3	
Mean journey times	1.4E0		1.1E-1		2.9E0	
(Mins)	(4.0E0)		(1.1E0)		(1.9E1)	

Note: All values are in log-mins and elasticities (denoted e) are unit-less, unless otherwise stated.

As in Chapter 6, the OD route from Oxford Circus station to Brixton station on the Victoria line southbound is used as a representative route to further illustrate the application of the results from the regression models. Journey times are analysed during one of the busiest times on the route, the PM peak period from 5.30pm to 6pm. A total of 194 trips are made during this period by regular commuters, and the average headway associated with the trips is approximately 1.94 minutes. Table 7.19 summarises the mean predicted values of each significant covariate in log-mins and average elasticities for each journey time component. The predicted values of the covariates are included to show the degree to which each covariate contributes to each journey time model. The average elasticities represent the average rate of change in journey time with respect to a unit increase in the continuous covariates that capture operational factors.

Beginning with the analysis of total journey times, the most influential covariates are the static covariates and fixed effects that capture the physical characteristics of Oxford and Brixton stations, the train route between the two stations including the five intermediate stops, and any residual characteristics specific to the Victoria line in the southbound direction. The physical characteristics aside, the next most influential covariate is train speed, followed by train headway, passenger volumes, and the measures of variance in headways are the least influential covariates. The individual passenger effects range from -0.04 log-mins (58 seconds) to 0.12 log-mins (67 seconds). The extreme ranges of the passenger effects rank behind the mean predicted values of the fixed effects capturing the effects of the stations, OD route, line/direction, and the number of stops along the route.

In the access and egress time components, the effects of passenger heterogeneity are greater. In the access time model, the individual passenger effects range from -0.36 log-mins (42 seconds) to 0.54 log-mins (103 seconds). The extreme ranges of the passenger effects are the most influential factors in the access time model, relative to the mean predicted values of the other variables in the model. Excluding passenger effects, the most influential factor is the mean predicted effect for line/direction, followed by headway, the station effect for Oxford Circus, passenger volumes, and the measures of headway variance.

For the egress time model, the individual passenger effects range from -0.13 log-mins (53 seconds) to 0.24 log-mins (76 seconds). Again, the extreme ranges of the individual passenger effects have a greater influence on egress times compared to the mean predicted values of the other covariates in the model. The passenger effects aside, the line/direction effect is the most influential factor in the model, followed by the station effect for Brixton, passenger volumes, and the covariates that capture variance in headways.

The residuals of the models capture the random sources of variance in journey times. For access times the mean value of the model residuals is 0.03 log-mins, indicating that average access times are slightly underestimated. In contrast, egress times are on average overestimated, with a mean residual value of -0.009 log-mins. The total journey times are also slightly overestimated, with a mean residual value of -0.001 log-mins.

The case study illustrates how the results from the regression analysis can be interpreted. Overall, the individual passenger effects for the Oxford Circus to Brixton route have the greatest degree of spread (or heterogeneity) in the access time model, followed by the egress time model, and the total journey time model.

7.5 Conclusions

In this chapter, semiparametric regression is undertaken to quantify the degree to which individual-specific passenger heterogeneity influences the variance of journey times. The data set covering three lines on the London Underground as used in Chapter 6 is filtered for regular commuters who undertake 30 or more repeat trips on the same origin-destination pair. The influence of individual passenger characteristics is quantified by modelling the individual passengers as random effects in the regression models. Twelve models are generated; one for each component of journey time, namely access, on-train, and egress times, and total journey times over the AM peak, inter-peak, and PM peak.

The passenger effects are significant in all models, with the exception of the on-train time model. For the access, egress, and total journey time models, the goodness-of-fit of all models increases with the inclusion of passenger effects, indicating that individual passenger characteristics are a source of journey time variance. The results for the remaining network supply and demand covariates are consistent with the models presented in Chapter 6; the covariates that capture the static physical characteristics of the network are the most influential covariates across all models.

Across all time periods, the passenger effects range from -0.87 log-mins (25.1 seconds) to 1.55 log-mins (282.2 seconds) for the access time model, -0.98 log-mins (22.6 seconds) to 0.62 log-mins (111.0 seconds) for the egress time models, and -0.23 log-mins (47.5 seconds) to 1.11 log-mins (182.3 seconds) for the total journey times models. The range and standard deviation of the passenger effects are greatest in the access time models, which indicates that heterogeneity in passenger effects have the greatest degree of influence in the access time component. This result is consistent with expectations; passenger behaviours and decisions relating to walking and waiting times at the origin station are captured in the access time model, while passenger characteristics relating to walking only as the passenger exits at the destination station are captured in egress time model. In terms of the influence of the time of day, the results show that the passenger effects are distributed with a lower degree of variance in the inter-peak across all models, compared to the morning and afternoon peak periods, where the passenger effects are distributed across a wider range with a higher degree of variance.

In the access and egress time models, the origin and destination stations are represented by group-specific effects variables that capture the static physical properties of the stations, for example, station dimensions, layouts, and locations. As an additional output of the models, the fixed station effects can be used to produce a ranking to compare the performance of the stations with respect to passenger access and egress walk times. The results show that access and egress times are fastest for North Greenwich (Jubilee line), Canary Wharf (Jubilee line) and White City (Central line) stations. The access and egress times tend to be slowest at Euston (Victoria line), Warren Street (Victoria line), and Notting Hill Gate (Central line) stations.

As with the analysis presented in Chapter 6, a case study of trips between Oxford Circus and Brixton on the Victoria line is included. The case study illustrates how the regression model results can be interpreted and applied to compare the impact of different service supply, demand, and individual passenger characteristics on journey time variance for a typical passenger route.

Chapter 8

Conclusions

8.1 Summary of thesis objectives

The focus of this thesis is the analysis of journey time variance on urban metro systems as experienced by passengers, using the London Underground as a case study. As introduced in Chapter 1, the objectives of the thesis can be summarised under three main areas:

1. Define and quantify the distributional properties of journey times.
2. Decompose total journey times into constituent sub-components.
3. Identify and quantify the factors that influence total journey times and the sub-components of journey time.

The thesis objectives are addressed through the application of statistical analysis methods to large volumes of passenger trip data, train movement data, incident data, and additional data on the physical and operational characteristics of the London Underground network. The statistical methods are applied to four representative lines on the London Underground, namely the Central, Jubilee, Victoria, and Waterloo and City lines.

8.2 Summary of thesis contributions

The research presented in this thesis contributes new empirical results on the variance of journey times on urban metro systems. The combination of multiple data sets for the London Underground enables a deeper level of analysis of journey time variance as experienced by passengers, compared to previous work undertaken in this field. The analysis presented in the thesis comprises of three general parts corresponding to the three main thesis objectives summarised in section 8.1. The following sections summarise the main contributions by chapter under the three research objectives.

The metrics used to quantify performance of the London Underground, and indeed most urban metro systems, tend to focus on the performance of network operations only using train movement data. To provide improved measurement of journey time performance from the passenger perspective, the first research objective is addressed in Chapter 4. Empirical journey time distributions are generated under different operating conditions, and the distributions are parametrically defined. From the moments of the distributions, three new metrics to capture journey time performance from the passenger perspective are proposed.

The second research objective is addressed in Chapter 5, where a probabilistic assignment algorithm is proposed to assign individual passenger trips to individual trains. Compared to existing assignment methods which require supplementary input of vast amounts of manually collected data, the probabilistic assignment method proposed in this thesis is based on inputs from automated data sources only, and is therefore more efficient in terms of computational time and cost for operators adopt in practice. The assignment method builds on the method proposed by Hörcher et al. (2017); a limitation of their method is that the assignment is based on the travel characteristics of the fastest passengers. This can therefore introduce a bias whereby passengers are allocated to favour train itineraries which produce the shortest egress times. The assignment algorithm presented in this thesis aims to correct for the potential bias through the use of egress times of passengers travelling in off-peak periods. A new method of assigning probabilities to each feasible train itinerary based on Bayes' Theorem is proposed, and statistical testing is undertaken to more rigorously verify that the assignment is non-biased.

Once passengers are assigned to trains, the total passenger journey times recorded by the automated fare collection (AFC) data are split into three constituent components: access, on-train, and egress times. The decomposition of journey times enables the third research objective to be addressed in Chapters 6 and 7. In the two chapters, semiparametric regression models are developed with each journey time component set as the response variable in the models. The models are used to derive elasticities of journey time with respect to different service supply and passenger demand factors, enabling a direct comparison of the relative degrees to which the different factors impact journey times.

The majority of existing regression-based studies of journey times apply linear regression methods, which are limited to representing linear relationships between the dependent and independent variables. In contrast, semiparametric regression methods are able to model flexible relationships generated from individual data points, resulting in parameter estimates with a higher degree of accuracy compared to the conventional linear models. The use of semiparametric regression methods in the analysis of journey time variance on urban metro systems has not been previously undertaken, and this is a key contribution of the thesis.

The regression models presented in Chapter 6 analyse the effects of network service supply and passenger demand factors on journey times, and covariates are included to capture properties of the stations, origin-destination routes, lines, and generic passenger attributes. The regression models in Chapter 7 focus on quantifying the effects of individual passenger-level heterogeneity on journey times. The analysis of the effects of individual passenger-specific characteristics via a semiparametric mixed model framework has not been previously undertaken, and this is the main contribution of the chapter.

The regression model of the access time component in Chapter 6 is further extended to investigate the impact of train headways on passenger wait times at the origin station. Marginal passenger wait times are estimated by computing the derivative of access time with respect to headways, and conclusions are drawn regarding passenger arrival patterns under different train operating frequencies. To current knowledge, this is the first application of advanced regression techniques to estimate the relationship between wait times and headways while conditioning for other service supply and passenger demand factors.

8.3 Summary of main findings

Beyond the initial introductory chapter, the first two chapters of the thesis, the literature review and data chapters, provide the context for the research presented in the subsequent analysis chapters. Four lines on the London Underground are analysed bidirectionally: the entire length of the Victoria and Waterloo and City lines, the central portion of the Central line from West Acton to Oxford Circus, and the central portion of the Jubilee line from Bond Street to North Greenwich. In this thesis, to avoid ambiguity arising from the inclusion of transfer movements and route choice, only single line trips that originate and terminate on the line of interest with no other probable routes under normal operating conditions are selected. The data cover weekday trips over a seven week period from October to December 2013.

The first thesis objective is addressed in Chapter 4; the properties of the passenger journey time distributions are parametrically characterised and quantified under regular and incident-affected conditions. Empirical journey time distributions on the four lines of the London Underground are generated for the entire operating day from 6am to 12am at 15 minute intervals. The form of the empirical distributions are parametrically defined. In line with the literature, journey time distributions under regular conditions are found to be predominately right-skewed with long tails. The Burr and log-logistic distributions are most representative, providing the best fit to approximately 86% of the distributions jointly. Under incident-affected conditions, approximately 35% of the distributions are found to adhere to a multimodal form, indicating the presence of mixed congested and non-congested travel conditions within a 15 minute time period.

Three new performance measures to assess the mean and variance of journey times as experienced by passengers are then proposed. The first indicator is a measure of the normalised journey times, while the second and third indicators provide measures of variance. The normalised reliability buffer time (RBT) indicates the degree to which journey times vary for a typical journey, and is taken as a measure of the 95th percentile journey time subtracted by the median journey times. The normalised standard deviation provides a measure of the total variance of journey times, and captures the extreme journey times that can arise from incident events and highly congested conditions. Each of the indicators are normalised by the free flow time, which is the journey time of a passenger travelling on a route in base uncongested conditions. The normalisation of the indicators allows comparisons to be made across routes and lines of differing lengths. The impact of incidents is assessed by computing the ratio of the value of the indicators under incident conditions to the value of the indicators under regular conditions.

The indicators are aggregated to compare the performance of journey times at a line level. Under regular conditions, mean journey times are found to be longer during the PM peak, compared to the AM and midday inter-peak periods across all lines. In terms of line performance, journey times are longest on the Waterloo and City line, with the mean journey time being on average 49% longer than the free flow time. Journey times on the Central line eastbound are quickest, with mean journey times 24% higher than the free flow time. In terms of variance, the RBT and standard deviation indicators show that journey times are generally more variable during the PM and midday inter-peak periods, depending on the line. Overall, journey times on the Waterloo and City line are most variable, and least variable on the Victoria line.

Under incident-affected conditions, mean journey times tend to be higher compared to regular conditions, ranging from 8% to 39% higher depending on the line. The variance indicators also generally report higher levels of journey time variance under incident-affected conditions when aggregated at a daily level. The RBT under incident-affected conditions ranges from 20% to 96% higher than regular conditions, depending on the line. The standard deviation of journey times also tends to be higher under incident-affected conditions, ranging from 7% to 79% higher depending on the line.

The second objective of the thesis is met via the development of a probabilistic assignment algorithm to assign individual passengers to individual trains, which is presented in Chapter

5. As a result of incomplete train movement data for the Waterloo and City line, only three lines, the Victoria, Central, and Jubilee lines are used in the analysis. The passenger trip data and train movement data are merged, and based on the timestamps in the two data sets, two assignment conditions arise. The first condition arises when the passenger trip can only be feasibly allocated to one train itinerary; this set of trips are termed unambiguous trips. The second condition captures trips that can be feasibly allocated more than one train itinerary; this set of trips are termed ambiguous trips. The ambiguous trips are allocated through a Bayesian assignment algorithm, which is based on the distribution of egress times for unambiguous trips in the off-peak. It is demonstrated that the use of unambiguous trips in the off-peak gives non-biased assignment results. Approximately 31% of trips in the data set are unambiguously allocated, and the remaining 69% of ambiguous trips are allocated using the Bayesian assignment algorithm.

Once all passengers have been allocated to trains, the total journey times from passenger entry at the origin station to passenger exit at the destination station can be decomposed into three constituent sub-components of total journey time. The components are defined as follows: 1) access time, which captures passenger walk times at the origin station from the entry point to the station platform, as well as the total waiting time at the platform prior to boarding, 2) on-train time, which represents the time taken for the train to travel between the origin and destination station platforms, and 3) egress time, which captures the walk time at the destination station from the platform to the exit point.

The final thesis objective is met through the application of semiparametric regression on the components of journey time in the final two analysis chapters of the thesis. In Chapter 6, the extent to which different service supply and demand factors influence journey times is quantified. Four regression models are developed; the components of journey time, namely, access, on-train, and egress time, as well as total journey time, are set as the response variables. The model covariates can be categorised as follows: 1) service supply characteristics, including dynamic operational characteristics and static physical characteristics of the stations, routes, and lines, and 2) passenger demand characteristics, including indicators of passenger volumes at different points of a journey, and general passenger attributes of age, card type, discount type, and trip frequency, as recorded by the Oyster system. Through semiparametric regression modelling, non-linear elasticity estimates of journey times are derived with respect to each covariate.

In all models, the static covariates capturing the effect of the stations, routes, and lines have the greatest effects. These variables capture the largely immutable physical characteristics of the system, and are therefore interpreted as representing the base constant characteristics of the network. The majority of the remaining variance in journey times is captured by dynamic operational factors.

After the static station, route, and line effects, train speed and train headways are the most significant covariates in the model of total journey times, with average elasticities of -0.54 and 0.05, respectively. For access times, headway is the second most influential covariate after the static station and line effects, with an average elasticity of 0.23. For on-train times, train speed is the second most significant covariate after the static route and line effects and has an average elasticity of -1.00. For egress times, the number of available exit points at the destination station platform is the second most influential set of covariates with an average elasticity of -0.16, after accounting for the static station and line effects.

As an extension of the results of the access time model, the influence of headways on passenger wait times at the origin station is able to be isolated and quantified. In line with the literature, the results show that passenger arrival patterns transition from random to non-random as headways increase. During peak times when train frequencies are higher, the transition from random to non-random occurs in the range of 2-3 minute headways. During off-peak periods where train frequencies are lower, the majority of passengers are found to arrive in a non-random way in order to minimise their wait times below a value of half of the headway.

In Chapter 7, the semiparametric regression models presented in Chapter 6 are extended to quantify to what extent individual-specific passenger heterogeneity influences the variance of journey times. A data set of regular commuters who undertake 30 or more repeat trips on the same origin-destination route over the seven week analysis period is extracted, and individual passengers are modelled as random effects in each journey time component model. A total of twelve models are developed; for each of the three components of journey time and total journey times, over the AM peak, inter-peak, and PM peak. Passenger effects are not found to influence on-train times. For the access, egress, and total journey time models, the passengers effects are significant, and the inclusion of the effects increases the goodness-of-fit of the models.

The passenger effects range from -0.87 log-mins (25.1 seconds) to 1.55 log-mins (282.2 seconds) for the access time models, -0.98 log-mins (22.6 seconds) to 0.62 log-mins (111.0 seconds) for the egress time models, and -0.23 log-mins (47.5 seconds) to 1.11 log-mins (182.3 seconds) for the total journey times models. The results show that the effects of individual passenger heterogeneity are most influential in the access time models, followed by the egress time models, and total journey time models. Across all models, the passenger effects have a lower degree of variance during the inter-peak compared to the peak, indicating that the influence of individual passenger heterogeneity is greater during the busier peak periods of travel.

As a by-product of the analysis, a ranking of stations in terms of their performance for passenger access and egress movements is derived. The station effects in these models capture the fixed physical characteristics of stations, such as station dimensions, layouts, and locations. The results show that access and egress times are lowest at North Greenwich (Jubilee line), Canary Wharf (Jubilee line) and White City (Central line) stations, while the access and egress times are longest at Euston (Victoria line), Warren Street (Victoria line), and Notting Hill Gate (Central line) stations.

In both Chapters 6 and 7, an additional case study of trips between Oxford Circus and Brixton on the Victoria line is presented to further illustrate how the results of the analyses can be interpreted and applied to infer the degree to which different service supply and demand factors, and individual passenger characteristics, affect journey time variance on a typical passenger route.

8.4 Potential applications

There are a number of potential applications of the research presented in the thesis, which can be broadly categorised into applications related to practical metro operations, and applications related to other transport research areas. The following sections highlight the applications of the results of the thesis by chapter.

In Chapter 4 of the thesis, three new passenger-oriented performance indicators are proposed: 1) a measure of mean journey times relative to uncongested free flow times, 2) a measure of the reliability of journey times through the normalised reliability buffer time indicator, and 3) a measure of the total variance of journey times. The indicators can be directly applied by operators to improve the measurement of journey time performance on their networks, as experienced by their customers. The indicators can be aggregated at different levels of temporal and spatial extents, from an individual origin-destination route level, to a wider line and network level, at different times of the day. The indicators can therefore be tailored to provide detailed information on the journey time performance of the system under different levels of aggregation as required. The results can be used to benchmark passenger journey time performance between lines, and between other urban rail operators worldwide.

The three proposed indicators are also evaluated for trips under incident conditions. The ratio of the values of each indicator under incident conditions versus regular conditions is used to quantify the effect of incidents in terms of the impact on the performance of journey times. This can be applied by operators to improve the assessment of the effects of incident events, and to improve response management under incident conditions.

In addition to being adopted by operators for performance reporting purposes, the indicators in Chapter 4 can also be used to provide improved real-time information for passengers. The metrics can be incorporated within the existing TfL Journey Planner tool to give real-time estimates of journey times and journey time variance on lines, to better inform passengers of travel conditions. This extra information would provide improved ability for passengers to plan their journeys, thereby improving the disutility of journey times, making urban metro systems a more attractive choice for travellers. Along with the passenger and operator oriented applications, the results can also be used in a broader sense to improve estimates of journey time variance in behavioural passenger utility models for metro travel, specifically in route-choice and departure time choice models.

The regression analyses in Chapters 6 and 7 assess the degree to which different service supply and demand factors influence each component of passenger journey time. Elasticities of journey times with respect to each supply and demand factor are derived, and these results can be used to inform operators of which operational and/or physical characteristics of the system should be targeted to achieve improvements in passenger journey times. Along with informing operators, the results can also be more widely applied to improve the valuation of operational and/or physical interventions, as well as the economic appraisal of new transport projects.

8.5 Future work

A number of potential areas of future research based on the analyses presented in this thesis have been identified, and are summarised in the following sections.

In this thesis, the empirical data used to demonstrate the application of the methods represents only single line trips that originate and terminate on the line of interest, with no other probable routes under normal operating conditions. To include all trips on the London Underground, the passenger flows for all possible origin-destination pairs within the network are required. To obtain this information, additional probabilistic methods to determine passenger route-choice must be applied. As the focus of the thesis is to demonstrate the methods used to quantify journey time variance, the additional step of incorporating route-choice is omitted. Therefore, a potential area for future research is the extension of all methods presented in this thesis to include all trips within a metro network via the application of route-choice models. Furthermore, as mentioned, the methods are restricted to the analysis of single line trips only, thus omitting the consideration of multi-line trips involving interchange movements when a passenger transfers from one line to another. Therefore, extension of the methods to incorporate passenger interchange movements for multi-line trips is also recommended.

Another area for future research is the treatment of the level of timestamp accuracy for the Oyster passenger trip records. As mentioned in Chapter 3, the Oyster data are accurate to a level of one minute only, and as a result, the passenger to train assignment and regression results are impacted. The train movement data are accurate to a level of one second, and so in Chapter 5 a conservative assumption is applied that the passenger could board the train at any time within the minute timestamp level of accuracy for the passenger trip, thereby increasing the number of train itineraries that a passenger could feasibly board. In the regression analyses in Chapters 6 and 7, an additional random error of magnitude 0-59 seconds for the access and egress time models and 0-118 seconds in the total journey time models is present as a result of the passenger trip timestamp truncation. This random error is incorporated within the random error regression term in the models. If higher quality data with a higher level of timestamp accuracy is able to be obtained, the accuracy of the passenger to train assignment could be improved and a higher level of precision could be obtained for the parameter estimates in the regression models.

Another limitation highlighted in Chapter 3 regarding the accuracy of the Oyster data timestamps is potential synchronisation issues between the card readers at stations across the network. As mentioned, the recorded tap-in and tap-out timestamps may be inaccurate, however, no information is available on the magnitude or location of the errors, and so systematic modelling of the error effect is not possible. In Chapters 6 and 7 any potential synchronisation errors are currently captured in the random error term of the regression models. Further liaison with TfL is recommended to establish the magnitude and location of the errors, so that the issue can either be resolved completely or so that the errors can be explicitly modelled in a systematic way.

In Chapter 4, empirical journey time distributions are defined under regular and incident-affected conditions. The analysis of incident-affected conditions is limited to quantifying the effect on journey times up to 15 minutes after the beginning of the incident event. The spatial extents are defined as being all origin-destination pairs downstream of the origin location of the incident, on the same line. Depending on the degree of incident severity, the impact on journey times could encompass a wider range of temporal and spatial extents. Potential areas of future research are therefore assessing the temporal and spatial propagation of incidents, specifically: quantifying the time taken for journey times to converge back to levels consistent with regular operating conditions, and defining the geographical extents to which incident events propagate through the network.

In Chapter 5, a Bayesian algorithm is proposed to assign passengers to trains. As mentioned in the chapter, the assignment is influenced by data quality issues, namely the absence of train arrival timestamps in the majority of the train movement data set. Only a minority of records in the train movement data set have fully recorded train arrival and departure timestamps at each station platform. As a result, to perform the assignment on the three lines within the defined study area, only the departure timestamps are used to perform the assignment. The addition of the train arrival timestamps could potentially increase the accuracy of determining whether a train itinerary is feasible, as the boundary conditions could be updated such that an itinerary is considered feasible if the passenger taps-in prior to both the train arrival and departure timestamps at the origin station platform, and taps-out after the train arrival and departure timestamps at the destination platform. Therefore, a potential area for future work would be to undertake the assignment on the smaller subset of data with both arrival and departure timestamps to determine whether a more accurate assignment can be produced with more complete data.

Another area for future work outlined in Chapter 5 is testing the accuracy of the proposed assignment algorithm on simulated data. A verification is undertaken comparing the proposed assignment algorithm with the algorithm developed by Hörcher et al. (2017), however, this is carried out on a sub-sample of Oyster passenger trip data and train movement data. Consequently, this exercise provides a comparative assessment between the two methods only, as the true train itinerary that the passenger boarded is not known. The development of high quality simulated data that is able to appropriately reflect the travel conditions that a passenger is likely to experience is outside the scope of this research, however, this is recommended as a direction for future work. Testing the proposed algorithm on simulated data would enable an additional assessment of the degree of accuracy of the method, relative to known “true” train itineraries.

In Chapters 6 and 7, semiparametric regression is performed to quantify the effect of different service supply, demand, and individual passenger characteristics on journey times. The selection of covariates in the models is limited to the data available. The goodness-of-fit results for the access and egress time models in particular show that there is scope to incorporate additional covariates to capture the residual variation in journey times. In terms of network demand characteristics, recommendations of additional covariates include covariates to capture passenger volumes at different points in the system, such as on trains and at pinch points where passenger flow is most congested. From a network service supply point of view, more detailed information can also be included on station layouts to obtain more accurate estimates of individual passenger walk distances within the origin and destination stations, which could vastly improve the explanatory power of the models. Further information on the operational aspects of the network, such as signalling schemes and rollingstock properties, could also be included.

In terms of the analysis of individual characteristics in Chapter 7, the models focus on the behaviour of regular passengers who undertake 30 or more trips over a seven week time period of analysis. The analysis could be extended to include individual-specific effects across all groups of passengers. Passengers could be potentially organised into different categories of trip frequency ranging from a minimum of 5 to 10 trips upwards to obtain a more comprehensive set of results across all passenger demographics.

In terms of spatial extents, the analysis of individual passenger effects is undertaken over different origin-destination routes on different lines. A potential area of further research could be to construct separate models for each origin and destination station for individual origin-destination routes. The resulting station and route-specific distribution of individual passenger effects could then be potentially used supplementary to or alternatively to the distribution of passenger egress times, to improve the assignment of passengers to trains as presented in Chapter 5.

As a final remark, the methods developed in the thesis apply to the analysis of journey time performance of a singular urban metro system. There is scope for all methods to be extended to enable comparisons across different metro systems. The results could then be used to undertake benchmarking, to identify best practices in metro journey time performance at a network level.

References

Abkowitz, M., Slavin, H., Waksman, R., Englisher, L., Wilson, N., 1978. Transit service reliability: final report. Tech. Rep. UMTA MA-06-0049-78-1.

Albrecht, T., B. A. . G. C., 2013. Applications of real-time speed control in rail-bound public transportation systems. IET Intelligent Transport Systems 7 (3), 305–314.

Anderson, R., Condry, B., Findlay, N., Brage-Ardao, R., Li, H., 2013. Measuring and valuing convenience and service quality: A review of global practices and challenges from mass transit operators and railway industries. International Transport Forum Discussion Paper No. 2013-16, prepared for OECD Roundtable on Valuing Convenience in Public Transport, Paris, September 2013.

Barkley, T., Hranac, R., Petty, K., 2012. Relating travel time reliability and nonrecurrent congestion with multistate models. Transportation Research Record 2278, 13–20.

Barron, A., Canavan, S., Anderson, A., Cohen, J., 2018. Operational impacts of platform doors in metros. Transportation Research Record 2672 (8), 266–274.

Barron, A., Melo, P. C., Cohen, J. M., Anderson, R. J., 2013. Passenger-focused management approach to measurement of train delay impacts. Transportation Research Record 2351, 46–53.

Bates, J., Polak, J., Jones, P., Cook, A., 2001. The valuation of reliability for personal travel. Transportation Research Part E 37, 191–229.

Benaglia, T., Chauveau, D., Hunter, D., Young, D., 2009. mixtools: An R package for analysing finite mixture models. Journal of Statistical Software 32 (6), 1–29.

Berry, D., January 1952. Evaluation of techniques determining overall travel time. In: Highway Research Board Proceeding, 31st Annual Meeting, Washington D. C.

Bertini, R. L., El-Geneidy, A. M., 2004. Modelling transit trip time using archived bus dispatch system data. Journal of Transportation Engineering 130 (1), 56–67.

Bowman, L. A., Turnquist, M. A., 1981. Service frequency, schedule reliability and passenger wait times at transit stops. Transportation Research Part A: General 15, 465–471.

Buchel, B., Corman, F., 2018. Modelling probability distributions of public transport travel time components. In: 18th Swiss Transport Research Conference.

Button, K., 2010. Transport economics, 3rd Edition. Edward Elgar Publishing Ltd, Cheltenham, UK.

Cambridge Systematics Inc., Texas Transport Institute, 2005. Traffic congestion and reliability: Trends and advanced strategies for congestion mitigation. Tech. rep., accessed at ops.fhwa.dot.gov.

Cameron, A. C., Trivedi, P. K., 2005. Microeconometrics: Methods and applications. Cambridge University Press, Cambridge, UK.

Casella, G., Berger, R., 2002. Statistical inference, 2nd Edition. Duxbury, Pacific Grove, CA.

Cepeda, M., Cominetti, R., Florian, M., 2006. A frequency-based assignment model for congested transit networks with strict capacity constraints: characterisation and computation of equilibria. Transportation Research Part B 40, 437–459.

Chan, J., 2007. Rail transit OD matrix estimation and journey time reliability metrics using automated fare data. Master's thesis, Massachusetts Institute of Technology.

Christoforou, Z., Chandakas, E., Kaparias, I., 2017. An investigation of dwell time patterns in urban public transport systems: the case of the Nantes tramway. In: 8th International Congress on Transportation Research in Greece.

Csikos, D., Currie, G., 2007. The impacts of transit reliability on wait time: insights from automated fare collection system data. In: 86th Annual meeting of the Transportation Research Record.

Csikos, D., Currie, G., 2008. Investigating consistency in transit passenger arrivals: insights from longitudinal automated fare collection data. Transportation Research Record 2042, 12–19.

Dandy, G., McBean, E., 1984. Variability of individual travel time components. Journal of Transportation Engineering 110 (3), 340–356.

de Ana Rodrguez, G., Seriani, S., Holloway, C., 2016. Impact of platform edge doors on passengers boarding and alighting time and platform behaviour. Transportation Research Record 2540, 102–110.

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

Diab, E., El-Geneidy, A., 2013. Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability. Public Transport 4 (3), 209–231.

Dougherty, C., 2016. Introduction to econometrics, 5th Edition. Oxford University Press, Oxford, UK.

Douglas Economics, 2012. Modelling train station demand and capacity. Tech. Rep. Final report for Transport for NSW - for distribution.

Duran-Hormazabal, E., Tirachini, A., 2016. Estimation of travel time variability for cars, buses, metro and door-to-door public transport trips in Santiago, Chile. Transportation Economics 59, 26–39.

Dutang, C., Goulet, V., Pigeon, M., 2008. actuar: An R package for actuarial science. Journal of Statistical Software 25 (7), 1–37.

El-Geneidy, A., Horning, J., Krizek, K., 2011. Analyzing transit service reliability using detailed data from automatic vehicular locator systems. Journal of Advanced Transportation 45 (1), 66–79.

El-Geneidy, A., Vijayakumar, N., 2011. The effects of articulated buses on dwell and running times. Journal of Public Transportation 14 (3), 63–86.

El-Geneidy, A. M., Strathman, J. G., Kimpel, T. J., Crout, D. T., 2006. Effects of bus stop consolidation on passenger activity and transit operations. Transportation Research Record 1971, 32–41.

Fan, W., Machemehl, R. B., 2009. Do transit users just wait for buses or wait with strategies? Some numerical results that transit planners should see. Transportation Research Record 2111, 169–176.

Fosgereau, M., Fukuda, D., 2012. Valuing travel time variability: characteristics of the travel time distribution on an urban road. Transportation Research Part C 24, 83–101.

Freemark, Y., 2013. Assessing journey time impacts of disruptions on London's Piccadilly line. Master's thesis, Massachusetts Institute of Technology.

Frumin, M., Zhao, J., 2012. Analysing passenger incidence behaviour in heterogeneous transit services using smart card data and schedule based assignment. Transportation Research Record 2274, 52–60.

Furth, P. G., Hemily, B., Muller, T. H. J., Strathman, J., 2006. Using archived AVL-APC data to improve transit performance and management. Tech. Rep. TCRP Report 113.

Furth, P. G., Muller, T. H. J., 2006. Service reliability and hidden waiting time: insights from AVL data. Transportation Research Record 1955, 79–87.

Gibson, S., Cooper, G., Ball, B., 2002. Developments in transport policy: the evolution of capacity charges on the UK rail network. Journal of transport economics and policy 36 (2), 341–354.

Glover, J., 2013. Principles of railway operation. Ian Allen Publishing Ltd, Surrey, UK.

Goldberg, G., 2018. Data-driven modelling of passenger response to disruption on the London Underground. Ph.D. thesis, Imperial College London.

Gong, H., Chen, X., Yu, L., Wu, L., 2016. An application-oriented model of passenger waiting time based on bus departure time intervals. Transportation Planning and Technology 39 (4), 424–437.

Goodman, C. J., S. L. K. . H. T. K., April 1998. Review of simulation models for railway systems. In: IEE International Conference on Developments in Mass Transit Systems. No. 543. pp. 80–85.

Gorman, M., 2009. Statistical estimation of railroad congestion delay. Transportation Research Part E 45, 446–456.

Greene, W. H., 2018. Econometric analysis, 8th Edition. Pearson, New York, NY.

Guessous, Y., Aron, M., Bhouri, N., Cohen, S., 2014. Estimating travel time distribution under different traffic conditions. In: 17th meeting of the EURO working group on Transportation, EWGT2014. No. 3. Transportation Research Procedia, pp. 339–348.

Guo, F., Li, Q., Rakha, H., 2012. Multistate travel time reliability models with skewed component distributions. Transportation Research Record 2315, 47–53.

Guo, S., Yu, L., Chen, X., Zhang, Y., 2011. Modelling waiting time for passengers transferring from rail to buses. Transportation Planning and Technology 34 (8), 795–809.

Hamdouch, Y., Lawphongpanich, S., 2008. Schedule-based transit assignment model with travel strategies and capacity constraints. Transportation Research Part B 42, 663–684.

Han, J., 2012. Multi-sensor data fusion for travel time estimation. Ph.D. thesis, Imperial College London.

Hansen, I. A., Goverde, R. M. P., van der Meer, D. J., 2010. Online train delay recognition and running time prediction. In: Proceedings from the 13th international IEEE annual conference on intelligent transport systems.

Harris, N. G., 2006. Train boarding and alighting rates at high passenger loads. Journal of Advanced Transportation 40 (3), 249–263.

Harris, N. G., Anderson, R. J., 2007. An international comparison of urban rail boarding and alighting rates. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit 221, 521–526.

Harris, N. G., Graham, D. J., Anderson, R. J., Li, H., 2014. The impact of urban rail boarding and alighting factors. In: 93rd annual meeting of the Transportation Research Board.

Hartigan, J., Hartigan, P., 1985. The dip test of unimodality. Annals of statistics 13, 70–84.

Hartigan, P., 1985. Algorithm AS217: Computation of the dip statistic to test for unimodality. Applied statistics 34, 320–325.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning, 2nd Edition. Springer, Dordrecht.

Heinz, W., 2003. Passenger service times on trains. Ph.D. thesis, KTH University of Stockholm.

Henderson, G., Kwong, P., Adkins, H., 1991. Regularity indices for evaluating transit performance. Transportation Research Record 1297, 3–9.

Herman, R., Lam, T., August 1974. Trip characteristics of journeys to and from work. In: Proceedings of the 6th International Symposium on Transportation and Traffic Theory, Sydney. pp. 57–86.

Holloway, C., Thoreau, R., Roan, T., Boampong, D., Clarke, T., Watts, D., Tyler, N., 2016. Effect of vertical step height on boarding and alighting time of train passengers. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit 230 (4), 1234–1241.

Holroyd, E. M., Scraggs, D. A., 1966. Waiting time for buses in Central London. Traffic Engineering and Control 8, 156–160.

Hong, S., Min, Y., Park, M., Kim, K. M., Oh, S. M., 2016. Precise estimation of connections of metro passengers from smart card data. Transportation 43, 749–769.

Hörcher, D., Graham, D. J., Anderson, R. J., 2017. Crowding cost estimation with large scale smart card and vehicle location data. Transportation Research Part B 95, 105–125.

Howlett, P., 2000. The optimal control of a train. Annals of operations research 98, 65–87.

Ingvardson, J. B., Nielsen, O. A., Raveau, S., Nielsen, B. F., 2018. Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: a smart card data analysis. Transportation Research Part C: Emerging Technologies 90, 292–306.

Jolliffe, J. K., Hutchinson, T. P., 1975. A behavioural explanation of the association between bus and passenger arrivals at bus stops. Transportation Science 9, 248–282.

Jun, J., 2010. Understanding the variability of speed distributions under mixed traffic conditions caused by holiday traffic. Transportation Research Part C 18, 599–610.

Karekla, X., Tyler, N., 2012. Reduced dwell times resulting from train-platform improvements: the costs and benefits of improving passenger accessibility to metro trains. Transportation Planning and Technology 35 (5), 525–543.

Kecman, P., Goverde, R. M. P., 2015. Predictive modelling of running and dwell times in railway traffic. Public Transport 7, 295–319.

Khmelnitsky, E., 2000. On an optimal control problem of train operation. IEEE Transactions on Automatic Control 45 (7), 1257–1266.

Kieu, L. M., Bhaskar, A., Chung, E., 2014. Establishing definitions and modelling public transport travel time variability. In: 93rd annual meeting of the Transportation Research Board.

Kim, D. N. Schonfeld, P. M., 1997. Benefits of dipped vertical alignments for rail transit routes. ASCE Journal of Transportation Engineering 123 (1), 20–27.

Kim, K. Chien, S., 2010. Simulation-based analysis of train controls under various track alignments. ASCE Journal of Transportation Engineering 136 (11), 937–948.

Kimpel, T. J., Strathman, J., Bertini, R. L., Callas, S., 2005. Analysis of transit signal priority using archived TriMet bus dispatch system data. Transportation Research Record 1925, 156–166.

Kimpel, T. J., Strathman, J. G., Dueker, K., 2000. Time point-level analysis of passenger demand and transit service reliability. Tech. rep., accessed online at tfl.gov.uk/corporate/publications-and-reports/london-connections-map.

Kittelson & Associates Inc, Parsons Brinckerhoff, KFH Group, Texas A&M Transportation Institute, Arup, 2013. TCRP Report 165: Transit Capacity and Quality of Service Manual, 3rd Edition. Transit Cooperative Research Program, Transportation Research Board, Washington, D.C.

Kittelson & Associates Inc, Urbitran Inc, LKC Consulting Services Inc, MORPACE International Inc, Queensland University of Technology, Nakanishi, Y., 2002. TCRP Report 88: A guidebook for developing a transit performance-measurement system. Transit Cooperative Research Program, Transportation Research Board, Washington D.C.

Krygsman, S., Dijst, M., Arentze, T., 2004. Multimodal public transport: an analysis of travel time elements and the interconnectivity ratio. Transport Policy 11, 265–275.

Kusakabe, T., Iryo, T., Asakura, Y., 2010. Estimation method for railway passengers' train choice behaviour with smart card transaction data. Transportation 37, 731–749.

Lam, W. H. K., Cheung, C. Y., Poon, Y. F., 1998. A study of train dwelling time at the Hong Kong mass transit railway system. Journal of advanced transportation 32 (3), 285–296.

Lehmann, E., Romano, J., 2005. Testing statistical hypotheses, 3rd Edition. Springer, New York, NY.

Leshem, G., Ritov, Y., 2007. Traffic flow prediction using Adaboost algorithm with random forests as a weak learner. World Academy of Science, Engineering and Technology International Journal of Mathematical and Computational Sciences 1 (1), 1–6.

Li, M., Liu, Z., Zhang, Y., Liu, W., Shi, F., 2017. Distribution analysis of train interval journey time employing the censored model with shifting character. Journal of applied statistics 44 (4), 715–733.

Li, R., Rose, G., Sarvi, M., 2006. Using automatic vehicle identification data to gain insight into travel time variability and its causes. Transportation Research Record 1945, 24–32.

Lin, T., Wilson, N. H. M., 1992. Dwell time relationships for light rail systems. Transportation Research Record 1361, 287–295.

Liu, R. Golovitcher, I. M., 2003. Energy-efficient operation of rail vehicles. Transportation Research Part A: Policy and Practice 37 (10), 917–932.

Liu, L., Chen, R., 2017. A novel passenger flow prediction model using deep learning methods. Transportation Research Part C: Emerging Technologies 84, 74–91.

Lomax, T., Schrank, D., Turner, S., Margiotta, R., 2003. Selecting travel time reliability measures. Tech. Rep. TTI-2003-3, accessed at tti.tamu.edu/documents.

Luethi, M., Weidmann, U. A., Nash, A., 2007. Passenger arrival rates at public transport stations. In: 86th Annual meeting of the Transportation Research Board.

Lukaszewicz, P., 2001. Energy consumption and running time for trains. Ph.D. thesis, KTH Stockholm.

Luthi, M., 2009. Improving the efficiency of heavily used railway networks through integrated real-time rescheduling. Ph.D. thesis, ETH Zurich.

Ma, Z., Ferreira, L., Mesbah, M., Hojati, A. H., 2015. Modelling bus travel time reliability with supply and demand data from automatic vehicle location and smart card systems. Transportation Research Record: Journal of the Transportation Research Board 2533, 17–27.

Ma, Z., Ferreira, L., Mesbah, M., Zhu, S., 2016. Modelling distributions of travel time variability for bus operations. Journal of advanced transportation 50, 6–24.

Maechler, M., 2016. Hartigans' dip test statistic for unimodality. Tech. Rep. 0.75-7.

Martchouk, M., Mannering, F., Bullock, D., 2011. Analysis of freeway travel time variability using bluetooth detection. Journal of Transportation Engineering 137 (10), 697–704.

Matloff, N., 2017. From linear models to machine learning: a modern view of statistical regression and classification. CRC Press, Boca Raton, F.L.

Mazloumi, E., Currie, G., Rose, G., 2008. Causes of travel time unreliability - a Melbourne case study. In: 2008 ATRF 31st Australasian Transport Research Forum. pp. 195–202.

Mazloumi, E., Currie, G., Rose, G., 2010. Using GPS data to gain insight into public transport travel time variability. Journal of Transportation Engineering 136 (7), 623–631.

Murali, P., Dessouky, M., Ordonez, F., Palmer, K., 2010. A delay estimation technique for single and double-track railroads. Transportation Research Part E 46, 483–495.

NCHRP, U., 1997. Quantifying congestion: Final Report. Tech. Rep. Report 398, accessed at onlinepubs.trb.org.

Nguyen, S., Pallottino, S., 1988. Equilibrium traffic assignment for large scale transit networks. European Journal of Operational Research 37, 176–186.

Noland, R. B., Polak, J. W., 2002. Travel time variability: a review of theoretical and empirical issues. Transportation Reviews 22 (1), 39–54.

Noland, R. B., Small, K. A., 1995. Travel-time uncertainty, departure time choice, and the cost of morning commutes. Transportation Research Record: Journal of the Transportation Research Board 1493, 150–158.

Nuzzolo, A., Russo, F., Crisalli, U., 2001. A doubly dynamic schedule-based assignment model for transit networks. Transportation Science 35 (3), 268–285.

Nygaard, M. F., Torset, T., 2016. Waiting time strategy for public transport passengers. In: Proceedings from the annual transport conference at Aalborg University.

O'Flaherty, C. A., Mangan, D. O., 1970. Bus passenger waiting times in central areas. Traffic Engineering and Control 11 (9), 419–421.

Oh, S., Kim, S., Hong, J., 2016. Comprehensive analysis of the influence of door width on the passenger flow time on Korean urban railways. Journal of Rail and Rapid Transit 230 (6), 1531–1539.

Osana, E. E., Newell, G. F., 1972. Control strategies for an idealised public transport system. Transportation Science 6, 52–72.

Parkinson, T., Fisher, I., 1996. TCRP Report 13: Rail Transit Capacity. Transit Cooperative Research Program, Transportation Research Board, Washington, D.C.

Paul, E. C., 2010. Estimating train passenger load from automated data systems: application to London Underground. Master's thesis, Massachusetts Institute of Technology.

Polus, A., 1979. A study of travel time and reliability on arterial routes. Transportation 8, 141–151.

Poon, M. H., Wong, S. C., Tong, C. O., 2004. A dynamic schedule-based model for congested transit networks. Transportation Research Part B 38, 343–368.

Preston, J., Wall, G., Batley, R. J., Ibanez, N., Shires, J., 2009. Impact of delays on passenger train services: evidence from Great Britain. Transportation Research Record: Journal of the Transportation Research Board 2117, 1423.

Profillidis, V. A., 2014. Railway management and engineering, 4th Edition. Ashgate, Farnham, UK.

Puong, A., 2000. Dwell time model and analysis for the MBTA Red Line. Tech. Rep. Working paper.

Rahman, M. M., Wirasinghe, S. C., Kattan, L., 2018. Analysis of bus travel time distributions for varying horizons and real-time applications. Transportation Research Part C 86, 453–466.

Richardson, A., Taylor, M., 1978. Travel time variability on commuter journeys. High Speed Ground Transportation 12, 77–99.

Schil, M., 2012. Journey time reliability in London using automated data collection systems. Master's thesis, Massachusetts Institute of Technology.

Schmocker, J., Bell, M. G. H., Kurauchi, F., 2008. A quasi-dynamic capacity constrained frequency-based transit assignment model. Transportation Research Part B 42, 925–945.

Scott, D. W., 2015. Multivariate Density Estimation: Theory, Practice, and Visualization, 2nd Edition. John Wiley & Sons, Inc., Hoboken, NJ.

Searle, S. R., Casella, G., McCulloch, C. E., 1992. Variance components. John Wiley & Sons, Inc., Hoboken, New Jersey.

Seddon, P. A., Day, M. P., 1974. Bus passenger waiting times in Greater Manchester. Traffic Engineering and Control 15, 442–445.

Seriani, S., Fernandez, R., 2015. Pedestrian traffic management of boarding and alighting in metro stations. Transportation Research Part C 53, 76–92.

Spiess, H., Florian, M., 1989. Optimal strategies: a new assignment model for transit networks. Transportation Research Part B 23B (2), 83–102.

Strathman, J., Dueker, K. J., Kimpel, T., Gerhart, R., Turner, K., Taylor, P., Callas, S., Griffin, D., Hopper, J., 1999. Automated bus dispatching, operations, control, and service reliability: baseline analysis. Transportation Research Record 1666, 28–36.

Strathman, J. G., Kimpel, T. J., Dueker, K. J., Gerhart, R. L., Callas, S., 2002. Evaluation of transit operations: data applications of Tri-Met's automated bus dispatching system. Transportation 29, 321–345.

Sun, L., Lee, D., Erath, A., Huang, X., August 2012. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing. ACM, pp. 142–148.

Sun, Y., Schonfeld, P., 2015. Schedule-based rail transit path-choice estimation using automatic fare collection data. Journal of Transportation Engineering 142 (1), 1–8.

Surprenant-Legault, J., El-Geneidy, A. M., 2011. Introduction of reserved bus lane: impact on bus running time and on-time performance. Transportation Research Record 2218, 10–18.

Susilawati, S., Taylor, M., Somenahalli, S., 2013. Distributions of travel time variability on urban roads. Journal of Advanced Transportation 47, 720–736.

Taylor, M., 1982. Travel time variability - the case of two public modes. Transportation Science 16 (4), 507–521.

Taylor, M., 1999. Dense network traffic models, travel time reliability and traffic management ii: Application to network reliability. Journal of advanced transportation 33 (2), 235–251.

Taylor, M., Susilawati, S., 2012. Modelling travel time reliability with the Burr distribution. Procedia - Social and Behavioural Sciences 54, 75–83.

Taylor, M. A. P., 2013. Travel through time: the story of research on travel time reliability. Transportmetrica B: Transport Dynamics 1 (3), 174–194.

Tetreault, P. R., El-Geneidy, A. M., 2010. Estimating bus run times for new limited-stop service using archived AVL and APC data. Transportation Research Part A 44, 390–402.

Thoreau, R., Holloway, C., Bansal, G., Gharatya, K., Roan, T., Tyler, N., 2016. Train design features affecting boarding and alighting of passengers. Journal of Advanced Transportation 50, 2077–2088.

Transport for London, 1999. Journey Time Metric (JTM): An overview. Tech. Rep. M(98)54, accessed at tfl.gov.uk.

Transport for London, 2013. Business Case Development Manual. Tech. Rep. V101.2013.05.

Transport for London, 2014a. Geographically accurate representation of London Tube and rail lines. Tech. rep., accessed online at tfl.gov.uk/corporate/publications-and-reports/london-connections-map.

Transport for London, 2014b. NACHS 2014 Explanatory Notes. Tech. Rep. Parts A and B, accessed at tfl.gov.uk.

Transport for London, 2015. Rail and Underground International Benchmarking Report. Tech. Rep. Item 8, accessed at tfl.gov.uk.

Transport for London, 2016. The Journey Time Metric (JTM). Tech. rep., accessed at tfl.gov.uk.

Transport for London, 2017. London Underground map, Version J - zonal map. Tech. rep., accessed online at tfl.gov.uk/info-for/suppliers-and-contractors/map-sizes-and-formats.

Transport for London, 2018a. Transport for London annual report and statement of accounts 2017/18. Tech. rep., accessed online at tfl.gov.uk/corporate/publications-and-reports/annual-report.

Transport for London, 2018b. Travel in London - Report 11. Tech. rep., accessed online at content.tfl.gov.uk/travel-in-london-report-11.pdf.

Transport for London, 2019a. London Underground performance reports. Tech. rep., accessed online at data.london.gov.uk/dataset/london-underground-performance-reports.

Transport for London, 2019b. London's transport - a history. Tech. rep., accessed online at tfl.gov.uk/corporate/about-tfl/culture-and-heritage/londons-transport-a-history/.

Transportation Research Board Committee on Highway Capacity and Quality of Service, 2010. Highway Capacity Manual, 5th Edition. Transportation Research Board, National Research Council, Washington, D.C.

Turnquist, M., 1978. A model for investigating the effects of service frequency and reliability on bus passenger wait times. Transportation Research Record 663, 70–73.

Uniman, D. L., 2009. Service reliability measurement framework using smart card data: application to the London Underground. Master's thesis, Massachusetts Institute of Technology.

Uniman, D. L., Attanuci, J., Mishalani, R. G., Wilson, N. H. M., 2010. Service reliability measurement using automated fare card data: application to the London Underground. Transportation Research Record 2143, 92–99.

van den Heuvel, J., 2016. Field experiments with train stopping positions at Schiphol airport train station in Amsterdam, Netherlands. Transportation Research Record 2546, 24–32.

Van Lint, J., Van Zuylen, H., 2005. Monitoring and predicting freeway travel time reliability. Transportation Research Record 1917, 54–62.

Van Lint, J., Van Zuylen, H., Tu, H., 2008. Travel time unreliability on freeways: Why measures based on variance tell only half the story. Transportation Research Part A 42, 258–277.

van Loon, R., Rietveld, P., Brons, M., 2011. Travel-time reliability impacts on railway passenger demand: a revealed preference analysis. Journal of Transport Geography 19 (4), 917–925.

Verma, S., 2010. Contracting for Ticketing Services. Tech. rep., prepared for The Rudin Center, New York University, Leveraging technology for transportation.

Wardman, M. R., 2004. Public transport values of time. Transport Policy 11 (4), 363–377.

Wardrop, J., 1952. Some theoretical aspects of road traffic research. Proceedings of the Institution of Civil Engineers Part 1 1 (3), 325–362.

Wei, Y., Chen, M. C., 2012. Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. Transportation Research Part C: Emerging Technologies 21 (1), 148–162.

Wen, B. W., 2017. Data-driven mesoscopic simulation of large-scale surface transit networks. Master's thesis, University of Toronto.

Weston, J. G., 1989. Train service model - technical guide. Tech. Rep. Operational research note 89/18.

Wiggenraad, P. B. L., 2001. Alighting and boarding time of passengers at Dutch Railway Stations. Tech. rep.

Wilson, N., Nelson, D., Palmere, A., Grayson, T., Cederquist, C., 1992. Service-quality monitoring for high frequency transit lines. Transportation Research Record 1349, 3–11.

Wong, H., Sussman, J., 1973. Dynamic travel time estimation on highway networks. Transportation Research 7, 355–370.

Wonnacott, T., Wonnacott, R., 1990. Introductory statistics. Wiley, London.

Wood, D. A., 2015. A framework for measuring passenger-experienced transit reliability using automated data. Master's thesis, Massachusetts Institute of Technology.

Wood, S. N., 2006. Generalized additive models: an introduction with R. Chapman & Hall, Boca Raton, FL.

Wood, S. N., Goude, Y., Shaw, S., 2015. Generalised additive models for large data sets. Journal of the Royal Statistical Society: Series C Applied Statistics 64 (1), 139–155.

Yetiskul, E., Senbil, M., 2012. Public bus transit travel-time variability in Ankara (Turkey). Transport Policy 23, 50–59.

Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X., Li, Z., 2017. Estimation of passenger route choice pattern using smart card data for complex metro systems. IEEE Transactions on Intelligent Transport Systems 18 (4), 790–801.

Zhu, Y., 2014. Passenger-to-train assignment model based on automated data. Master's thesis, Massachusetts Institute of Technology.

Zhu, Y., Koutsopoulos, H., Wilson, N., 2017. A probabilistic passenger-to-train assignment model based on automated data. Transportation Research Part B: Methodological 104, 522–542.

Appendix A

Regression studies on bus journey times

The following table provides a summary of the regression studies undertaken for the analysis of bus journey times, as referenced in the literature review in Chapter 2.

Table A.1: Regression studies on bus journey times

Location and source	Model	Response variables	Significant explanatory variables
Bus, Brisbane, Australia (Ma et al., 2015)	Linear regression - SURE	1) Mean travel time 2) Buffer time: 95% − 50% *JT* 3) CV	1) Route length, delay at start of route, stops, boardings, alightings, no. signals, recurrent congestion index (i.e. ratio of mode to free flow speeds), weather, road type, dummy for CBD area 2 & 3) Route length, SD delay at start of route, stops, SD boardings, SD alightings, no. signals, recurrent congestion index (i.e. ratio of mode to free flow speeds), road type, dummy for CBD area, weather (only 3)
Bus, Montreal, Canada (Diab and El-Geneidy, 2013)	Linear regression - OLS	1) Journey time 2) Journey time deviation 3) CV 4) CV JT deviation	1 & 2) Boardings, alightings, direction, stops, time of day, start delay, route ID, bus type, weather, smart card usage, dummy for use of reserved bus lane, dummy for transit priority signalling 3 & 4) CV Boardings, CV alightings, direction, stops, time of day, route ID, bus type, weather, smart card usage, dummy for use of reserved bus lane, dummy for transit priority signalling

Table A.1: Regression studies on bus journey times

Location and source	Model	Response variables	Significant explanatory variables
Bus, Ankara, Turkey (Yetiskul and Senbil, 2012)	Linear regression - MLE	1) Standard deviation journey time 2) CV journey time	(Both models) Dummy variables for time of year (month), route, weekday, and time of day, average no. of stops, average bus load, service frequency, ratio of new buses in fleet
Bus, Minneapolis, US (El-Geneidy et al., 2011)	Linear regression - OLS	1) Journey time 2) Journey time deviation (i.e. ratio actual to scheduled journey time) 3) Headway deviation (i.e. ratio actual to scheduled headway) 4) CV journey time	Models 1 & 3) Route distance, number of stops, direction, time of day, boardings, alightings, average bus load, delay at start of route, headway delay at start of route, driver experience Model 2) Route distance, number of stops, direction, time of day, boardings, alightings, delay at start of route Model 3) direction, time of day, CV stops, CV bus load, CV delay at start of route, CV driver experience
Bus, Montreal, Canada (El-Geneidy and Vijayakumar, 2011)	Linear regression - OLS	1) Dwell time 2) Journey time	(Both models) Boardings, alightings, bus load, time of day, route distance, start delay, route ID, bus type, weather
Bus, Quebec, Canada (Surprenant-Legault and El-Geneidy, 2011)	Linear regression - OLS	Journey time	Direction, dummy variable for implementation of smart card and limited stop service, dummy for limited stop service, dummy for reserved lane operation, delay at start of route, snow, boardings and alightings, rear door boarding
Bus, Montreal, Canada (Tetreault and El-Geneidy, 2010)	Linear regression - OLS	Journey time	Bus load, boardings and alightings by door (front or back), day of week, dummy for time of day, direction, weather (rain or snow), delay at start of route, stops, whether bus is low floor.
Bus, Melbourne, Australia (Mazloumi et al., 2008)	Linear regression - OLS	Journey time variability: $90\% - 10\%\ JT$	Route length, stops, signalised intersections, delay, dummies for time of day (peak vs off peak), dummy for land use (industrial area)
Bus, Portland, Oregon, US (El-Geneidy et al., 2006)	Linear regression - OLS	1) Sum of boardings and alightings 2) Journey time 3) CV journey time 4) CV headway delay	1) Population and income of persons residing within 0.25 miles of bus stop, route ID 2) Route ID, dummy for time of day and direction, dummy for consolidated bus stops route or not, boardings, alightings, no. signals, bus signal priority, stops, route length 3) CV boardings and alightings, CV stops, route length 4) CV stops, CV delay at start of route

Table A.1: Regression studies on bus journey times

Location and source	Model	Response variables	Significant explanatory variables
Bus, Portland, Oregon, US (Kimpel et al., 2005)	Linear regression - OLS	Journey time	No. stops, no. lift operations, driver experience, delay at start of route, dummy variable for time of day, scheduled run time, dummy for transit signal priority
Bus, Portland, Oregon, US (Bertini and El-Geneidy, 2004)	Linear regression - OLS	1) Dwell time 2) In-vehicle time 3) Journey time	1) Boardings, alightings 2) Route length 3) No. stops, boardings, alightings
Bus, Portland, Oregon, US (Strathman et al., 2002)	Linear regression - OLS	Journey time	Route distance, no stops, no lift operations, dummy variables for time of day, route ID, boardings and alightings, headway, dummy for summer season, dummy variables to index driver ID
Bus, Portland, Oregon, US (Strathman et al., 1999)	Linear regression - OLS	1) Arrival delay 2) Headway delay 3) Run time delay	(All models) No. stops, length of route, boardings, alightings, delay at start of route, scheduled headway, scheduled run time, dummy for peak time/direction

Note: JT is journey time, SD is standard deviation, CV is coefficient of variation, OLS is ordinary least squares, MLE is maximum likelihood estimation, SURE is seemingly unrelated regression equations.

Appendix B

Map of London Underground

The total extents of the London Underground metro system are illustrated in the map overleaf.

Source: Transport for London (2017).

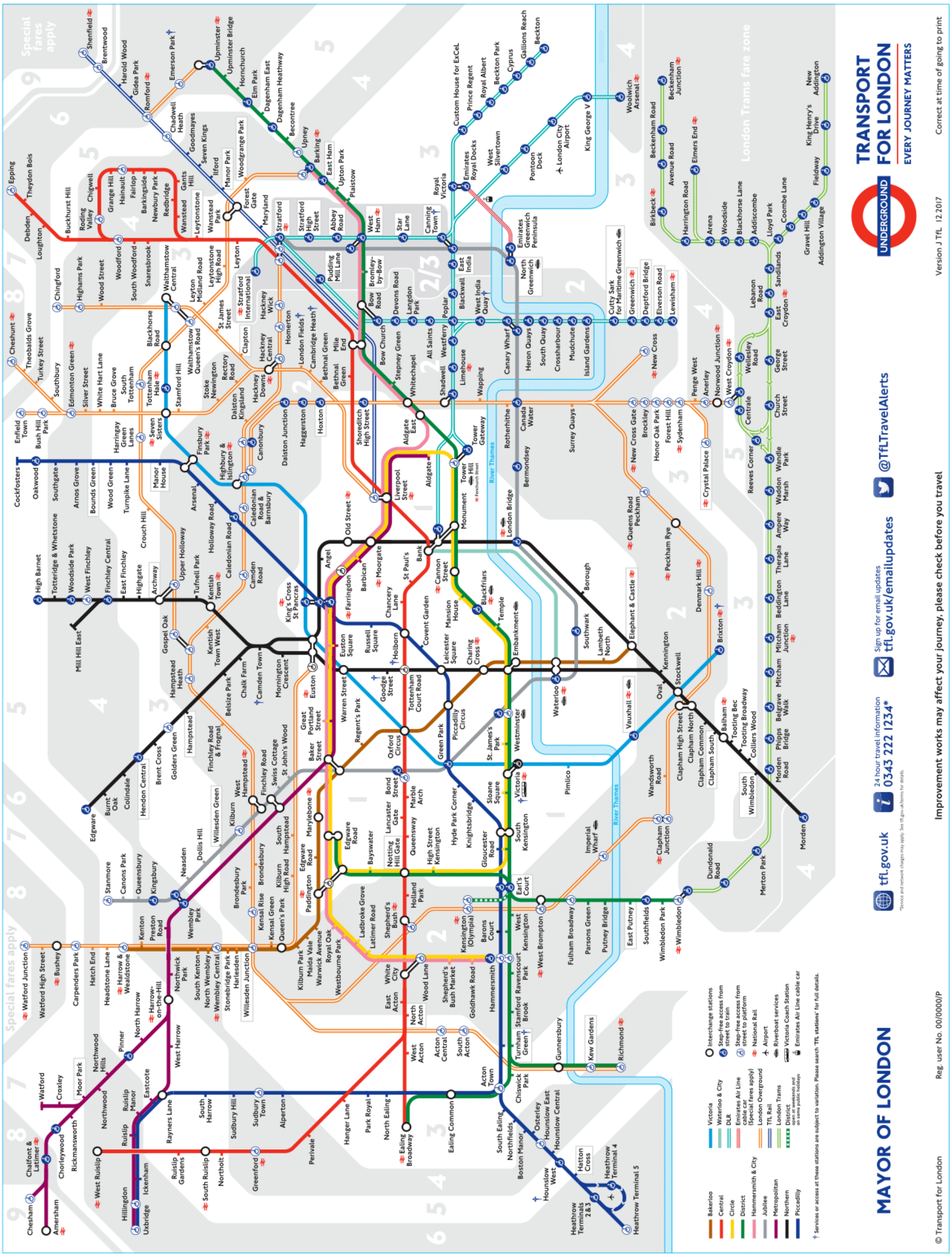

Figure B.1: Map of London Underground

Appendix C

List of TfL station and line codes

The following tables provide a list of the line, direction, and station codes that are used by Transport for London (TfL), as referenced in Chapter 3.

Table C.1: TfL codes for lines and directions

Line code	Line	Direction code	Direction
6	Jubilee	0	Westbound
6	Jubilee	1	Eastbound
3	Victoria	0	Northbound
3	Victoria	1	Southbound
2	Central	0	Westbound
2	Central	1	Eastbound
11	Waterloo and City	0	Westbound
11	Waterloo and City	1	Eastbound

Table C.2: TfL codes for stations

Station	Station NLC	Line
Bank	513	Waterloo and City
Bermondsey	787	Jubilee
Blackhorse Road	522	Victoria
Bond Street	524	Central & Jubilee
Brixton	778	Victoria
Canada Water	788	Jubilee
Canary Wharf	852	Jubilee
East Acton	563	Central
Euston	574	Victoria
Finsbury Park	580	Victoria
Green Park	590	Jubilee & Victoria
Highbury and Islington	603	Victoria
Holland Park	608	Central
Kings Cross St Pancras	625	Victoria
Lancaster Gate	629	Central
London Bridge	635	Jubilee
Marble Arch	640	Central
North Acton	653	Central
North Greenwich	789	Jubilee
Notting Hill Gate	663	Central
Oxford Circus	669	Central & Victoria
Pimlico	776	Victoria
Queensway	681	Central
Seven Sisters	698	Victoria
Shepherds Bush	700	Central
Southwark	784	Jubilee
Stockwell	716	Victoria
Tottenham Hale	729	Victoria
Vauxhall	777	Victoria
Victoria	741	Victoria
Walthamstow Central	742	Victoria
Warren Street	745	Victoria
Waterloo	747	Waterloo and City
Waterloo	796	Jubilee
West Acton	753	Central
Westminster	761	Jubilee
White City	764	Central

Note: NLC is National Location Code.

Appendix D

Train assignment method comparison

The following analysis illustrates a comparison of the train assignment algorithm proposed in Chapter 5 and the method proposed by Hörcher et al. (2017). To test the absolute accuracy of both methods, a simulated data set is required where the "true" train itinerary associated with each trip is known. The generation of suitable simulated data accounting for an accurate reflection of realistic metro travel conditions including variations in passenger congestion, train frequencies, and missed boardings is complex, and is outside the scope of this thesis.

As appropriate simulated trip and train movement data are not available, the assignment is performed on a subset of the London Underground Oyster card data and train movement data. The limitation of performing the comparative assignment on this data set is that the "true" train itinerary associated with each trip is unknown, and so this comparison highlights the relative performance of the two assignment methods compared against each other only. The sample data represent all recorded weekday (Monday to Friday) trips and train movements between Finsbury Park and Oxford Circus on the Victoria line, from October to December 2013. Summary statistics of the trips and train headways are presented in Table D.1.

The passenger trip data and vehicle location data are merged, and the timestamps are matched according to the entry and exit conditions defined in equations 5.4 and 5.5 in Chapter 5. Through this process, trips with one feasible itinerary are termed "unambiguous" trips, and trips with more than one feasible itinerary are termed "ambiguous" trips.

Table D.1: Finsbury Park to Oxford Circus data sample summary statistics (mins)

Quantity	Time period	n	Min.	Median	Mean	99%	Max.	s
Passenger journey times	AM	36784	11.00	16.00	17.15	28.00	81.00	3.37
	PM	8931	11.00	15.00	15.13	25.00	77.00	2.84
	Off-peak	3134	11.00	15.00	15.26	31.00	53.00	3.43
	Total	75098	11.00	15.00	16.11	27.00	93.00	3.24
Train headways	AM	2822	1.15	1.85	2.04	4.85	18.35	0.79
	PM	3017	1.10	1.85	1.99	4.43	11.20	0.70
	Off-peak	3355	1.07	2.58	2.89	8.29	18.80	1.36
	Total	14665	1.07	2.33	2.41	5.62	18.80	1.00

Note: n refers to the number of observations and s is the sample standard deviation.

For the Finsbury Park to Oxford Circus sample data set, 17,676 trips (24%) are unambiguously allocated out of the total 75,098 trips on the origin-destination (OD) pair. The egress times of the unambiguous trips are used to generate the probability distribution functions used as the basis of the assignment methods. The assignment method proposed by Hörcher et al. (2017) makes use of the total set of unambiguously allocated trips, while the new method proposed in Chapter 5 additionally incorporates the distribution of egress times in the off-peak period to generate updated probabilities to minimise potential bias toward the selection of the fastest egress times. The equations to assign probabilities for egress times associated with each feasible itinerary are presented in equations D.1 and D.2 (refer to Chapter 5 for a detailed discussion of the methods).

Hörcher et al. (2017) method:

$$P(C_i|I) = \frac{P(C_i)}{\sum_{j\in S} P(C_j)} \tag{D.1}$$

where

C_i refers to the itinerary of candidate train i

I is the event of observing a total set S of feasible itineraries

$P(C_i)$ and $P(Cj)$ are calculated by applying the probability distribution function for the total set of unambiguously allocated egress times.

New assignment method proposed in Chapter 5:

$$P(y_{ij}^{eg}|r_{ij}^{eg}) = \frac{P(r_{ij}^{eg}|y_{ij}^{eg})P(y_{ij}^{eg})}{P(r_{ij}^{eg})}, \qquad y_{ij}^{eg} \in \mathcal{A}, r_{ij}^{eg} \in \mathcal{U}_{op} \tag{D.2}$$

where

y_{ij}^{eg} is the egress time for passenger i and train itinerary $j = 1, ..J_i$ in the set of ambiguous trips, i.e. $y_{ij}^{eg} \in \mathcal{A}$.

r_{ij}^{eg} is a randomly sampled egress time from the set of unambiguous trips in the off-peak period, i.e. $r_{ij}^{eg} \in \mathcal{U}_{op}$.

$P(y_{ij}^{eg})$ is the prior probability for the given itinerary, calculated by applying the probability distribution function for the total set of unambiguously allocated egress times, i.e. $f_1(r_{ij}^{eg})$ (see equation 5.9 in Chapter 5).

$P(r_{ij}^{eg})$ is the marginal probability of observing the randomly sampled egress time, calculated by applying the probability distribution function for the total set of unambiguously allocated egress times, i.e. $f_1(r_{ij}^{eg})$ (see equation 5.9 in Chapter 5).

$P(r_{ij}^{eg}|y_{ij}^{eg})$ is the updated probability for the given itinerary, applying the probability distribution function for the off-peak set of unambiguously allocated egress times, i.e. $f_2(r_{ij}^{eg})$ (see equation 5.11 in Chapter 5).

Table D.2: Summary statistics for egress times of unambiguous trips (mins)

Time period	n	Median	Mean	99%	Max.	s
Off-peak	1297	1.32	1.38	3.27	38.37	1.25
Total	17676	1.30	1.33	3.18	38.37	0.74

Note: n refers to the number of observations and s is the sample standard deviation.

As shown in Table D.2, the egress times for the off-peak set of unambiguous trips are longer across all descriptive statistics compared to the total set of unambiguous egress times. Therefore, the new method of assignment will tend to allocate higher weightings to longer egress times with the application of the off-peak egress time distribution. A total of 133,298 train itineraries are identified as feasible options for 57,422 ambiguous trips (76% of data set). The mean number of feasible train options for each trip is 3.1. The two methods of assignment are performed on the set of ambiguous trips. The summary statistics comparing the resulting passenger egress times for the set of ambiguous trips are given in Table D.3. Figure D.1 illustrates kernel density plots comparing the distributions of egress times by assignment method.

Table D.3: Summary statistics for egress times of assigned ambiguous trips (mins)

	Summary statistics (mins)					Mann-Whitney test		Wilcoxon test	
Method	Median	Mean	99%	Max.	s	U-stat	P-value	W-stat	P-value
New (Chapter 5)	1.42	1.41	2.60	18.20	0.60	2.5E9	2.2E-16	2.4E6	2.2E-16
Hörcher et al. (2017)	1.35	1.34	2.60	18.20	0.60				

Note: s is the sample standard deviation.

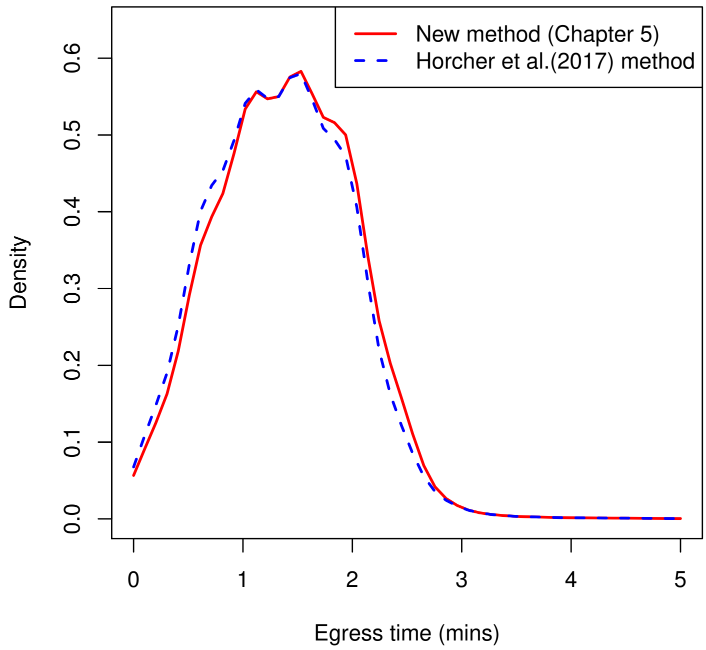

Figure D.1: Distribution of ambiguous egress times by assignment method

As shown in Table D.3 and Figure D.1, egress times corresponding to the new assignment algorithm proposed in Chapter 5 are longer than egress times produced from the assignment method proposed by Hörcher et al. (2017). To verify the level of disparity between the egress times, the unpaired non-parametric Mann-Whitney U test and the paired Wilcoxon signed-rank test are performed. The unpaired test assumes that the data are drawn from independent populations, while the paired test assumes that the data are drawn from the same underlying population. The paired test therefore provides a more conservative assessment of the degree of matching between data sets compared to the unpaired test (Lehmann and Romano, 2005). The null hypothesis for both tests is that the egress times from the two assignment methods are drawn from the same underlying distribution, while the alternative hypothesis is that egress times from the new assignment method are longer than egress times from the Hörcher et al. (2017) method.

As shown in Table D.3, both the unpaired and paired tests reject the null hypothesis, and so it can be concluded that there is a statistically significant difference between the two sets of egress times and that egress times from the new assignment method are greater than the egress times from the Hörcher et al. (2017) method. This result verifies the assumptions of the proposed algorithm; the use of egress times during the off-peak period in the new method reduces bias toward the selection of faster egress times produced by the Hörcher et al. (2017) method. As mentioned previously, this analysis compares the performance of the two methods relative to one another only. Future research is recommended to trial the two methods on high-quality simulated data, to compare how well each method performs relative to the known "true" train itinerary.

Appendix E

REML optimisation algorithm

The semiparametric regression models in Chapters 6 and 7 are fitted by penalised iteratively reweighted least squares (PIRLS), and the model smoothness parameters are estimated by restricted maximum likelihood (REML) optimisation. The algorithm presented here is an excerpt of the method detailed in Wood et al. (2015) for semiparametric regression models fitted to large data sets using the "bam" function in the package "mgcv" in the R statistical analysis software. To enable quicker computation, the algorithm estimates the model parameters without forming the total model matrix $\mathbf{X}$. As a result, computer memory requirements are in the order of $O(np)$, rather than $O(np^2)$ if the entire matrix is computed.

Initialisation

In the following iterative optimisation algorithm, $\widehat{\beta}_\lambda$ and λ are the coefficient and smoothing parameters to be estimated for the model. In the algorithm, model deviance D is calculated as per equation E.1.

$$D = 2[L(\widehat{\beta}_{max}) - L(\widehat{\beta})]\phi \tag{E.1}$$

where

$L(\widehat{\beta}_{max})$ is the maximised likelihood score of the saturated model. This is the highest possible likelihood score for the data, and it is derived from the model that assigns a parameter for each data point.

$L(\widehat{\beta})$ is the maximised likelihood of the model under consideration, and ϕ is the scale parameter of the model.

The restricted maximum likelihood (REML) score of the model $\nu_r(\lambda)$ is calculated as per equation E.2.

$$\nu_r(\lambda) = \frac{||f - \mathbf{R}\widehat{\beta_\lambda}||^2 + ||r||^2 + \widehat{\beta}_\lambda^T \mathbf{S}_\lambda \widehat{\beta}_\lambda}{2\phi} + \frac{n - M_p}{2} log(2\pi\phi) + \frac{log|\mathbf{R}^T\mathbf{R} + \mathbf{S}_\lambda| - log|\mathbf{S}_\lambda|_+}{2} \quad \text{(E.2)}$$

where

ν_r is the restricted maximum likelihood (REML)

λ is the smoothness parameter

n is the number of rows in the model matrix which represents the number of observations of the dependent variable, and p is the number of columns in the model matrix or the number of model covariates

ϕ is the scale parameter

f is the smooth function based on penalised thin plate regression splines

$\mathbf{R}$ is the upper triangular $p \times p$ factor when the total model matrix $\mathbf{X}$ is decomposed into $\mathbf{X} = \mathbf{QR}$ to enable faster computation, where $\mathbf{Q}$ is the column orthogonal $n \times p$ factor

$||r||^2$ is evaluated as $||r||^2 = ||y||^2 - ||f_k||^2$ where $f_k = \mathbf{Q}^T y$

$\widehat{\beta_\lambda}$ is the matrix of unknown model parameters

$\mathbf{S}_\lambda$ is the matrix of known coefficients and $|\mathbf{S}_\lambda|_+$ is the product of the strictly positive eigenvalues of $\mathbf{S}_\lambda$

M_p is the formal degree of rank deficiency of $\mathbf{S}_\lambda$

In the next iteration step, x_i denotes the vector of covariates that are associated with response variable y_i where $i = 1, .., n$, and divide the integers from 1 to n into M non-overlapping subsets $\gamma 1, .., \gamma M$ of approximately equal size (so $\cup_i \gamma_i = 1, .., n$ and $\gamma_j \cap \gamma_i = \emptyset$ for all $i \neq j$. M is allocated based on computer memory. Let $\bar{\eta}_i = g(y_i + \xi_i)$ (with ξ_i as defined in the previous section). Set the PIRLS iteration index $q = 0$ and $D = 0$ (or any constant). Perform any initialisation steps that are necessary to set up the thin plate regression spline bases for the smooth terms.

Iteration

Step 1: set $D_{old} = D$, $\mathbf{R}$ to be a $0 \times p$ matrix, f a 0-vector, $D = 0$ and $r = 0$.

Step 2: repeat the following steps (a) to (f) for $k = 1, .., M$.

(a) Set $f_0 = f$ and $\mathbf{R}_0 = \mathbf{R}$.

(b) Form the model submatrix $\mathbf{X}_k$ for the covariate set $x_i : i \in \gamma_k$.

(c) if $q > 0$ form $\widehat{\eta} = \mathbf{X}_k \widehat{\beta}_\lambda$; otherwise $\widehat{\eta} = \bar{\eta}_{\gamma k}$.

(d) Form $\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i)$, $z_i = g^{'}(\widehat{\mu}_i)(y_i - \widehat{\mu}_i + \widehat{\eta}_i)$ and $w_i = V(\widehat{\mu}_i)^{-0.5} g^{'}(\widehat{\mu}_i)^{-1}\ \forall\ i \in \gamma_k$. Let z be the vector containing these z_i values and $\mathbf{W}$ be the diagonal matrix of corresponding w_i values.

(e) Set $r \leftarrow r + ||\mathbf{W}z||^2$ calculate the deviance residuals for the current subset of data and add the sum of squares of these to D.

(f) Form $\mathbf{QR} = \begin{pmatrix} \mathbf{R}_0 \\ \mathbf{W}\mathbf{X}_k \end{pmatrix}$ and $f = \mathbf{Q}^T \begin{pmatrix} f_0 \\ \mathbf{W}z \end{pmatrix}$ and discard $\mathbf{Q}$.

Step 3: set $||r||^2 = r - ||f||^2$.

Step 4: if $q > 0$ test for convergence by comparing the current deviance D with the previous deviance D_{old}. Stop if convergence has been reached (or q has exceeded some predetermined limit suggesting failure).

Step 5: estimate λ by optimising $\nu_r(\lambda)$. This also yields $\widehat{\beta}_\lambda$.

Step 6: $q \leftarrow q + 1$

At convergence the final $\widehat{\beta}_\lambda$ and λ are the coefficient and smoothing parameter estimates for the model.

Appendix F

Chapter 6 supplementary material

As discussed in Chapter 6, plots of excluded covariates where journey times do not vary in a systematic way are illustrated in Figure F.1 to F.6.

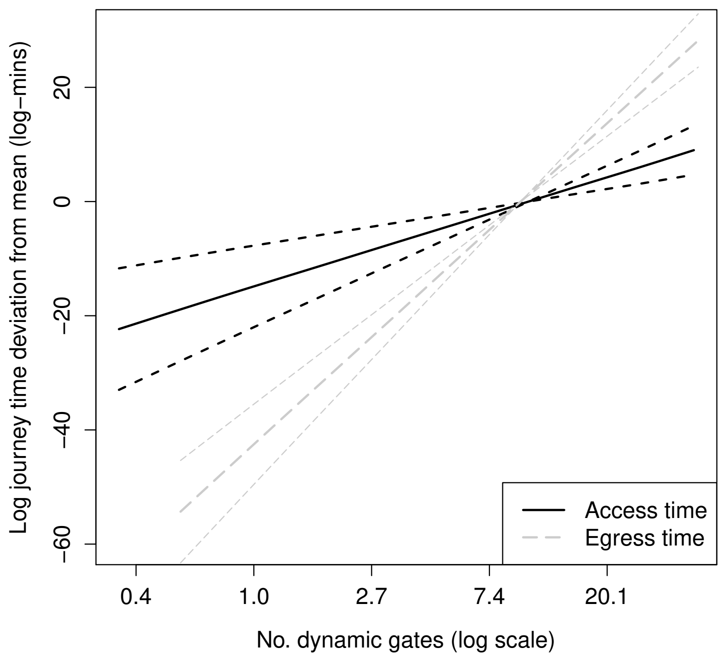

Figure F.1: Regression plot of dynamically available gates

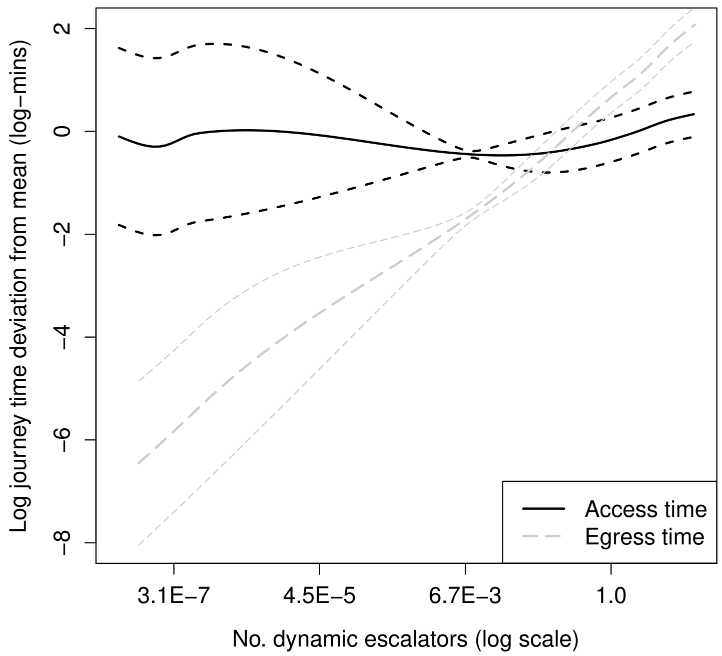

Figure F.2: Regression plot of dynamically available escalators

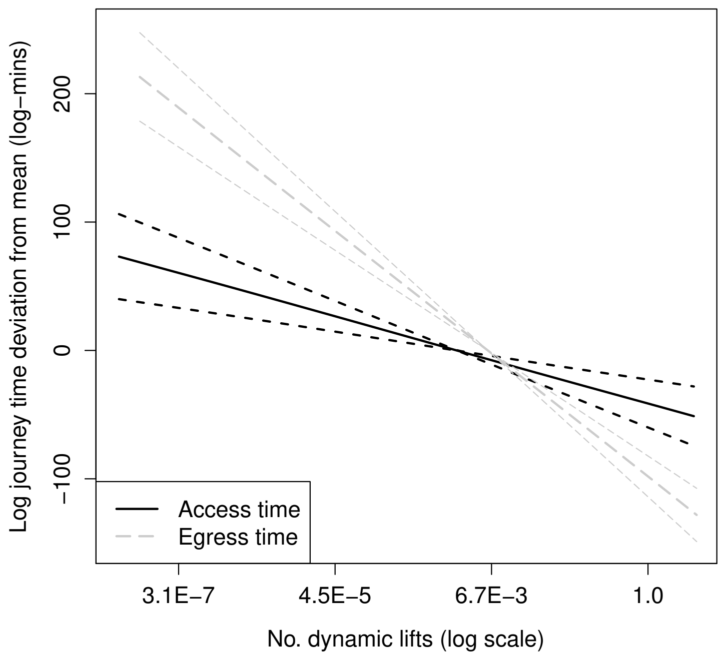

Figure F.3: Regression plot of dynamically available lifts

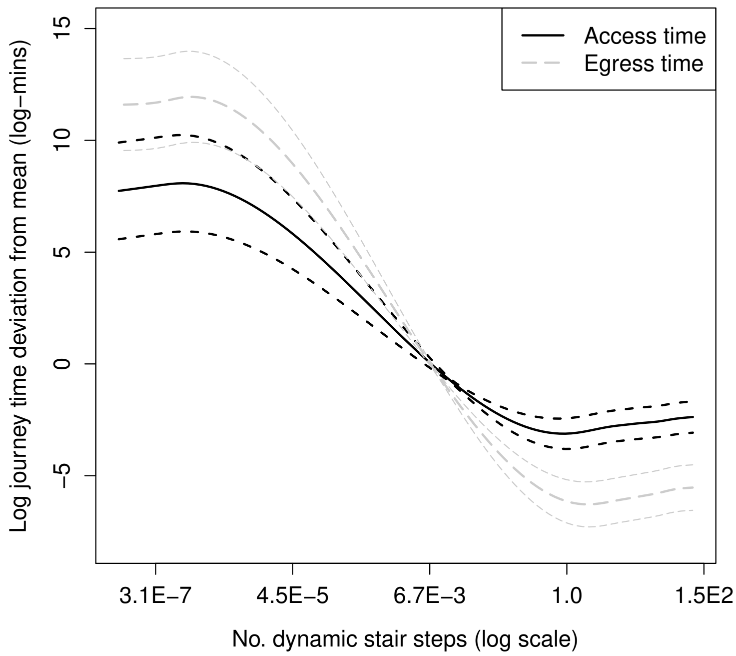

Figure F.4: Regression plot of dynamically available stairs

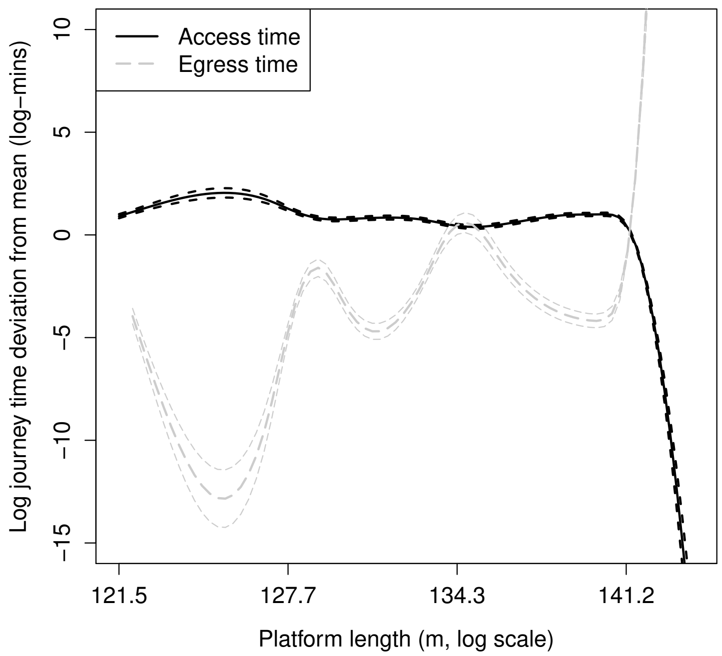

Figure F.5: Regression plot of platform length

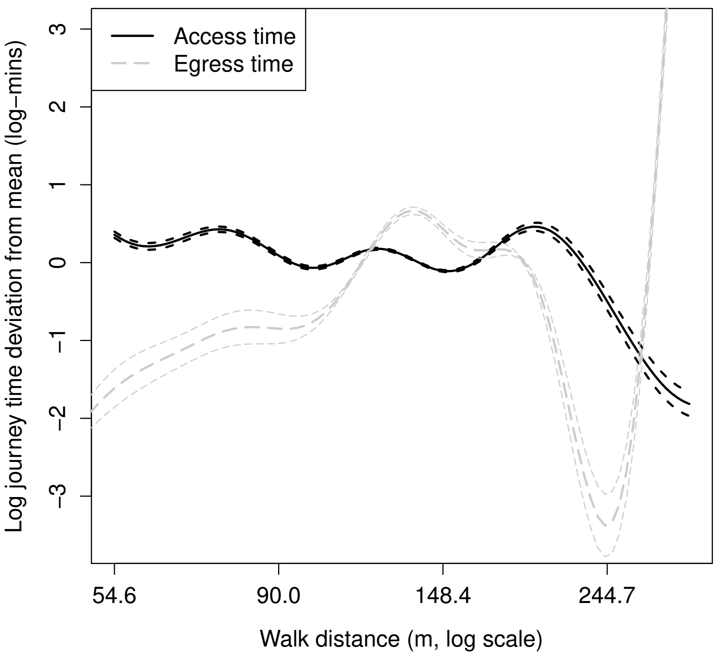

Figure F.6: Regression plot of station walk distance

The following table and figures provide information on the residuals for the semiparametric models presented in Chapter 6.

Table F.1: Model residuals summary statistics (mins)

Model	Min.	Median	Mean	Max.	s
Access time	-4.09	0.046	-5.6E-11	3.41	0.53
Egress time	-3.33	0.090	5.1E-10	3.62	0.57
On-train time	-4.5E-07	3.5E-10	-1.1E-12	4.9E-07	5.1E-08
Total journey time	-0.955	-0.007	2.4E-12	2.14	0.14

Note: s is the sample standard deviation.

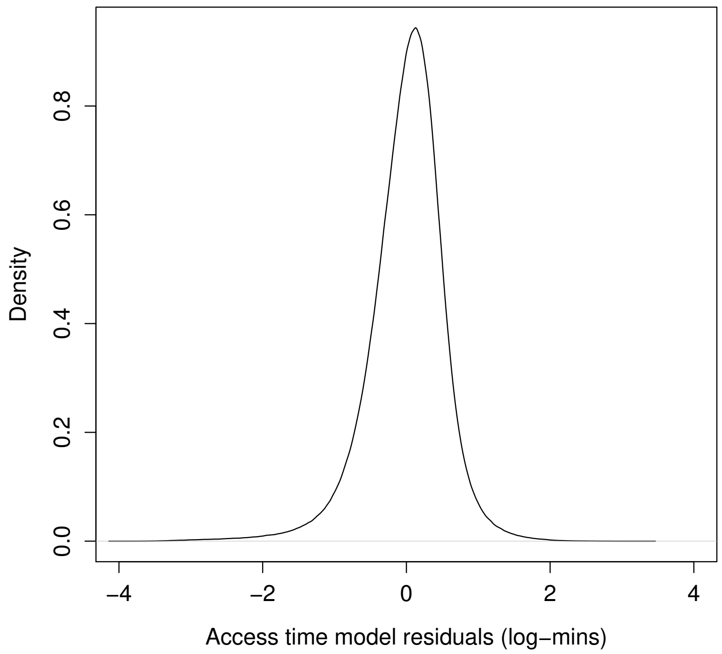

Figure F.7: Access time model residuals

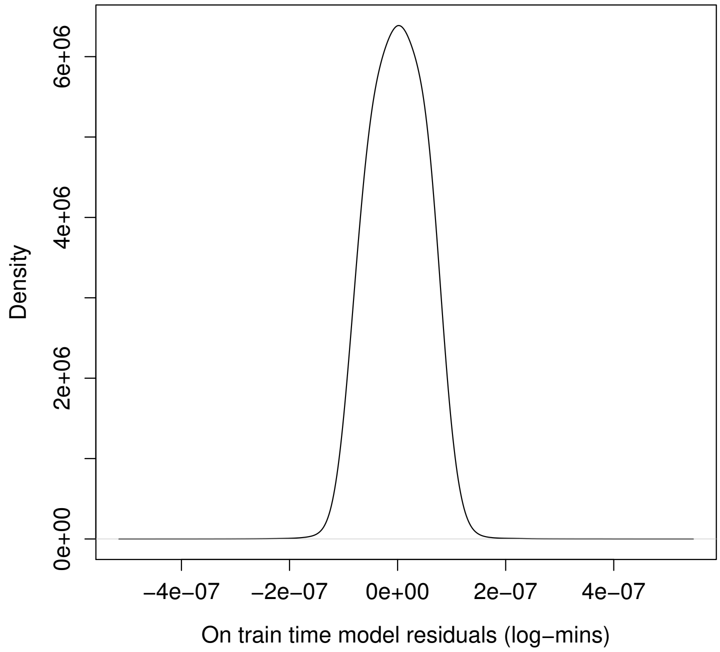

Figure F.8: On train time model residuals

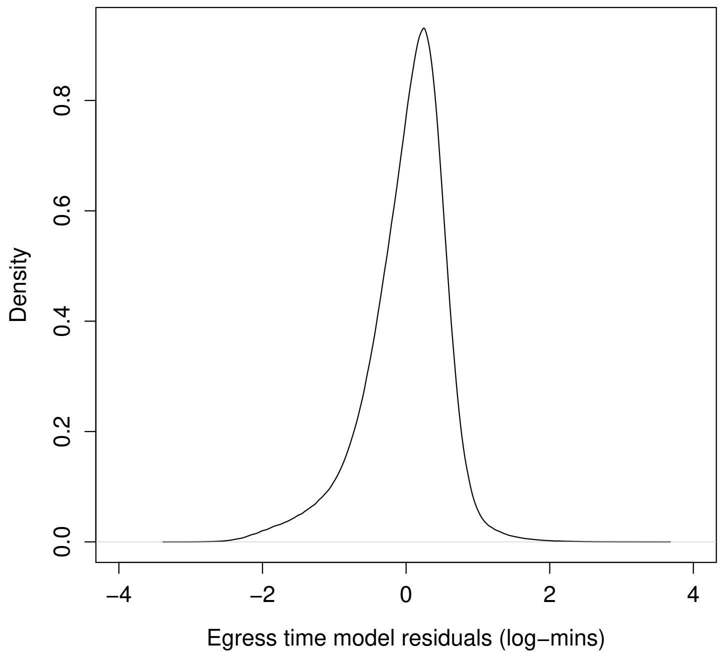

Figure F.9: Egress time model residuals

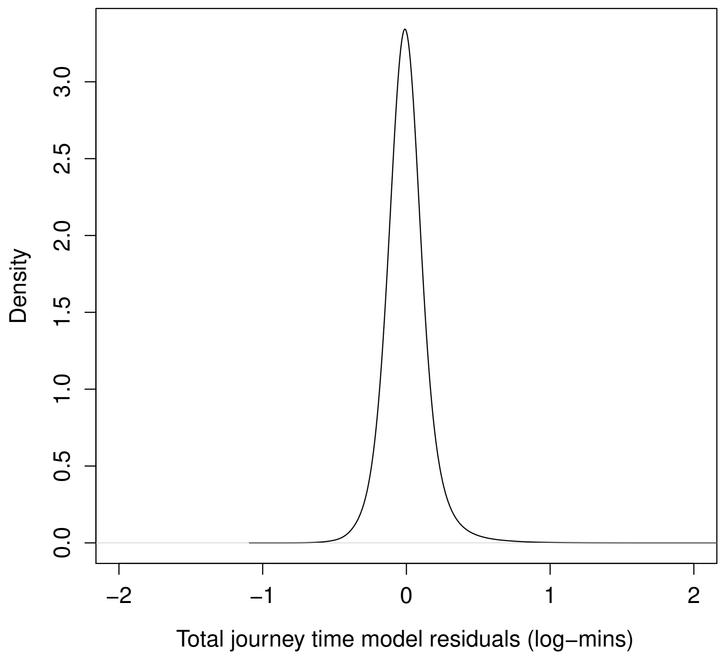

Figure F.10: Total journey time model residuals

Appendix G

Chapter 7 supplementary material

The following table and figures provide information on the residuals for the semiparametric models presented in Chapter 7.

Table G.1: Model residuals summary statistics (mins)

	Time period	Min.	Median	Mean	Max.	s
Access time	AM peak	-3.65	0.06	-1.2E-13	2.50	0.52
	Inter-peak	-3.32	0.08	2.4E-13	3.06	0.50
	PM peak	-3.28	0.05	2.2E-13	2.14	0.48
Egress time	AM peak	-2.87	0.08	1.6E-14	3.35	0.50
	Inter-peak	-2.61	0.07	-5.3E-14	2.25	0.53
	PM peak	-2.71	0.08	2.2E-14	3.03	0.53
Total journey time	AM peak	-0.86	0.00	-2.9E-8	1.45	0.11
	Inter-peak	-0.70	0.00	7.4E-9	1.56	0.11
	PM peak	-0.51	-0.01	-1.7E-8	1.12	0.11

Note: s is the sample standard deviation.

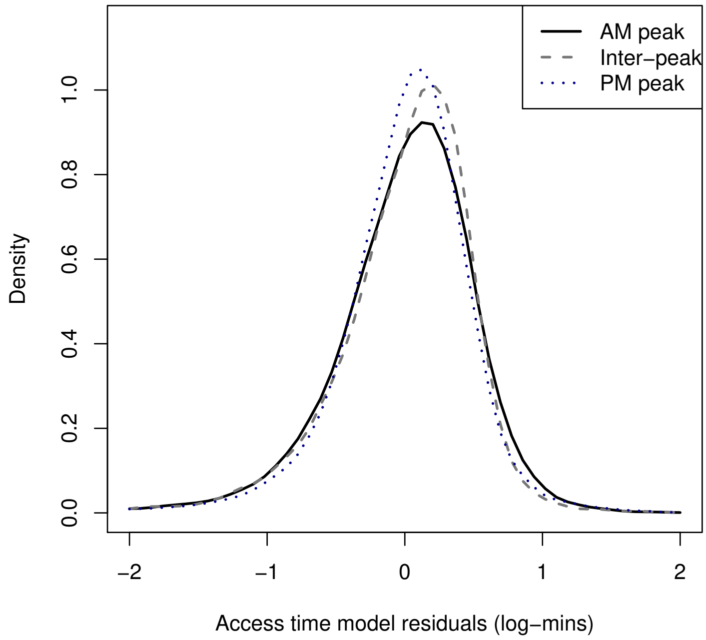

Figure G.1: Access time model residuals

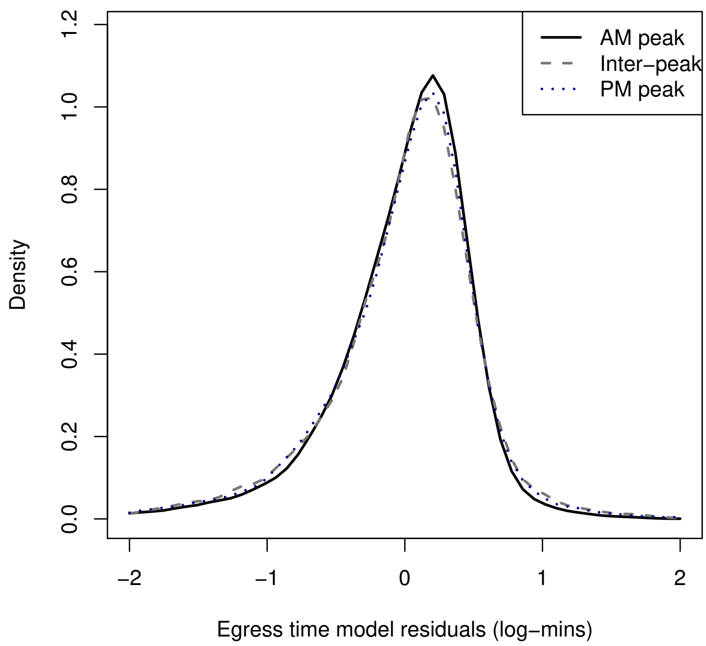

Figure G.2: Egress time model residuals

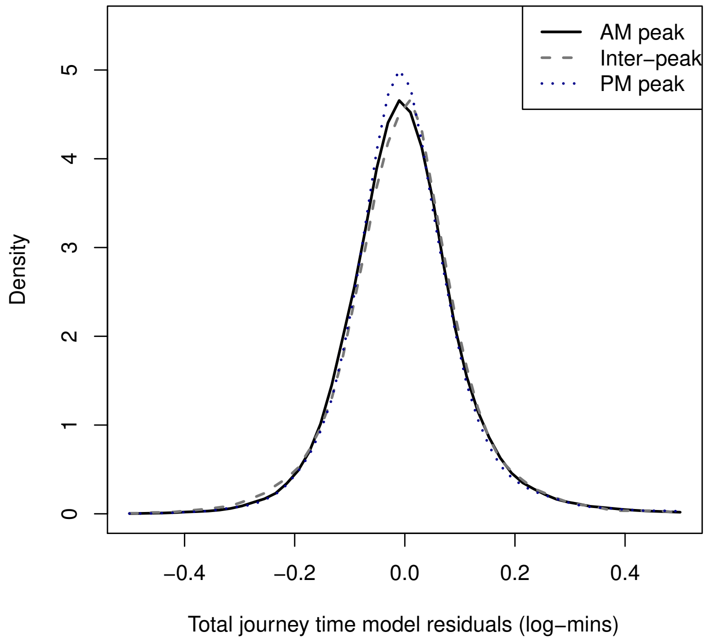

Figure G.3: Total journey time model residuals