UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Civil, Maritime and Environmental Engineering and Science

Modelling railway station choice: can probabilistic catchments improve demand forecasts for new stations?

by

Marcus Adrian Young

ORCID ID 0000-0003-4627-1116

Thesis for the degree of Doctor of Philosophy

February 2019

When the train starts, and the passengers are settled
To fruit, periodicals and business letters
(And those who saw them off have left the platform)
Their faces relax from grief into relief,
To the sleepy rhythm of a hundred hours.
Fare forward, travellers! not escaping from the past
Into different lives, or into any future;
You are not the same people who left that station
Or who will arrive at any terminus,
While the narrowing rails slide together behind you

T. S. Eliot, “The Dry Salvages”

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND THE ENVIRONMENT
Civil, Maritime and Environmental Engineering and Science

Doctor of Philosophy

MODELLING RAILWAY STATION CHOICE: CAN PROBABILISTIC CATCHMENTS IMPROVE DEMAND FORECASTS FOR NEW STATIONS?

by Marcus Adrian Young

The aim of this thesis is to determine whether the performance of the aggregate rail demand models that are commonly used to forecast demand for new railway stations can be improved by incorporating probabilistic station catchments derived by means of station choice models. The current approaches to forecasting demand for new railway stations have been examined and their limitations identified, and previous work to develop station choice models and incorporate them into demand models has been reviewed. A series of station choice models able to predict station choice at small-scale origin zones were calibrated using revealed preference data from passenger surveys carried out in Scotland and Wales. An automated data processing framework, incorporating a bespoke multi-modal route planner, was developed to derive the model predictor variables from disparate sources of open transport data. The station choice models were found to perform substantially better at predicting station choice than a base model where the nearest station was assumed to be chosen. Trip end models were calibrated for Category E and F stations in Great Britain, using both deterministic and probabilistic station catchments, and a methodology was developed to apply these models to predict demand for new stations and to assess the effect of abstraction on existing stations. The methodology was used to forecast demand at several recently opened stations, including a newly opened line. The models with probabilistic catchments were found to perform better than those with traditional deterministic catchments, and to produce more accurate forecasts than those made during the scheme appraisal process. This is the first known example of successfully incorporating probabilistic station catchments into an aggregate rail demand model, and represents a significant advance over previous work in this area. These findings have important policy implications. They can be used to update industry guidance on best-practice for implementing this type of model in a local context and, more importantly, provide the basis of a robust and transferable national trip end model for Great Britain.

Contents

List of Figures

List of Tables

List of Accompanying Material

Information about the availability of, and access to, supporting data for this research project is available from the University of Southampton repository at:

https://doi.org/10.5258/SOTON/D0825

Declaration of Authorship

I, Marcus Adrian Young, declare that this thesis entitled *Modelling railway station choice: can probabilistic catchments improve demand forecasts for new stations?* and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: Young and Blainey (2018a), Young and Blainey (2018b), Young (2017b), Young (2017a), Young (2016), and Young and Blainey (2016).

Signed:

Date:

Acknowledgements

This research project was supported by the Engineering and Physical Sciences Research Council (DTG Grant EP/M50662X/1). Their funding not only enabled me to undertake this PhD, but also supported my attendance at a number of academic conferences where I was able to share my work and meet many other PhD students and researchers.

I would like to thank my supervisor, Dr Simon Blainey, for initially having confidence in my ability to complete this research project, which was his original idea, and for all the advice, support and feedback he has provided over the past four. I am sure I would not have achieved so much, especially beyond the strict requirements of the PhD, had it not been for his level of engagement and encouragement. I am also grateful for the guidance provided by Prof. John Preston, in particular his invaluable feedback on the draft of this thesis.

Finally, this undertaking would not have been possible without the encouragement, understanding and patience of my partner, family and friends; for which I am immensely thankful.

Data acknowledgements: the support of the following data providers is gratefully acknowledged. The Welsh Government and Transport Scotland for providing passenger survey data; Paul Kelly for permission to use the brfares.com API; Dan Saunders at Basemap Ltd for providing a TRACC educational license; and Ordnance Survey Ltd for providing AddressBase data. Map data, Code.Point and Code.Point Polygons ©Crown Copyright and Database Right. Ordnance Survey (Digimap Licence). This work uses public sector information licensed under the Open Government Licence v3.0. Basemap data ©OpenStreetMap contributors.

Abbreviations

AIC	Akaike information criterion
ASC	alternative specific constant
ATOC	Association of Train Operating Companies
BIC	Bayesian information criterion
CDM	competing destinations model
CNL	cross-nested logit
DAFNI	Data and Analytics Facility for National Infrastructure
DfT	Department for Transport
GDP	gross domestic product
GEV	generalised extreme value
GJT	generalised journey time
GNL	generalized nested logit
GSCL	Generalised Spatially Correlated Logit
GTFS	General Transit Feed Specification
HHI	Herfindahl-Hirschman Index
IIA	independence from irrelevant alternatives
IIGD	independent and identically extreme value (Gumbel) distribution
IV	inclusive value
LATIS	Land-Use and Transport Integration in Scotland
LENNON	Latest Earnings Networked Nationally Overnight
LL	log-likelihood
MAUP	modifiable areal unit problem
ML	mixed logit
MNL	multinomial logit
MSE	mean squared error
NaPTAN	National Public Transport Access Nodes

NL	nested logit
NRE	National Rail Enquiries
NRPS	National Rail Passenger Survey
NRTS	National Rail Travel Survey
OD	origin-destination
ONSPD	ONS Postcode Directory
ORR	Office of Rail and Road
OS	Ordnance Survey
OSM	OpenStreetMap
OTP	OpenTripPlanner
PCL	paired combinatorial logit
PDFC	Passenger Demand Forecasting Council
PDFH	Passenger Demand Forecasting Handbook
PLD	PLANET Long Distance
POI	Ordnance Survey Points of Interest
RAM	random access memory
RP	revealed preference
RPL	random parameter (mixed) logit
RSQI	rail service quality index
RUM	random utility model
SD	standard deviation
SP	stated preference
SWEC	spatially weighted error correlation
TfL	Transport for London
TNDS	Traveline National Dataset
WG	Welsh Government

Chapter 1

Introduction

The railway in Britain has experienced considerable growth in recent years, with total passenger journeys increasing by 51% (an additional 584 million journeys) over the past decade alone (Office of Rail Regulation, 2017). This has been accompanied by an expansion in the network, with 56 new stations and several new lines opening over the same period (Alderson & McDonald, 2017). This growth looks set to continue, with further new lines and stations currently under construction or planned, and campaigns being run nationwide by communities eager to be connected to the rail network (Campaign for Better Transport, 2017). However, there are concerns about the accuracy of the station demand forecasts that are used to determine the viability of proposed new schemes. A report commissioned by the UK Government to investigate the issue, compared forecast and observed demand at 23 newly opened stations (Steer Davies Gleave, 2010). It found that forecast demand was above or below observed demand by a margin of more than 20% in 14 cases, including an under-prediction in excess of 100% at three stations. More recently, the demand forecast for the new Borders Railway line in Scotland was described as a 'shocking failure' (Campaign for Borders Rail, 2016), after usage figures revealed that passenger trips in the first year of operation were up to eight times higher than forecast for three of the new stations, and lower than predicted for the other four. Inaccurate forecasts can have potentially serious consequences. Under-prediction might lead to the unnecessary rejection of a proposal on the grounds of the perceived benefit-cost ratio, or to the inadequate provision of station and network infrastructure. Conversely, over-prediction, or not adequately accounting for abstraction from existing stations, could result in a new station that fails to deliver the expected economic and societal benefits.

1.1 The station catchment problem

Although the UK Department for Transport (DfT) has published some general guidance for those carrying out or commissioning demand forecasts for new local railway stations

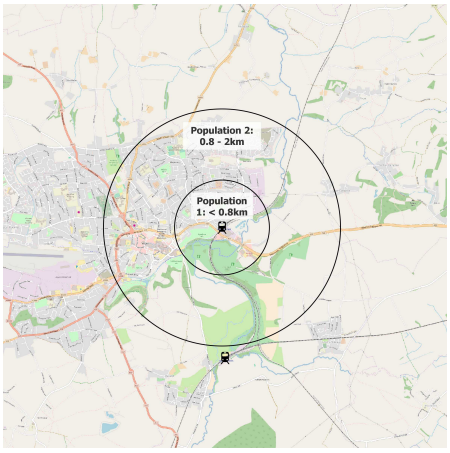


FIGURE 1.1: Example of a radial station catchment that is divided into two bands.

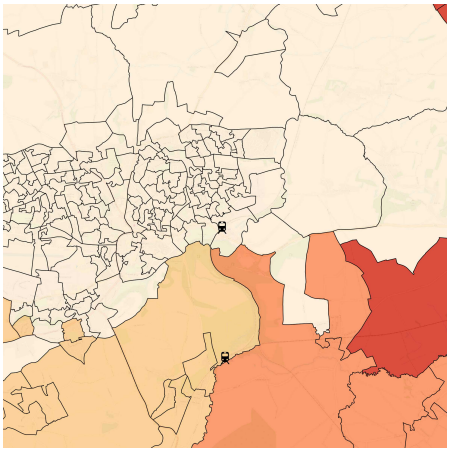


FIGURE 1.2: Example of zone-based station catchments (using census output area as zone).

(Department for Transport, 2011), the models used are usually developed for, and applied to, a specific local context. In most cases trip rate or trip end models are adopted, as was the case in two-thirds of the stations/lines considered in the Steer Davies Gleave report. Trip rate models assume the number of trips to be some function of the population in the area surrounding a station (its catchment), while trip end models include additional variables relating to station services, facilities or the locality. Previous work by Blainey (2010) and Blainey and Preston (2010) successfully calibrated national trip end models suitable for general application in forecasting demand for new local rail stations in England and Wales. However, a weakness of this previous work, in common with trip end models generally, lies in how the station catchments are defined. Two methods are typically used; either a distance- or time-based buffer is placed around the station (see Figure 1.1), or the study area is divided into zones and each zone is assigned to its nearest station (see Figure 1.2). The latter was the method adopted by Blainey (2010), with census output areas used as the zonal units. Both approaches produce discrete non-overlapping catchments which imply that station choice is a deterministic process (anyone beginning a trip from within a zone will always choose to board at the same station) and that stations do not compete with each other for passengers.

While there will be some trip origins where the choice of station is effectively deterministic, with the probability of the nearest station being chosen at or very close to one, it is not difficult to find real-world examples where this is unlikely to be the case, either based on personal experience of the author, or through conversations with other rail users. For example, consider the potential choice of station for someone beginning their trip from the Boldrewood Innovation Campus in Southampton. The five most likely alternative stations, listed in order of road distance from the campus, are: Swaythling, St. Denys, Southampton Airport Parkway, Southampton Central, and Eastleigh. The location of the campus and each of these stations, together with some key characteristics of the stations, are shown on the map in Figure 1.3. A deterministic catchment would assign all trips originating from the campus to the nearest station, which in this example would be Swaythling. For those walking to the station, only

Swaythling or St. Denys are realistic options, with the access journey taking around 30 minutes. This is likely to be too far for many travellers, especially given the availability of a regular and reliable bus service as an alternative main access mode, suggesting that walk-only mode will account for only a small proportion of station access journeys. Although it is the nearest station, Swaythling has the lowest daily service frequency, and is only served by a single train operating company. While Swaythling is easy to access by bus from the campus, so are the other stations (apart from St. Denys) and these may be preferred due to more frequent services and the greater range of destinations served by direct trains. If driving to a station by car from the campus then all five stations are relatively easy to get to, but parking provision is very limited at Swaythling and St. Denys, suggesting that one of the other three stations will be chosen instead, especially given their superior service levels. The ultimate destination is also likely to influence the station chosen. For example, if travelling to the West Country via Salisbury then fast services to Salisbury and beyond are only available from Southampton Central; while someone travelling to London might choose Southampton Airport Parkway or Eastleigh, thus avoiding a potentially congested drive through central Southampton. These stations are also in the direction of travel, thus reducing the on-train journey time and fare (for example, at the time of writing an off-peak return from Eastleigh to London is £4.10 cheaper than the same ticket from Southampton Central). Taking all these factors into account, it seems likely that a relatively small proportion of travellers beginning their journey at the Boldrewood Innovation Campus would actually choose to depart from Swaythling, with the majority preferring to depart from other stations.

This anecdotal evidence is supported by prior research, discussed in detail in Chapter 2, which shows that in reality station catchments are far more complex entities than the simple catchment definitions allow. Simplistic catchments have been found to account for only 50–60 percent of observed trips (Blainey & Preston, 2010). Station choice is not homogeneous within zones, with catchments overlapping (Mahmoud, Eng, & Shalaby, 2014) and varying by access mode (Mahmoud et al., 2014) and station type (Lythgoe & Wardman, 2004).

If station catchments are not correctly defined in the aggregate demand models, then inappropriate weight will be given to other model variables, such as service quality measures, as drivers of trip generation, rather than the catchment population. By defining more realistic catchments, the parameter estimates will be more robust, and the models will be more transferable (Wardman & Whelan, 1999). A potential mechanism to improve the representation of station catchments would be to define them probabilistically, by using an appropriate station choice model to calculate the probability of a particular station being chosen for each zone. While there is a substantive body of prior station choice research (which is comprehensively reviewed in Chapter 3), the studies have primarily focussed on identifying and understanding the relevant explanatory factors, with relatively little attention given to how station choice models could be used to improve rail demand models. There are two notable exceptions. The first is the attempt by Wardman and Whelan (1999) to incorporate probabilistic station

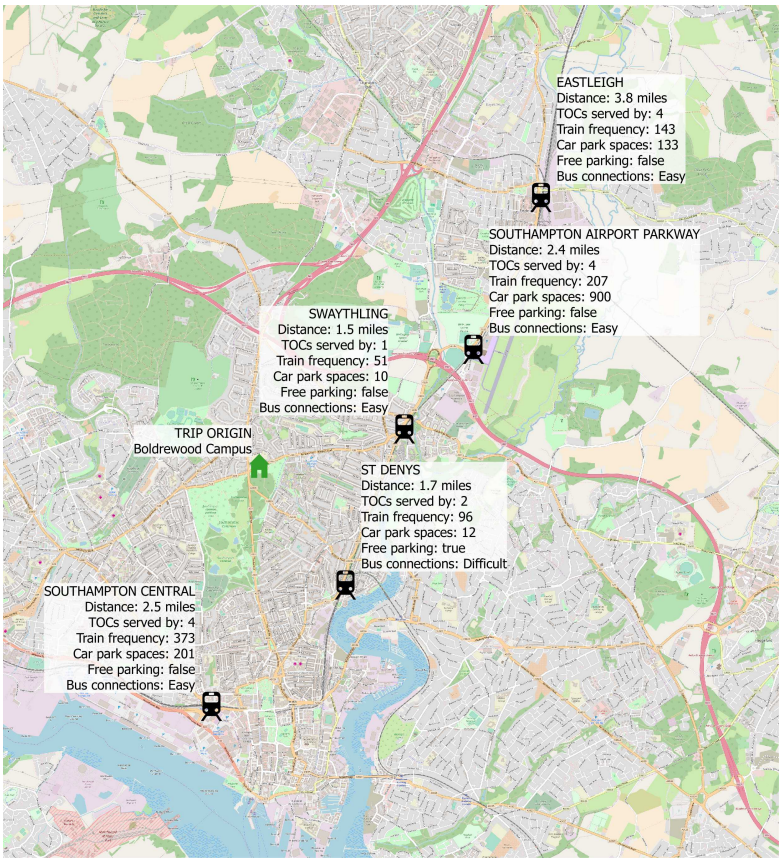


FIGURE 1.3: Railway stations that might be chosen by someone beginning a trip at Boldrewood Innovation Campus, showing key characteristics (TOC = train operating company).

catchments into a flow model[1] by apportioning the population of postal sectors[2] to one of five competing stations. However, due to time and computer resource constraints they had to use a subset of the flow data, which resulted in the model failing to converge. They recommended further work, noting that they had 'seriously underestimated the complexity of [the] task and the computing and time resources required'. However, this approach has not been revisited since, despite the considerable advances in computing power over the past two decades. The second, and probably the most refined methodologically, is that proposed by Lythgoe and Wardman (2002, 2004), where station choice is an intrinsic component of a flow model, with a station's generation potential represented by the population within 40 km allocated to a grid of zones. However, this approach was intended to forecast demand for parkway stations and is limited to modelling inter-urban journeys greater than 80 km (subsequently reduced to 40 km by Lythgoe, Wardman, and Toner (2004)).

The research described in this thesis will seek to address the problem by developing station choice models that can be used to generate probabilistic station catchments, which can subsequently be integrated into trip end or flow models. For example, a set of alternative stations could be allocated to each unit postcode, and the probability of each station being

[1] Flow models forecast trips from each origin station to each destination station and additionally take account of the train leg and characteristics of the destination.

[2] There are approximately 3,000 addresses in a postcode sector.

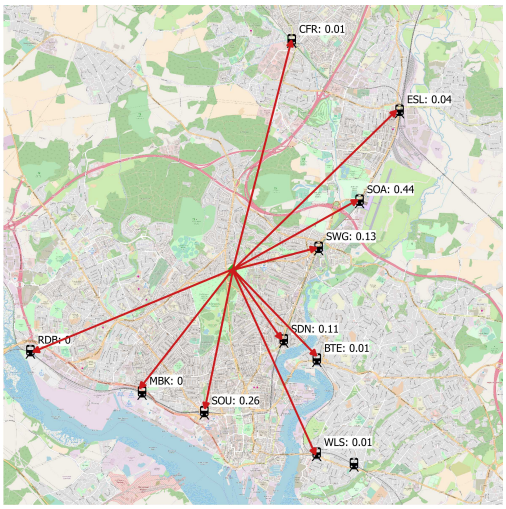


FIGURE 1.4: The concept of assigning 10 alternative stations to each zone and calculating the probability of each being chosen (assuming Boldrewood campus is the zone centroid).

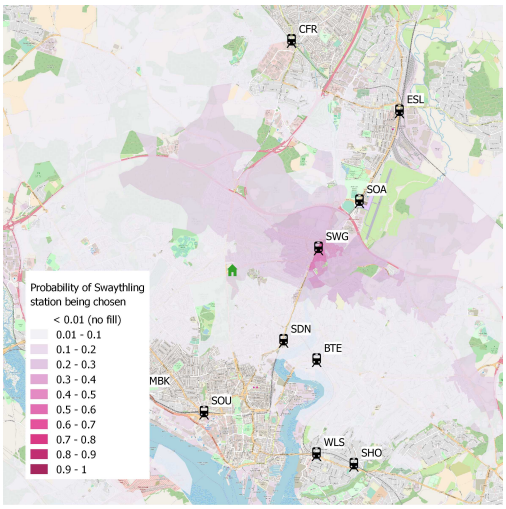


FIGURE 1.5: How a probabilistic catchment for Swaythling station might look, with the probability of the station being chosen shown for each postcode (using simulated probabilities).

chosen would then be calculated using a station choice model. Figure 1.4 illustrates the concept, using Boldrewood Innovation Campus as an example zone centroid with ten alternative stations. The probabilistic catchment for a specific station could then be 'generated', as illustrated in Figure 1.5, where the probability of Swaythling station being chosen for each postcode polygon is represented using a choropleth. The population of each postcode, obtained from census data, would be apportioned to each station based on these probabilities, thus forming the population explanatory variable for a trip end model. Ideally, the calibrated station choice models would be readily transferable, and not limited to modelling station choice in specific situations or local contexts. Collecting the necessary observed choice and explanatory data to calibrate a rigorous and effective predictive model on a case-by-case basis is an expensive and time-consuming process, and this could be avoided if a single generalised model with wide applicability was developed.

1.2 Research aim and objectives

The overall aim of this thesis is to determine whether the performance of the aggregate railway station passenger demand models can be improved by incorporating probabilistic station catchments derived by means of station choice models. The key objectives that must be met to achieve this aim are as follows:

1. Obtain, process and validate suitable survey datasets that can reveal observed station choice behaviour, ideally covering more than one region of GB.

2. Derive candidate predictor variables for the station choice models, with a particular focus on maximising the potential of open transport data sources that have recently become available. This should include an accurate representation of access journeys and train-leg components, obtained using a suitable multi-modal route planner.

3. Calibrate station choice models appropriate for integrating into aggregate rail demand models, and assess their predictive performance and transferability.

4. Develop a methodology to incorporate probability-based station catchments into aggregate demand models and apply this methodology to calibrate a national-scale model for local railway stations in GB.

5. Develop a practical methodology for generating demand forecasts for new stations using the national-scale model, and for estimating abstraction effects from existing stations.

6. Apply the demand forecasting methodology to several case studies, and carry out a performance appraisal, including an assessment of models with either deterministic or probabilistic station catchments.

1.3 Research scope

The main focus of the original contribution of this thesis is to challenge the long-standing convention that station catchments in aggregate rail demand models should be defined in a deterministic manner, by developing and appraising an alternative approach where station catchments are defined probabilistically using models of station choice. To accomplish this overarching contribution, three core elements of research were completed: calibration of station choice models; development of a national trip end model for GB in which station catchments are defined probabilistically using the station choice models; and the application and appraisal of the national trip end model. Each core element can be divided into several sub-elements each with a specific research focus. These research elements are summarised in Figure 1.6. To help guide the reader, the original contribution and/or significant outputs or findings that arose from each element are also presented in this diagram. A full discussion of the empirical and methodological contributions to knowledge in the field of rail demand forecasting arising from this thesis is reserved for the final chapter (Section 9.3.3).

1.4 Thesis structure

Following on from this introduction, Chapter 2 considers the current state of practice with regard to forecasting demand for new railway stations in the UK, identifying weaknesses in how station catchments are defined in the aggregate demand models that are typically

Core research element	Sub-element	Chapter	Research focus	Original contribution or significant outputs/findings
Station choice models Dissagregate models with choice modelled at unit postcode level and accounting for spatial correlation.	Revealed preference survey data	4	Validate reported trips and maximise usability of responses with incomplete address information	Novel methods to process and validate OD survey data: maximising usability and identifying common errors.
	Predictor variables	5	Derive predictor variables from open transport data with emphasis on accurately representing choice attributes faced by decision maker.	A framework to automatically generate variables from open transport data supported by an API wrapper to query a bespoke trip planner.
	Calibration	6	Use established model forms to calibrate using two regional datasets; account for spatial correlation; assess transferability and predictive performance.	MNL models ready for integration into trip end or flow models with choice modelled at high spatial resolution. Fotheringham's CDM used to capture spatial competition effects - first application in this context.
Trip end models A national trip end model for new local railway stations that incorporates probabilistic station catchments.	New model form and calibration dataset	7	Develop suitable model form to incorporate probabilistic catchments, building on previous work by Blainey and Preston (2013b).	First example of incorporating probabilistic catchments into aggregate rail models. Enhances earlier work with higher resolution zones, decay functions from observed data and larger dataset.
	Generating probabilistic catchments	7	Develop methods to generate station choice probabilities for 1.5 million postcodes in mainland GB.	SQL procedural code to efficiently calculate 14 million accessibility terms. Development of station choice predictor application in R Shiny to sense-check probabilities and visualise catchments.
	Performance assessment	7	Compare models with deterministic and probabilistic catchments. Examine residuals and investigate geographic variation in performance.	Models with probabilistic catchments performed better, with greater weight given to the population variable. Analysis of standardised residuals indicated no systematic basis at regional level.
Application and appraisal Methods to apply the trip end model to forecast demand for new stations and assess abstraction effects.	New methodology	8	Develop methodology for applying trip end model.	Methodology developed based around observed data showing that <1% of access journeys exceed 60 minutes. Streamlined version proposed when forecasting a new line.
	Case studies	8	Generate demand forecast for several case study stations using both deterministic and probabilistic approaches.	Performance assessed for ten stations opened since the calibration base year. The probabilistic catchment model out-performed the deterministic model and official scheme forecasts in most cases.
	Abstraction analysis	8	Develop methodology to forecast abstraction from existing stations and assess impact of the accessibility term on proportional substitution.	Methodology developed and tested for a proposed new station in Wales. Analysis of the effect of accessibility term showed an appropriate adjustment to the proportional substitution pattern.

FIGURE 1.6: Diagram summarising the elements of research undertaken as part of this thesis.

adopted, and assessing the accuracy of demand forecasts produced over recent years. Chapter 3 then presents a review of prior station choice research, and considers the extent to which station choice models have been used to improve rail demand models. Chapter 4 is concerned with the observed station choice data that forms the basis for this research, describing how it was cleaned and validated, and providing some descriptive analysis, including the visualisation of station catchments. Chapter 5 deals with the data sources for the predictor variables used in the station choice models, including the development of a multi-modal route planner and an automated framework to derive model variables from open transport datasets. Chapter 6 then describes the calibration and appraisal of a range of station choice models that have the potential to be incorporated into both trip end and flow rail demand models. Chapter 7 then proposes a methodology for integrating a station choice component into a trip end model, and explains its use to calibrate a national-scale trip end model for GB. In Chapter 8 a methodology to forecast station demand using the national-scale model is proposed, and it is then applied to forecast demand for a number of recently opened stations, including a new railway line. The predictive performance of the models, using both deterministic and probability-based catchment definitions, is then assessed and compared with official scheme forecasts. Finally, in Chapter 9, the outcomes of the research project are summarised, and several areas for potential future work are identified.

Chapter 2

Forecasting demand for new railway stations: the status quo

2.1 Introduction

This chapter begins by considering the state of rail demand in the UK today, and the growth in stations and lines that is expected to continue into the future (Section 2.2). This is followed by a review of the established approaches to modelling demand for new stations, concentrating on the aggregate models that are typically adopted in the UK (Section 2.3). In Section 2.4, the methods used to define station catchments in these aggregate models are examined, focussing on their ability to produce realistic representations of actual station catchments. The accuracy of demand forecasts produced by aggregate models during the scheme appraisal process is then assessed in Section 2.5. The chapter closes by summarising the main findings and drawing some conclusions (Section 2.6).

2.2 Rail demand today

Rail passenger demand can be measured in terms of the number of passengers who choose to travel by train, rather than using an alternative transport mode or not travelling at all. Where passengers choose to travel from and to, and what route they decide to take, will in turn determine the level of demand generated by, or attracted to, each railway station on the rail network. The level of rail passenger demand can be influenced by a range of factors that are either outside of the control of the rail industry (exogenous factors) or within the control of the rail industry (endogenous factors). Some examples of these factors are shown in Table 2.1.

Exogenous factors	Endogenous factors
Gross domestic product (GDP) of a country or region	Rail fares
Level of employment	Punctuality
Population	Reliability
Levels of private car ownership and operating costs	Station facilities
Availability and costs of other public transport modes	Service frequency
Travel time of other modes (e.g. effect of congestion)	Journey time
Integration of rail with other modes	Level of crowding on trains
	Station location
	New stations or new lines

TABLE 2.1: Examples of exogenous and endogenous factors that may influence rail passenger demand. Based on information provided in the PDFH (Association of Train Operating Companies, 2013).

In Great Britain, travel by rail has experienced something of a resurgence in recent decades, with a rapid growth in passenger journeys replacing the declines of the 1960s and 1970s and the modest growth of the 1980s. The annual number of journeys has more than doubled over the past 20 years, as shown in Figure 2.1. The average annual growth in passenger journeys was 3.95% between 1997/98 and 2016/17, compared to 0.54% between 1980 and 1996/97[1] (Office of Rail and Road, 2017). During the last two decades growth in rail travel substantially out-paced growth in GDP, with the number of passenger journeys rising 104%, while GDP increased by only 48%. This is the reverse of the relationship seen in the previous 20-year period, when GDP rose 59% and passenger journeys increased by only 14%[2]. While rail use has been increasing over recent years, travel by other modes has been falling. For example, in England the total number of trips made by rail increased by 56% between 2002 and 2016, but trips by car/van and bus fell by 13% and 19% respectively (Department for Transport, 2017b).

The rail network has also been expanding, with more than 100 stations either reopened or newly built since the privatisation process was completed in April 1997, including 56 in the past ten years alone (see Figure 2.2) (Alderson & McDonald, 2017). Many more stations are currently under construction, proposed or being campaigned for by local communities (Railfuture, 2018). New lines have been built, ranging from local services such as the Borders Railway between Edinburgh and Tweedbank which opened in 2015 with seven new stations, to major infrastructure projects such as High Speed 1 between London and the Channel Tunnel which fully opened in 2007. Further new or extended lines of both local and national significance are currently being planned or actively considered (RailEngineer, 2016), and there are local campaigns seeking to get former rail lines across the country re-opened

[1] Annual passenger journeys were reported by calendar year from 1950 to 1984 (inclusive), and by financial year from 1985/86.

[2] GDP data obtained from `https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/abmi/qna`. Q1 2017 compared with Q2 1997; and Q1 1997 compared with Q4 1977.

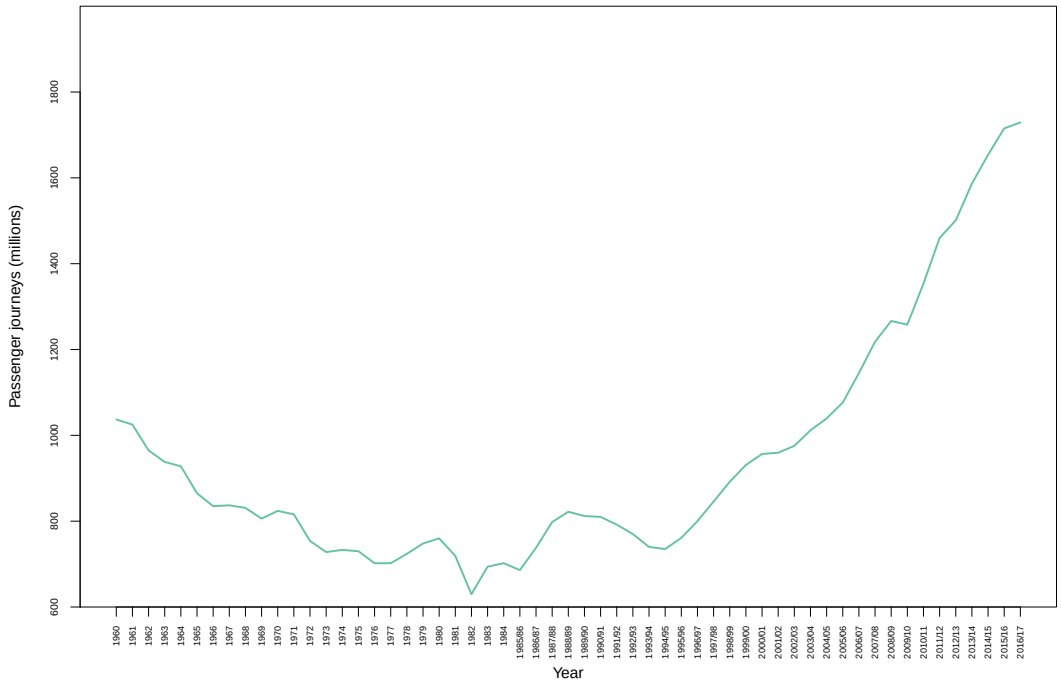


FIGURE 2.1: Annual passenger journeys by rail in Great Britain over the period 1960–2016/17. Based on data provided by Office of Rail and Road (2017).

(Campaign for Better Transport, 2017). Figure 2.3 (RailEngineer, 2016) shows the new passenger lines that have opened since 2000, and those currently under development, or at various stages of consideration (as at May 2016).

Against this backdrop, the potential to meet local or regional transport needs, and also economic growth objectives, by investing in new rail stations, routes or services, is increasingly being recognised by UK local authorities, Passenger Transport Executives and Local Enterprise Partnerships (Department for Transport, 2011). However, in order to assess whether a particular scheme will achieve the required objectives, it is necessary to produce accurate forecasts of the effect on demand of any such proposal. In the next section, the modelling techniques currently available to produce such forecasts will be discussed.

2.3 Modelling station demand

The main source of advice on passenger demand forecasting for the rail industry in Britain is the Passenger Demand Forecasting Handbook (PDFH), which is maintained and developed by the Passenger Demand Forecasting Council (PDFC) (Association of Train Operating Companies, 2013). The PDFC consists of the train operating companies, Network Rail, DfT, Transport Scotland, Office of Rail and Road (ORR), Transport for London (TfL), the Urban Transport Group, the Rail Safety and Standards Board, HS1, HS2, Rail North and the Welsh Government. The most recent version of the PDFH (v5.1) was published in April 2013, and is said to summarise 'over twenty years of research on rail demand forecasting' and be

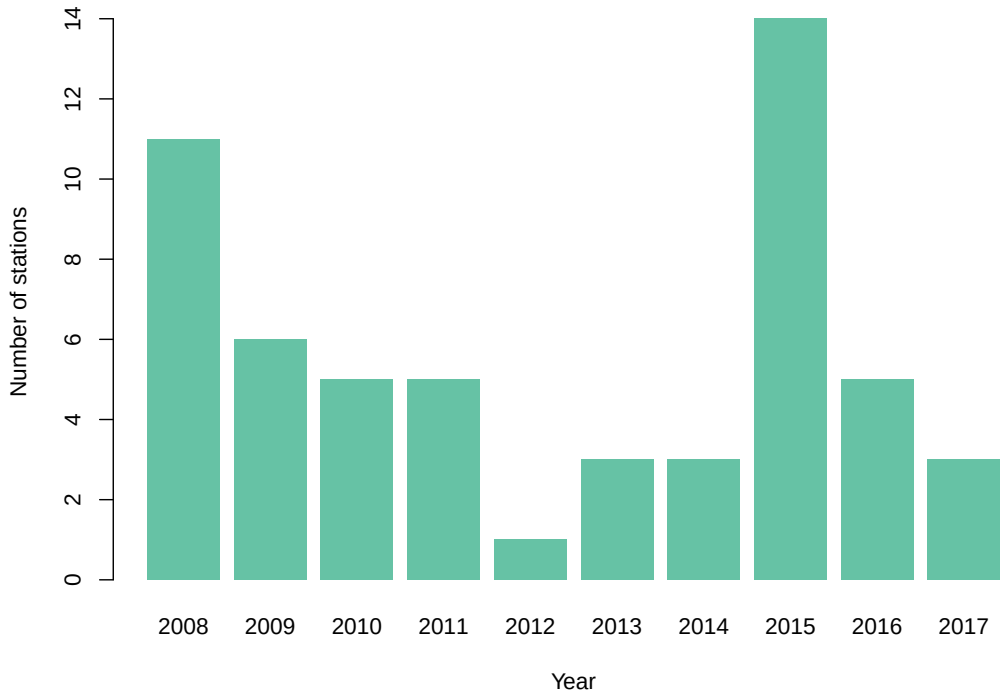


FIGURE 2.2: Number of new railway stations opened in Britain on the national rail network for each of the last ten years. Based on information provided in Alderson and McDonald (2017).

'recognised within the industry as the key source of evidence in this area' (Association of Train Operating Companies, 2015).

The primary focus of the PDFH is on the elasticity-based approach to forecasting demand which forms the core of the industry's model, MOIRA, which is overseen by the PDFC. MOIRA contains two sets of timetable data, one for the base year and one for the forecast year, with the latter incorporating all the service changes that are to be modelled. Other key data inputs to MOIRA are the Latest Earnings Networked Nationally Overnight (LENNON) database which contains details of all tickets sold in the base year, and the EDGE database that holds forecasts of the drivers of exogenous demand (Worsley, 2012). The forecast number of journeys (J_{new}) between a pair of zones can be given by the following:

$$J_{new} = I_E I_P I_T J_{base}, \tag{2.1}$$

where I_E, I_P, I_T are indexes representing the proportionate increase in journeys due to external factors, fares and journey times respectively, and J_{base} is the journey data for the base year obtained from LENNON. I_E and I_P are both composed of a range of variables with different elasticities, for example a simplified I_E incorporating variables for GDP and population would take the following form:

$$I_E = \left(\frac{GDP_{new}}{GDP_{base}}\right)^g \times \left(\frac{POP_{new}}{POP_{base}}\right)^p, \tag{2.2}$$

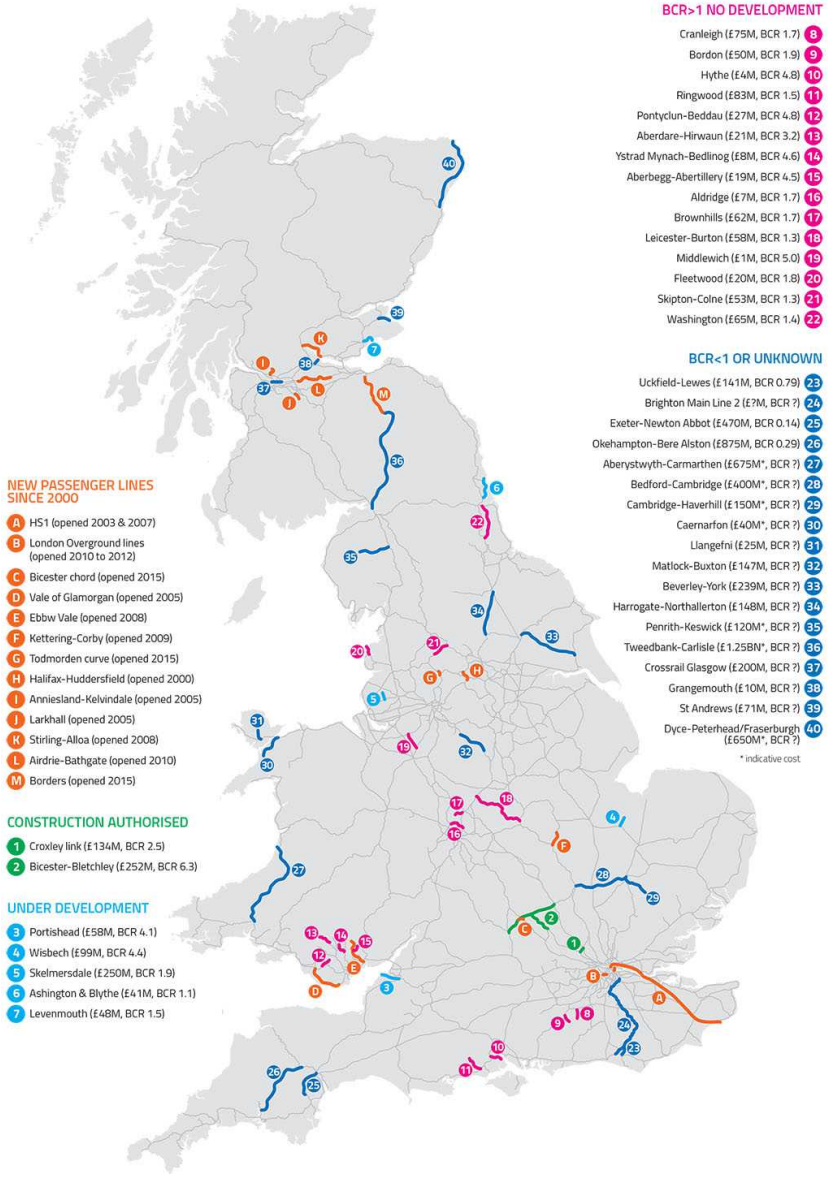


FIGURE 2.3: New passenger rail lines opened in Britain since 2000, currently under development or at various stages of consideration (as at May 2016). Note: Reprinted from 'After Borders, what next?', by RailEngineer, 2016, April 22. Image reproduced with permission of the rights holders, Graeme Bickerdike and Rail Media.

where g and p are the elasticities of GDP and population respectively. I_T consists of a single variable known as generalised journey time (GJT), and its associated elasticity. GJT is the sum of at least three elements: station to station time in minutes; a penalty for service frequency expressed in equivalent minutes of journey time; and a penalty for interchange expressed in equivalent minutes of journey time. Other factors can be incorporated within GJT, for example crowding, and these will also be expressed in terms of equivalent minutes of journey time (Association of Train Operating Companies, 2013). The elasticity values for each variable that are used in MOIRA are predominantly based on analysis of time series data and are discussed in detail in the PDFH.

It can be seen that this approach is an incremental one, forecasting how base demand will change as a result of changes in fare, service or external factors. This presents a problem when forecasting demand for a new station where there is no base demand, or when a significantly improved service is to be offered at a station where base demand is currently very low. The PDFH suggests that the elasticity-based approach is probably only appropriate for changes of up to 20% in explanatory variables. If the incremental approach is inappropriate, alternative modelling techniques are needed that can forecast the absolute level of demand, and the PDFH devotes two chapters to this subject, one outlining the types of model available and when they might be used, and a second discussing the available evidence on modelling absolute demand, derived from both industry and academic research. The types of model are discussed in the sections that follow.

2.3.1 Simple demand models

The simplest form of model to forecast absolute demand is the trip rate model, which considers station demand to be some function of the population of its catchment (Preston, 1991a). A trip rate model might take the following linear form, suggested by Blainey (2010) as an initial basic model:

$$V_i = \alpha + \beta P_i, \tag{2.3}$$

where V_i is the passenger entries and exits at station i, P_i is the catchment population of station i, and α and β are parameters to be estimated. Such a model would need to be calibrated based on existing stations, with the dependent variable the observed number of entries and exits over a time period (usually a year). The selection of comparable stations on which to estimate the model is critical for producing a successful trip rate model, as they are known to lack spatial transferability and are unlikely to be useful unless the demand forecast scenario is very similar to the one used to calibrate the model (Preston, 1991b). This is because they do not take account of the differences in relevant factors that influence trip rates, such as socio-economic characteristics of the catchment population, the level of train service at the station, or how attractive destinations are (Preston, 1991a).

The transferability of the trip rate model can be improved by incorporating a range of these additional explanatory factors into the model, and it then becomes known as a trip end model. A series of trip end models were developed by Blainey (2010) to forecast the number of trips made from local stations in England and Wales, with the following linear additive model an example:

$$V_i = \alpha + \beta \sum_{a}^{A} P_a w_a + \delta F_i + \lambda T + \tau Job_{i4} + \rho Pk_i, \tag{2.4}$$

where P_a is the resident population of census output area a, A is all output areas whose closest station by car travel time is station i, w_a is a distance decay function, F_i is the number

of trains calling at station i on a normal weekday, T is the distance in km from station i to the nearest category A–D station[3], Job_{i4} is the number of jobs located within four minutes drive of station i, Pk_i is the number of parking spaces at station i, and α, β, δ, λ, τ, and ρ are parameters.

2.3.2 Spatial interaction (flow) models

A weakness of the trip end model, in common with the trip rate model, is that it does not take account of the attractiveness of destinations. This requires a more complex spatial interaction model that is able to estimate passenger flows between origin station i and destination j for all origin-destination (OD) pairs. A spatial interaction model takes the following general form:

$$T_{ij} = f(V_i W_j S_{ij}), \tag{2.5}$$

where T_{ij} is the number of trips between origin i and destination j, V_i represents attributes of origin i (e.g. population), W_j represents attributes of destination j (e.g. number of work places), and S_{ij} represents the separation between origin i and destination j (e.g. distance) (Rodrigue, Comtois, & Slack, 2013). Preston (1991a) developed log-linear and semi-log direct demand models which were calibrated using data on 99 flows for small town, suburban and rural stations in West Yorkshire, and nine model variants of a multiplicative form were developed by Blainey and Preston (2010) based on some 2,400 flows for small stations in South Wales, with the following one of the basic model forms tested:

$$T_{ij} = \alpha \left(\sum_a P_a w_a \right)^{\beta} Job_{i4}^{\tau} Pk_i^{\rho} J_{ij}^{\delta} F_{ij}^{\eta}, \tag{2.6}$$

where T_{ij} is the number of trips made from station i to j, $P_a w_a$ is population weighted by distance for each census area in the station's catchment, Job_{i4} is the number of jobs located within four minutes drive of station i, Pk_i is the number of parking spaces at station i, J_{ij} is the average journey time for direct trains from station i to j, F_{ij} is the number of direct trains from station i to j on a normal weekday, and α, β, τ, ρ, δ and η are parameters to be estimated.

[3]Stations were divided into six categories (A – E) when the GB rail industry was privatised in 1996. Category A stations are national hubs, Category B are national interchanges, Category C are important feeder stations, Category D are medium staffed stations, and Category E and F are small stations, staffed and unstaffed respectively (Green & Hall, 2009).

2.4 But what about choice?

2.4.1 Aggregate models and catchment definitions

The models discussed in Sections 2.3.1 and 2.3.2 rely on data for some explanatory variables that is aggregated, and it is necessary to explicitly define the unit of aggregation before any models can be developed. The unit of aggregation is the station catchment, which will often be divided into a collection of zones that are used to aggregate relevant data, such as population or socio-economic characteristics. A variety of approaches have been adopted to define station catchments, while the zones can be defined by the researcher or an existing zone structure may be applied, for example one based on a national census data unit.

Preston and Aldridge (1991) included the population within a 2 km radius of each station in trip end models calibrated for 36 stations within the Greater Manchester area, and in subsequent direct demand models, Preston (1991b) divided the catchment into two radial zones, up to 0.8 km and 0.8 km to 2 km from the station (see Figure 1.1). In the USA a half mile circular catchment around a station is considered the 'de facto standard' for planning transit developments, and its outer circumference is intended to represent the distance that a passenger can walk at 3 mph and reach the station within 10 minutes (Guerra, Cervero, & Tischler, 2012). In a variation of the circular catchment, a series of concentric 'doughnut shaped' bands delimiting zones of population where travel time to the station is up to 4, 6, 8, 10, 15, 20, 30 and 45 minutes were used in a spatial interaction model developed by Wardman and Whelan (1999).

Another approach is to divide population into zones and then allocate each zone to its nearest station (see Figure 1.2). This was adopted by Blainey (2010) in several trip end models, where census output areas were assigned to the nearest station measured by road travel time. A more sophisticated variation of this method developed by Blainey and Preston (2010) created flow-specific catchments in spatial interaction models, on the premise that passengers will seek to minimise the total journey time from origin to destination. Each census output area was allocated to one of four alternative stations for each destination, on the basis of minimising total journey time. With this method, a station can have a different catchment (a different set of census output areas) for each destination, and while the model is deterministic in the choice of station, this choice can vary by destination.

The fact that choice of station is deterministic in these aggregate models gives rise to two implicit assumptions: first, any trip originating from a location will use the single station that has a catchment encompassing that location; and second, each station has a discrete catchment that does not overlap with any other station's catchment. The models do not allow for the possibility that a particular locality falls within the catchment of more than one station and that different passengers starting a trip from the same locality might choose different stations. Nor do they explicitly recognise that stations might be in competition,

with improved facilities or services at one station resulting in passengers being abstracted from another. It may be possible to introduce explanatory variables into a model to act as a proxy for choice or competition. For example, in a study of demand for small local stations, Blainey (2010) attempted to account for the effect on demand of proximity to larger stations by included the distance to larger (Category A–D) stations as a single variable. Although the variable was significant, its inclusion improved model fit, as measured by adjusted R-squared, by only 0.003.

2.4.2 Catchments in reality

It is important to understand whether the artificially constructed catchments described in the previous section are representative of real station catchments, and whether the implied assumptions hold. A number of studies have sought to explore this and some of their findings are reviewed in the sections that follow.

2.4.2.1 Radial catchments

Blainey and Evens (2011) used data for some 114,000 trip ends obtained from the National Rail Travel Survey (NRTS)[4] to investigate the extent to which observed station catchments correspond with the common catchment definitions used in aggregate models. They found that a 0.8 km radial catchment based on straight-line distance accounted for 40.7% of observed trips, increasing to 68.3% of trips for a 2 km catchment. When based on road network distance performance deteriorated, with the 0.8 km and 2 km catchments accounting for 32.9% and 65% of trips respectively. When the 2 km catchments were restricted to being non-overlapping (sometimes referred to as 'cropped'), there was a further reduction in performance, with only 57% of observed trips included. Furthermore, these average results masked considerable variations at the individual station level, with a 2 km buffer capturing 80–100% of trips for 34 stations, but only 0–20% of trips for seven stations.

2.4.2.2 Nearest station-based catchments

A number of studies have found that not all passengers choose to use their nearest station, a phenomenon commonly referred to in the UK as 'railheading'. For example, Debrezion, Pels, and Rietveld (2007a) reported that 47% of passengers in the Netherlands did not use their nearest station; Mahmoud et al. (2014) found that over 30% of cross-regional commuters who accessed a station by car did not choose the station closest to their home; and Chakour and Eluru (2014) observed that the nearest station was often not chosen, and

[4]The NRTS was a survey carried out by the Department for Transport to gather data on passenger rail trips in Great Britain on weekdays outside school holidays. The London and the South East area was surveyed during 2001, with Wales, Scotland and the remainder of England surveyed in 2004 and 2005.

in some cases not even the third closest station was selected. Using OD data from Dutch Railway Company customer satisfaction surveys aggregated by postcode area, Givoni and Rietveld (2014) ranked stations by the number of departures originating from each postcode area. They found that out of 83 postcode areas, in 56 the first ranked station was the nearest, but in 27 the first ranked station was not the nearest, and on average was 2.3 km further away.

Analysis reported in the PDFH, based on a large OD dataset of some 230,000 observations obtained from passenger surveys carried out in the 1980s and 1990s, showed that the likelihood of a passenger using their nearest station varied by journey purpose, with those on business or holiday trips the least likely to use their nearest station, and commuters the most likely. The surveys also revealed that around 50% of inter-city passengers did not use their nearest station, compared with just 20% of travellers in the South East where the network is very dense. Referring to the same research, Lythgoe et al. (2004) note that some parkway stations have an extremely high proportion of railheaders, such as Birmingham International at 92% and Bristol Parkway at 85%. Variation in realheading by station type has also been reported by Blainey and Evens (2011), who ranked stations for each individual based on access or egress distance and found that most passengers boarding at a Category A station (a national hub), had at least one other station closer to their origin or destination, while the vast majority boarding at a Category F station (small unstaffed) were using their nearest station.

Blainey and Preston (2010) carried out an OD survey on the Cardiff–Rhymney line in South Wales, with the primary aim of comparing theoretical catchments with observed catchments to inform the work on trip rate models discussed in Section 2.3.1. They found that only 53% of trip ends were located within catchments defined by assigning census areas to their nearest station based on theoretical road access time, with this improving to 63% when catchments were constructed using the flow-specific method. They also experienced problems with the catchment of some city centre stations, where low population densities resulted in large census areas with shapes that meant the main shopping area was not assigned to the nearest station, because the census area centroid was closer to another station.

2.4.2.3 Overlapping catchments

Several studies have produced visualisations of observed station catchments using geographical information systems, to explore the extent of catchment overlap. For example, Mahmoud et al. (2014) produced approximate station catchments, shown in Figure 2.4, based on observed choices by commuters in the Toronto area, Canada, and found 'substantial overlap' indicating that commuters living in the same locality make differing station choices. Fan et al. (1993) generated plots of observed station catchments by determining the trip origin furthest from the station for each 30 degree arc, and then joining the points together to

create polygons representing each station's catchment, and also found there was significant overlap between them.

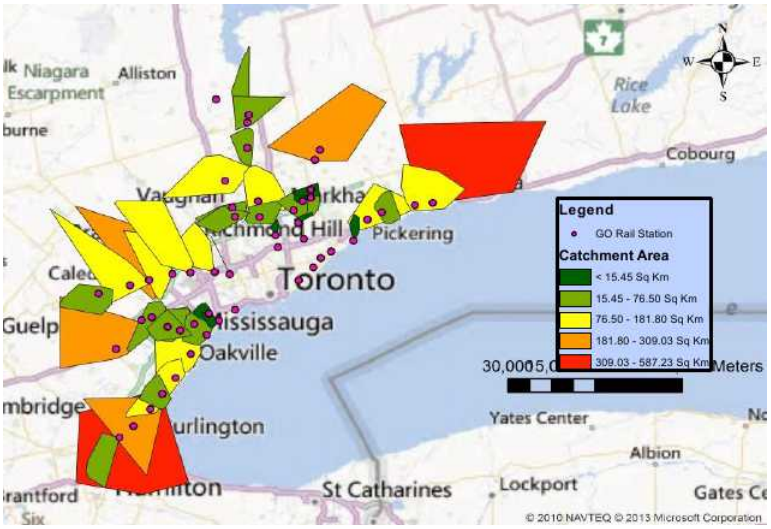


FIGURE 2.4: Observed park and ride catchment areas in Toronto, Canada. Note: Reprinted from '*Park-and-ride access station choice model for cross-regional commuter trips in the Greater Toronto and Hamilton Area*', by Mahmoud, M. S., Eng, P., & Shalaby, A., 2014, paper presented at Transportation Research Board 93rd Annual Meeting. Image reproduced with permission of the rights holder, SAGE Publications.

Adcock (1997) reported finding 'major instances' of station pairs having a common catchment area, and other studies have examined the homogeneity of station choice within catchment zones. For example, Desfor (1975) assigned census blocks to a particular station's catchment based on the majority observed choice of the commuters resident within each block, and if a block did not have a majority choice it was assigned to an unallocated category. This method resulted in a 71% correct allocation, and Desfor concluded that the 'concept of homogeneous non-overlapping market areas is a serious oversimplification'. Givoni and Rietveld (2014) ranked stations by the number of departures originating from each postcode area, and found that on average 71% of departures were from the top-ranked station and 18% and 9% were from the second and third ranked stations respectively, again indicating that choice is not homogeneous within zones.

2.4.2.4 Mode-specific catchments

It is intuitive to expect that station catchments will vary depending upon access mode, for example the catchment for walk access will be considerably smaller than the catchment for motorised vehicle access. For example, Blainey and Evens (2011) disaggregated access and egress distances by main mode for the North East region and found that the average distance was 1.6 km for walk mode, 8.2 km for bus mode, and 11.6 km for car mode. The catchment for public transport access will reflect the routes that serve a particular station, as was found

by Givoni and Rietveld (2014), where the presence of tram and metro lines clearly influence the shape of station catchments in Amsterdam.

2.4.2.5 Catchments by station type

The geographic size and shape of station catchments will also differ based on the type of station and its position within the rail network. Passengers travel further on average to stations that offer inter-urban or inter-regional services, where the access journey is a smaller component of the total journey, than they do to suburban stations that provide short services to or from a major urban centre or connectivity to the wider rail network. Stations built outside of urban centres with good road links that provide easy accessibility by car to mainline train services, so-called parkway stations, are known to have the longest average access distances (Lythgoe & Wardman, 2004). Some stations are particularly well-connected to the rail network, providing direct services on several routes to many destinations, whilst others offer the only access to the rail network for large geographic areas that have no rail service, perhaps as the result of branch-line closure in the 1960s. In both cases these stations would be expected to have larger catchments, for both access and egress journeys, than stations that are poorly connected or in areas that are well served by a dense station network.

2.5 Does choice matter — are existing models good enough?

Having established that the catchment definitions used in the aggregate demand models are unlikely to be a realistic representation of real-world station catchments, it is important to consider the extent to which this deficiency might be impacting the ability of these models to produce accurate demand forecasts.

The main source of information on the performance of station demand forecasts in recent years is the 'Station Usage and Demand Forecasting for Newly Opened Railway Lines and Stations' report produced by Steer Davies Gleave (2010). This was commissioned by the Department for Transport, reflecting a general concern about the perceived poor performance of station demand forecasts. Although the report sought to examine the 40 stations that had opened since privatisation of the industry in 1997, demand forecast information was only available for 27 stations. The modelling methodologies adopted for individual stations, and for three lines, are shown in Figure 2.5 (Steer Davies Gleave, 2010, p. 15)[5]. Of the 16 stations or lines where the methodology was known, a 'trip rate' model was used in ten cases, a mode choice (logit) model was the predominant approach in three cases, and a four-stage strategic model was used in two cases. The trip rate approach clearly dominates, being used

[5]The Ebbw Vale line includes Ebbw Vale Parkway, Llanhilleth, Newbridge, Crosskeys, Risca & Pontyminster and Rogerstone stations; Edinburgh Crossrail line consists of Brunstane and Newcraighall stations; Larkhall to Milngavie line includes Larkhall, Merryton and Chatelherault stations; and the Vale of Glamorgan line includes Llantwit Major and Rhoose stations.

TABLE 3.1 SUMMARY OF DEMAND FORECASTING METHODOLOGY

New Station/Line	Methodology Used	Abstraction modelled?	Exogenous growth modelled?	Extent of documentation
Alloa	No Information supplied	Unclear	Unclear	None provided
Aylesbury Vale Parkway	Trip rate and accessibility modelling (using HEXs)	Yes	Yes	Good
Chandlers Ford	Logit model, trip rate model and MOIRA	Yes	Yes	Good
Coleshill Parkway	Trip rate model and logit mode choice	Unclear	Yes	No description of demand modelling
Corby	Trip rate, MOIRA and station access model	Yes	Yes	Good
East Midlands Airport Parkway	GIS catchment analysis, elasticity based model & airport mode share assumptions	Yes	Yes	Good
Ebbw Valley Line	Logit model and uplift for trip generation	N/A	Yes	Reasonable
Edinburgh Crossrail	No Information supplied	N/A	Unclear	No description of demand modelling
Edinburgh Park	Trip rate and logit mode choice	Yes	Yes	Rather poor
Glasshoughton	Trip rate	Unclear	Yes	No description of demand modelling
Imperial Wharf	RAILPLAN strategic forecasting model	Yes	Yes	Good
Larkhall-Milngavie	4 stage land use model	Yes	Yes	Good
Laurencekirk	Trip rate	Partially	No	Reasonable
Liverpool South Parkway	Elasticity based model, airport accessibility model, mode switch (logit) model	Yes	Yes	Good
Mitcham Eastfields	Trip rate	Yes	Yes	Good
Shepherds Bush	Trip rate	Yes	Yes	Good
Vale of Glamorgan Line	Trip rate	Unclear	Yes	Poor
Warwick Parkway	Parkway Access Model and Mode/Route Choice models	Yes	Yes	Good

FIGURE 2.5: Summary of modelling methodology used to forecast demand for new stations. Note: Reprinted, with highlights added to identify trip rate models, from '*Station usage and demand forecasts for newly opened railway lines and stations*', by Steer Davies Gleave, 2010, p. 15. Reproduced under the Open Government Licence v3.0.

to assess two-thirds of the schemes. Detailed information about the nature of the trip rate models is not provided in the report, although it is noted that they varied in complexity and additional explanatory variables were present in some cases (and should therefore more accurately be referred to as trip end models). Only one example of the method used to define the station catchment is given, for Mitcham Eastfields, where the population centroids of census Enumeration Districts were assigned to their closest station.

Table 2.2 compares forecast demand and actual demand for the reviewed stations, based on data published in the Steer Davies Gleave report[6]. The bar chart in Figure 2.6 relates only to the stations where a trip rate model was used, and shows the percentage difference between actual and forecast demand. In only three cases was observed demand within 20% of the forecast. The forecast was particularly poor for Glasshoughton, where observed demand was 2.65 times higher; Edinburgh Park, where observed demand was 1.8 times higher; and Aylesbury Vale Parkway, where observed demand was less than half that expected. The report does suggest several reasons that might, at least partly, explain the poor performance for these stations: no attempt was made to forecast demand generated by a leisure complex at Glasshoughton; Edinburgh Park may have abstracted demand from South Gyle; and planned housing development near Aylesbury Vale Parkway did not materialise due to the 2007–2008 financial crisis.

A more recent, and widely publicised, example of inaccurate station demand forecasts is the new Borders Railway line in Scotland, which opened in 2015 with seven new stations. As shown in Figure 2.7, the final scheme appraisal severely under-forecast demand at the three

[6]Note that Llantwit Major and Rhoose stations are combined; and Corby and Laurencekirk stations are excluded because actual demand data was not available (they opened during 2009).

New Station	Forecast	Actual	Difference
Aylesbury Vale Parkway	29000	13066	-55%
Brunstane	129920	121758	-6%
Newcraighall	467600	176975	-62%
Chandlers Ford	290237	236145	-19%
Ebbw Vale Parkway	45858	252607	451%
Crosskeys	62982	67347	7%
Newbridge	82951	115733	40%
Risca and Pontyminster	105412	101624	-4%
Rogerstone	58087	71041	22%
Llanhilleth	37529	40967	9%
Imperial Wharf	437760	256000	-42%
Liverpool South Parkway	640652	465324	-27%
Mitcham Eastfields	179115	239040	33%
Shepherds Bush	922717	1219167	32%
Alloa	120000	335687	180%
Warwick Parkway	201000	238654	19%
Glasshoughton	50989	135279	165%
Llantwit Major + Rhoose	395650	401192	1%
Edinburgh Park	209619	382823	83%
Coleshill Parkway	119000	98903	-17%
Larkhall	276993	334015	21%
Chatelherault	48399	40922	-15%
Merryton	215191	99500	-54%

TABLE 2.2: Forecast and observed demand for new stations, produced from data published in Steer Davies Gleave (2010).

Scottish Borders stations (Tweedbank, Galashiels and Stow), with actual demand in the first 12 months up to eight times higher than forecast, while over-predicting demand, to a lesser extent, at the four Midlothian stations (Transport Scotland, 2017). There is only limited information publicly available on the models used to generate these forecasts. It is known that two methods were used: a stated preference survey[7] of residents living near the line (which was presumably used to estimate the switch to rail from other modes); and a trip rate model that applied 'generic trip rates' to the 'population within a defined area' (Transport Scotland, 2012). The stated preference approach produced a higher estimate than the trip rate model and the mid point between the two models was used as the demand estimate. The forecasts shown in Figure 2.7 are therefore higher than they would have been if the trip rate approach alone had been used.

[7] Stated preference surveys elicit what individuals *say* they would do under hypothetical choice situations and can be unreliable for forecasting purposes. See Section 3.4.1 for a more detailed discussion.

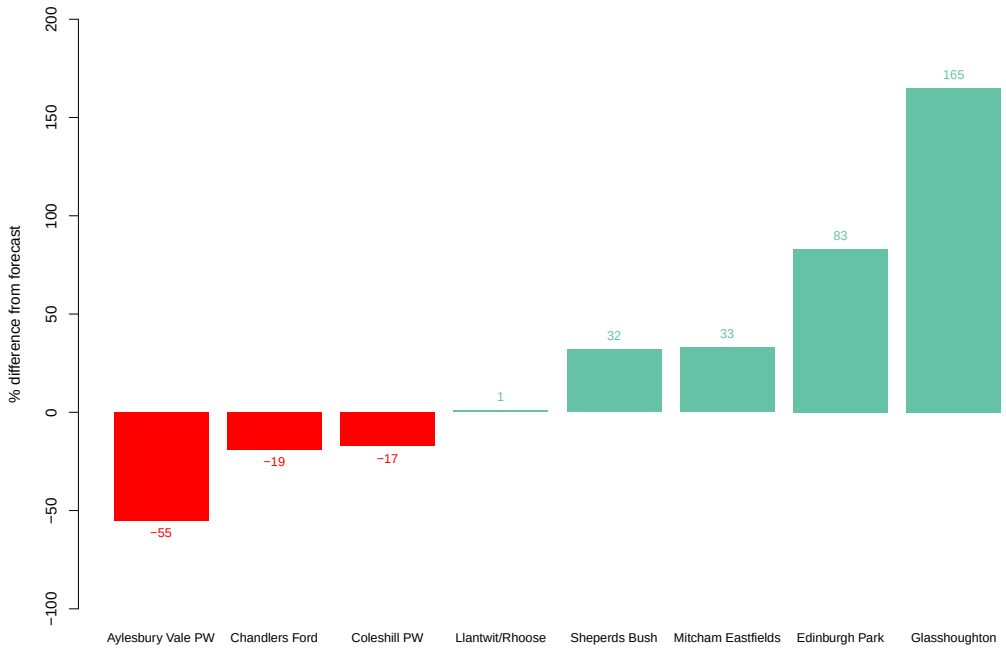


FIGURE 2.6: Difference between forecast demand and actual demand for those stations where a trip rate model was used. Based on figures provided in Steer Davies Gleave (2010).

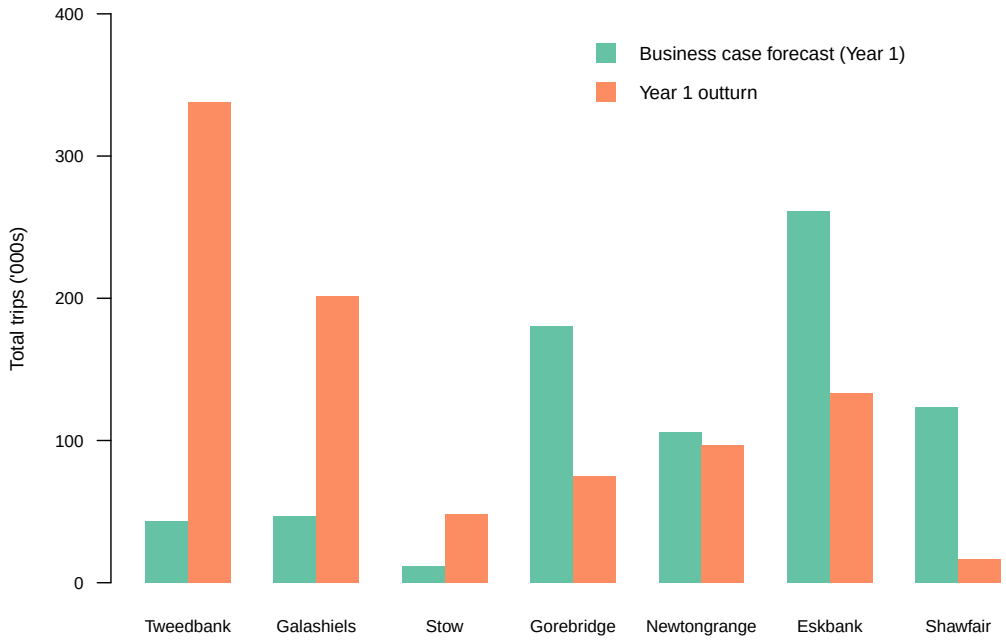


FIGURE 2.7: Final business case demand forecast for stations on the Borders Railway (first 12 months of operation) compared with actual station usage.

The Borders Railway example serves as a good illustration of the potential implications of inaccurate forecasts, as demand is a crucial driver of scheme benefits. The low benefit-cost ratio of 0.5:1, which was revealed when the final business case document was released (Transport Scotland, 2012), led the scheme to be described as 'one of the worst-performing major transport projects to be funded in recent times' (Local Transport Today, 2013). Such a low benefit-cost ratio could have resulted in the scheme not gaining approval, and appears to have resulted in a less ambitious scheme than originally envisaged. For example, the length of dynamic loop (where two trains can pass without having to stop) was reduced by 6.5 miles, limiting the ability to run more services and potentially causing service reliability issues; and the specified width of road bridges over sections of single track was reduced, precluding the cost-effective provision of additional double-track in the future should growth in demand warrant it (Spaven, 2017).

These findings confirm that the models being used to forecast demand for new stations are in many instances not producing accurate forecasts, and in some cases the discrepancies are so large that they could undermine the case for an otherwise viable scheme (due to under-forecasting), or result in a new station being built that fails to deliver the expected economic and societal benefits (due to over-forecasting). It is not possible to conclude that the simplistic method of defining catchments is the primary cause of the poor predictive performance of trip end models. However, if station catchments are not correctly defined, then inappropriate weight will be given to other explanatory variables, such as service quality measures, as drivers of trip generation, rather than the catchment population. By defining more realistic catchments, the parameter estimates will be more robust, and the models will be more transferable (Wardman & Whelan, 1999). If trip end models with greater geographic transferability can be calibrated, then this increases the likelihood that an improved nationally applicable model can be developed, building on the previous work by Blainey (2010) to develop a national trip end model with deterministic catchments. This would reduce the need for local solutions, which will inevitably vary in approach, robustness and performance. Even if a model specific to a local context was considered desirable, the national model would be available to act as a sense-check of the demand forecasts. In formulating an appraisal framework for new local railway stations, Blainey and Preston (2013b) recognised the shortcomings of the simplistic catchment definition methods and acknowledged that station choice models are likely to produce more accurate catchments. These were not incorporated into the appraisal procedures, due to their added complexity and the absence of a calibrated station choice model for the whole country. However, the authors note that their inclusion is crucial if the reliability and transferability of the framework are to be increased.

2.6 Conclusions

The number of passenger journeys made by rail is expected to continue rising in the years ahead, potentially increasing 40% by 2040 (Network Rail, 2018), and the UK Government

has recently set out an ambition of 'reversing the historic contraction of the rail network' with an emphasis on new local connections and stations that support housing development or economic growth, or that address urban congestion (Department for Transport, 2017a). There will, therefore, be a continuing need to assess proposals for new railway stations and lines. A crucial part of this evaluation process is generating accurate demand forecasts, as predicted station patronage is a key driver of the benefits that will determine whether or not a scheme is considered viable. However, this chapter has shown that the aggregate models that are most commonly used to forecast demand do not always perform well, and a contributory factor could be the relatively simple way that station catchments are defined. These do not represent the true nature of station catchments, and the implicit assumptions — that catchments do not overlap and stations do not compete for passengers — do not hold in reality. This suggests that passenger demand forecasting models might be improved if a probabilistic station choice element could be incorporated into them. In a trip end model, for example, this would allow the population in a zone to be weighted by the probability of a particular station being chosen by a rail passenger in that zone, with each zone having a probability for each competing station (see Figure 1.4). This would then allow a probabilistic catchment for each station to be generated (see Figure 1.5). In a flow model, the population in a zone could be weighted by the probability of a particular station being chosen, given the journey destination. In such a model, each zone would have a probability for each competing station, conditional on each journey destination. By incorporating more realistic station catchments into these models they should become more transferable, enabling an improved national model for GB to be developed. To achieve this goal, an appropriate model to forecast station choice at the zonal level will be required.

The next chapter will consider the body of prior station choice research, reviewing the modelling approaches adopted, the factors found to influence station choice, and previous efforts to incorporate a station choice element into rail demand models. This review will inform the subsequent research that will seek to develop station choice models and devise a methodology for incorporating them into the aggregate rail demand models.

Chapter 3

Railway station choice modelling: methods and evidence

3.1 Introduction

This chapter provides a comprehensive review of previous station choice research. It begins with a brief history of prior research to set the scene (Section 3.2), before moving on, in Section 3.3, to consider the theoretical basis of discrete choice models alongside their application in the field of station choice modelling, including approaches to validation and testing. Section 3.4 then looks at issues relating to obtaining and preparing data on observed choice, as well as the approaches taken to define choice sets. The factors that might explain observed station choice, how these have been selected and measured, and what influence they have been found to have on decision makers, are discussed in Section 3.5. How station choice models have been used in the context of rail passenger demand forecasting is then considered in Section 3.6. Finally, the conclusions drawn from the body of previous work are presented in Section 3.7.

3.2 A brief history of station choice modelling

The earliest published examples of station choice research date back to the mid 1970s in North America. Liou and Talvitie (1974) modelled access mode and station choice in the Chicago area of the Illinois Central Railroad using a sequential multinomial logit (MNL) approach which pre-dated the formal development of the nested logit (NL) model. Desfor (1975) used binary probit and weighted linear regression to explore choice between station pairs on the Lindenwold high-speed line (PATCO) - a rapid transit system predominantly based around the park and ride concept which carries commuters into the Philadelphia central business district.

No further work in this area has been found until a study in Japan modelled main travel mode, access mode and station choice before and after the opening of a new station on the Yokosuka line near Tokyo (Harata & Ohta, 1986). This appears to be the first station choice study to implement the NL model. In subsequent research, the NL model was adopted in several studies of joint access mode and station choice (Davidson & Yang, 1999; Debrezion, Pels, & Rietveld, 2009; Fan et al., 1993; Givoni & Rietveld, 2014) while the MNL model has been used for modelling station choice alone (Adcock, 1997; Blainey & Evens, 2011; Debrezion et al., 2007a; Kastrenakes, 1988; Mahmoud et al., 2014). Lythgoe and Wardman (2002) extended a direct-demand model for parkway stations to include a station choice element using nested logit and later enhanced this model by widening its applicability to shorter journeys and by developing a form of cross-nested logit (CNL) model (Lythgoe et al., 2004). In an unusual approach, Chakour and Eluru (2014) proposed a latent segmentation model where the observations are split using binary logit into those assumed to choose the station first or access mode first, with mode choice and station choice modelled using MNL in the order determined by this segmentation.

Within the last few years researchers have begun to develop models using more complex, open-form, discrete choice models which must be estimated using simulation techniques. These include work by Chen et al. (2014) to develop a framework for modelling park and ride station choice under uncertainty (or risk) and to model station choice specifically under parking search time uncertainty (Chen et al., 2015); a random parameter mixed logit (ML) model of park and ride lot choice (Pang & Khani, 2018); and an error components ML model of station choice that accounts for the unobserved spatial correlation between pairs of alternatives (Weiss & Habib, 2017). There has also been a willingness to consider alternatives to the utility maximisation behavioural assumption that underlies the vast majority of discrete choice models, with Sharma, Hickman, and Nassir (2017) developing models of parking lot choice based on the random regret minimisation approach proposed by Chorus (2012).

In addition to academic research, models have been developed, usually by consultancy firms, for use in central or local government transport models and as part of specific rail development proposals. Examples in the UK include: a binary logit model for West Coast Main Line track access assessment (MVA Consultancy, 2011); an MNL model to assess the demand for and benefits of High Speed 2 (Atkins Limited, 2011); and incorporating a station choice element into regional transport models (Fox, 2005; Fox et al., 2011). A summary of prior station choice research is given in Table 3.1

3.3 Application of discrete choice models

The basis of discrete choice models is that an individual can choose from a number of alternatives which are collectively known as the choice set. Three characteristics of the alternatives are assumed: the decision maker must choose only one (i.e. they must be

Author	Country	Focus of study	Main statistical approach	Data type	Survey type	Survey size
Liou and Talvitie (1974)	USA	AM/SC	Sequential MNL	Disaggregate	RP	3,694
Desfor (1975)	USA	SC	Probit	Disaggregate	RP	150
Harata and Ohta (1986)	Japan	AM/SC	NL	Disaggregate	RP	1,358
Kastrenakes (1988)	USA	SC	MNL	Aggregate	RP	26,000
Fan et al. (1993)	Canada	AM/SC	NL	Disaggregate	RP	1,824
Adcock (1997)	UK	SC	MNL	Disaggregate	RP	230,000
Wardman and Whelan (1999)	UK	AM/SC	NL	Disaggregate	RP & SP	32,525
Davidson and Yang (1999)	USA	AM/SC	NL	Disaggregate	RP	11,000
Lythgoe and Wardman (2004)	UK	SC	NL	Aggregate	n/a	n/a
Lythgoe et al. (2004)	UK	SC	Cross-NL	Aggregate	n/a	n/a
Fox (2005)	UK	SC (PnR)	NL	Disaggregate	RP	8,508
Debrezion et al. (2007a)	Netherlands	SC	MNL	Aggregate	RP	unknown
Debrezion et al. (2009)	Netherlands	AM/SC	NL	Aggregate	RP	unknown
Blainey and Evens (2011)	UK	SC	MNL	Disaggregate	RP	113,518
Atkins Limited (2011)	UK	SC	MNL	n/a	n/a	n/a
Fox et al. (2011)	Australia	SC (PnR)	NL	Disaggregate	RP	unknown
MVA Consultancy (2011)	UK	SC	Binary logit	Disaggregate	RP	3,957
Chakour and Eluru (2014)	Canada	AM/SC	Latent segmentation	Disaggregate	RP	3,902
Chen et al. (2014)	Australia	SC	ML	Disaggregate	n/a	n/a
Givoni and Rietveld (2014)	Netherlands	AM/SC	NL	Disaggregate	RP	8,491
Mahmoud et al. (2014)	Canada	SC	MNL	Disaggregate	RP	2,297
Chen et al. (2015)	Australia	SC	ML	Disaggregate	SP	600
Sharma et al. (2017)	Australia	PL (PnR)	MNL (RUM vs RRM)	Disaggregate	RP	2575
Weiss and Habib (2017)	Canada	SC (PnR/KnR)	ML	Disaggregate	RP	2,807
Pang and Khani (2018)	USA	PL (PnR)	ML	Disaggregate	RP	418

SC = station choice; AM = access mode choice; PL = parking lot choice; PnR = Park and ride; KnR = Kiss and ride; RP = revealed preference; SP = stated preference, RUM = random utility maximization; RRM = random regret minimization

TABLE 3.1: Summary of published research into railway station choice.

mutually exclusive); all alternatives must be included; and there must be a finite number of alternatives[1].

Discrete choice models are usually based on the assumption of utility maximisation, and are then known as random utility models (RUMs). An individual obtains utility from each alternative in the choice set and will choose the alternative that provides them with the maximum utility. The researcher does not know the perceived utility of each alternative, that is only known by the individual. The researcher attempts to measure the utility by identifying attributes of the alternatives and/or of the individual. That part of the utility that the researcher does not know is called the unobserved portion of utility and is treated as a random (stochastic) component. The utility that an individual obtains from an alternative can therefore be expressed using the following formula:

$$U_{ni} = V_{ni} + \varepsilon_{ni}, \tag{3.1}$$

where U_{ni} is the utility for individual n of alternative i, V_{ni} is the utility measured by the researcher, and ε_{ni} is the unobserved portion of utility. In practice V, which is known as the representative or deterministic component of utility, will be a function consisting of the selected attributes of the alternatives and the individual and their respective coefficients (or parameters). The function is commonly linear-additive in parameters and the representative utility for individual n of alternative i can be given by

$$V_{ni}(\boldsymbol{X}, \beta) = \sum_{k=1}^{K} \beta_k X_{kni}, \tag{3.2}$$

where $\boldsymbol{X}$ is a matrix of attributes and β is a vector of parameters of those attributes. The parameters, if unknown, are obtained statistically, for example by maximum likelihood estimation.

If faced with a choice set J then an individual will choose alternative i when:

$$U_{ni} > U_{nj} \quad \forall j \neq i. \tag{3.3}$$

However, as there is an unknown component to the utility it is not possible to say for certain what alternative an individual will choose, it is not deterministic. The probability of individual n choosing alternative i is:

$$\begin{aligned} P_{ni} &= Prob(U_{ni} > U_{nj} \quad \forall j \neq i) \\ &= Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \quad \forall j \neq i) \\ &= Prob(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \quad \forall j \neq i). \end{aligned} \tag{3.4}$$

For example, suppose an individual is choosing between two railway stations and the representative utility of station i is 5 and station j is 4. Although station i has the highest

[1] This introduction to discrete choice models, and the notation that follows, is largely based on Train (2009).

observed utility, it cannot be assumed that the individual will choose this station as the impact of unobserved factors on utility is not known. Station j would instead be chosen if its unobserved utility is more than 1 unit greater than the unobserved utility of station i. The probability of the individual choosing station i is therefore the probability that $\varepsilon_j - \varepsilon_i < 1$.

Assumptions made about the characteristics of the unobserved portion of utility will determine what form of statistical model is appropriate to calculate the probability of an individual choosing a particular alternative. If the unobserved portion of utility is assumed to follow an independent and identically extreme value (Gumbel) distribution (IIGD) then logit or NL models are suitable. These are closed-form models where the choice probabilities can be calculated exactly. For probit models the unobserved portion of utility is assumed to follow a multivariate normal distribution, and for mixed logit it is assumed to consist of two parts, one of which follows the Gumbel distribution and the other which follows a distribution that is specified by the researcher. Both probit and ML models are open-form and choice probabilities are approximated by simulation (Train, 2009).

3.3.1 Binomial and multinomial logit

Binomial logit is the simplest RUM-based discrete choice model which is used when there are only two alternatives under consideration. Based on the assumption that the stochastic utility component follows an IIGD, the probability of choosing alternative i over j can be calculated using the following derived equation:

$$Pr_{(i)} = \frac{e^{V_i}}{e^{V_i} + e^{V_j}}, \tag{3.5}$$

and as the sum of the probabilities for the two alternatives must equal one, the probability of alternative j is:

$$Pr_{(j)} = 1 - Pr_{(i)}. \tag{3.6}$$

The binomial logit model has seldom been used in modelling station choice, as in most research the number of alternatives in the choice set exceeds two. However, it has been applied in work carried out for the Office of Rail Regulation to assess applications made by Open Access Operators to run services on the two main lines running between London and Scotland (MVA Consultancy, 2011; Prior et al., 2011). These new operators planned to provide certain stations with direct services to major destinations such as London where these services did not currently exist. As part of a wider modelling framework a station choice model was developed to assess the extent to which passengers might be abstracted from the current 'primary' stations to these 'secondary' stations as a result of the service improvement. Pairs of primary and competing secondary stations were identified and binomial logit models used to forecast the proportion of passengers choosing each station in the pair under two different fare structures (walk-up fare and advanced fare). The station (dis)utility was

represented in the model by a single composite term: the GJT from the trip origin to the destination station. The components of GJT were: weighted access time to departure station by car; GJT obtained from MOIRA (which includes in-vehicle time, a frequency penalty, and an interchange penalty); fare; and car park cost. A spread parameter[2] for the model was estimated using data from the NRTS by taking the difference in station access time to a primary station (Preston) and four nearby potential secondary stations, plotting this against the percentage of passengers that chose the primary station, and then fitting a logit curve (see Figure 3.1). The model observations were trips from each primary station obtained from the NRTS, adjusted with an expansion factor to represent actual demand at the station. Using this model, the probability of choosing primary station p over secondary station s can be shown as:

$$Pr_{(p)} = \frac{e^{-\gamma GJT_p}}{e^{-\gamma GJT_p} + e^{-\gamma GJT_s}}, \tag{3.7}$$

where γ is the spread parameter. There are potential weaknesses with this approach as the spread parameter calibration assumes that the observed choice behaviour can be explained purely by differences in access distance and that the sensitivity of passengers to changes in access distance will remain similar when the competing stations are offering direct services on the mainline.

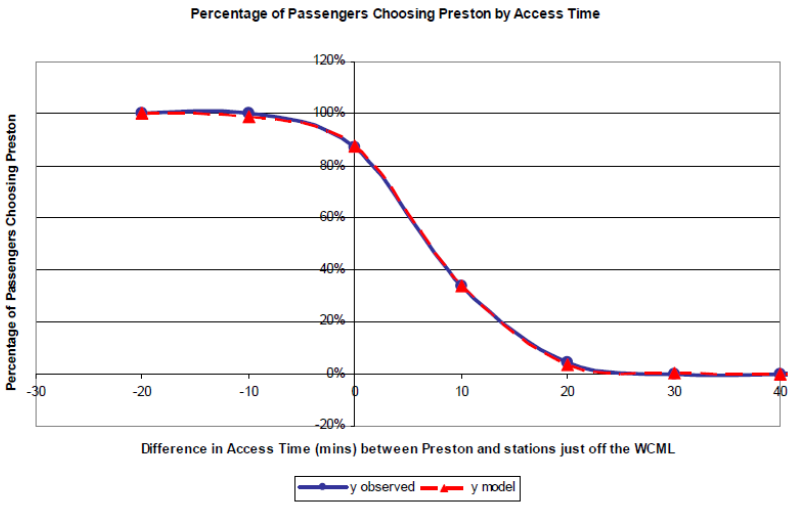


FIGURE 3.1: Difference in access time between Preston and nearby stations (not on the WCML) against percentage of passengers choosing Preston station, showing observed data and fitted logit curve. Note: Reprinted from '*Making better decisions. Assessment of aspirations for track access on the West Coast Main Line*', by MVA Consultancy, 2011, p. 4.6. Reproduced under Open Government Licence v3.0.

Most station choice studies have used larger choice sets, with many applying the MNL model. This is an extension of the binomial logit model that allows choice probabilities for any

[2]In a model of this type where various cost factors are combined into a single GJT, it is usual practice to incorporate a spread parameter which reflects the sensitivity of passengers' choices to changes in GJT (or a component of GJT). The proportion choosing each alternative will be split equally as the value of the spread parameter approaches zero, and the proportion choosing the alternative with the lowest GJT will move towards one as the spread parameter increases (Whelan et al., 2001). The spread parameter is entered into the model as a negative value which ensures that a higher GJT corresponds to a lower utility.

number of alternatives to be calculated. The probability of choosing alternative i from a choice set of J alternatives is then given by the following equation:

$$Pr_{(i)} = \frac{e^{V_i}}{\sum_{j=1}^{J} e^{V_j}}. \tag{3.8}$$

The earliest station choice model to adopt an MNL approach was developed for New Jersey Transit by Kastrenakes (1988). It was used to forecast the proportion of travellers from each minor civil division[3] using each station in that division's observed choice set, and to feed that information into a mode choice model. Due to the lack of suitable software, the logit model was transformed into a linear in parameters form that was then developed as a regression equation.

A research project carried out at TCI Operational Research (formerly the British Rail Operational Research Unit), sought to develop a station choice model that could be incorporated into the UK rail industry demand model, MOIRA, although no progress beyond this preliminary work has been publicly reported (Adcock, 1997). This is probably the most ambitious piece of research to date in terms of the size of the dataset used to estimate an MNL model, with some 230,000 detailed trip observations from the entire UK mainline network and London Underground[4]. The only other study to approach this number of observations is Blainey and Evens (2011) where some 114,000 trip ends covering the Wales and North East regions of the UK were obtained from the NRTS. The Adcock research is unusual in considering the entire passenger trip from the ultimate origin to the ultimate destination based on unit postcode[5], with both access and egress distance included as factors in the models. While Blainey and Evens (2011) included distance from ultimate origin to destination station, most other disaggregate studies have concentrated primarily on the access part of the journey.

In the Netherlands, Debrezion et al. (2007a) developed three MNL models based on different approaches to defining the utility function, including a cross-effect and a translog function (see Section 3.5.3 for more details); and in Canada the station choice of park and ride commuters taking cross-regional trips in the Greater Toronto and Hamilton area was investigated, with separate models calibrated for three market segments based on the type of station (subway and/or commuter rail) considered to be within 'reasonable reach' of the commuter (Mahmoud et al., 2014).

[3]Minor civil divisions are the primary governmental or administrative divisions of a county in many states of the USA. In New Jersey these will refer to townships, cities, towns, boroughs and villages of varying population size.

[4]The dataset was compiled from routine passenger surveys carried out in the late 1980s and early 1990s, including the 'Network South East and London Underground Origin & Destination Surveys', the 'InterCity Monitor', and the 'Regional Railways Monitor' (Association of Train Operating Companies, 2013).

[5]In the UK a unit postcode represents the most detailed spatial unit available from postcode data. For small postal users (i.e. not business addresses), a unit postcode typically represents around 15 addresses, though it is possible to contain up to 100 addresses in densely populated areas.

An unusual approach was used to develop station choice models to assess demand for stations on the planned high speed railway line between London and the West Midlands in the UK (Atkins Limited, 2011). The models are not based on observed station choice data, either in the aggregate or disaggregate, and parameters are not estimated for utility function variables as part of model development. The station choice models use information from the PLANET Long Distance (PLD) multi-modal model (which models long distance journeys above 50 miles by rail, air and car), and two local transport models, RAILPLAN (in London) and PRISM (in the West Midlands). As an example, the London station choice model consists of the following main steps:

1. The GJT from each non-London PLD zone[6] to each London strategic station (which includes existing and proposed HS2 stations) is calculated using the PLD model with established parameters/elasticities for in-train time, waiting time, boarding penalty and a representation of access journey at the origin end.

2. The GJT from each of the strategic London stations to each RAILPLAN zone[7] in London is calculated using the RAILPLAN model.

3. The two GJTs are then summed to derive an end-to-end GJT from each non-London PLD zone to each RAILPLAN zone, via each London strategic station.

4. An MNL probability equation is then used to calculate the share of demand that each London strategic station will attract for each non-London PLD zone to RAILPLAN zone pair. The (dis)utility for each alternative is the end-to-end GJT which is multiplied by a negative spread parameter.

5. The shares are then aggregated into London PLD zones by weighting them using data on the proportion of long-distance demand each RAILPLAN zone is expected to generate. The result of the aggregation is the share of demand between each London PLD zone and each non-London PLD zone that each London strategic station will account for. This data feeds back into the PLD model.

This approach is fairly simplistic as it assumes that the only factor impacting station choice is the end-to-end GJT, which does not include fare and is based on standard elasticities in the PLD model, and there is no estimation or calibration based on observed station choice behaviour.

[6]The PLD aggregates data into large zones, for example the City of Birmingham is a single zone and Greater London is divided into 7 zones, and there are 238 zones in mainland GB. Each zone has a single centroid that represents where 'on average' a traveller starts or ends their trip.

[7]RAILPLAN zones are much smaller than PLD zones, for example the central London PLD contains 493 RAILPLAN zones. RAILPLAN only considers public transport access costs.

3.3.1.1 Assumptions of the MNL model

The key assumption that underlies the relatively simple and easy to understand closed-form logit model is that the unobserved (random) components of utility of the alternatives are independent of each other and have an identical (Gumbel) distribution. The distribution assumption implies that the variance of the random components is the same across all alternatives, and the independence assumption implies that there is no correlation between the random components for any pair of alternatives. Based on a choice set of four alternatives the IIGD assumption can be represented by a 4 by 4 covariance matrix:

$$\begin{array}{c} \text{alternative} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix} \end{array}, \tag{3.9}$$

where the constant variance σ^2 appears on the diagonal and the covariance (correlation) between each pair of alternatives is zero (Jones & Hensher, 2008).

As a practical example, consider a situation where a passenger assigns a higher utility to stations with a staffed ticket office but this factor is not included in the representative utility function. If several stations within a choice set had a staffed ticket office then they would share a common unobserved factor affecting the utility of alternatives and the assumption of independence would be violated. In another example, the choice of station might be influenced by the time taken to find a parking space, and this factor is not accounted for in observed utility. If the time taken shows little variation for some stations in the choice set but varies greatly for others then the assumption of constant variance across the alternatives would be violated.

Train (2009) notes that while the independence assumption may appear very restrictive, it can also be considered the effective outcome of a well specified model, where the representative utility is captured so well by the measured factors that any remaining random component of utility is just 'white noise'. However, if there is correlation present then the researcher must either look for an alternative model, improve the specification of representative utility, or accept that the model is 'only an approximation' and use it anyway.

The MNL model has two further assumptions. The first is that every individual responds to attributes in the same way, known as response homogeneity. This means that the model is unable to account for *individual* 'taste' differences that are due to unobserved characteristics of the individual. For example, if a passenger chooses a station that involves a longer access journey because the drive is more scenic or passes by their child's school. The second is that the variance and covariance of the random components of the alternatives are identical

for all individuals. Consider a model where access distance to a station is included in the utility function, acting as a proxy for unobserved travel time which is likely to be the real determinant of utility. Depending on the access mode used by the individual (for example, walk, cycle or car), the unobserved travel time will differ between individuals and its variance might not be identical across individuals, thus violating the assumption.

3.3.1.2 Independence from irrelevant alternatives and proportional substitution behaviour

As a consequence of the assumptions discussed above, the MNL model exhibits the independence from irrelevant alternatives (IIA) property and displays substitution behaviour that may not be realistic in some circumstances. The IIA property means that the ratio of logit probabilities for any two alternatives, and therefore the odds of choosing one alternative over another, remains the same irrespective of any other alternatives or their attributes. As a consequence, if the probability of an alternative increases due to improved utility then the increase in probability is 'taken from' the remaining alternatives in proportion to their probabilities prior to the change. This is known as proportional substitution (Train, 2009). For example, if there was a choice between three stations with logit probabilities of 0.4, 0.4 and 0.2 and the probability of station 1 increased from 0.4 to 0.6 due to an improved access bus link, the probability increase of 0.2 would be taken two-thirds from station 2 and one-third from station 3, with the new probabilities becoming 0.6, 0.27, and 0.13.

3.3.2 Nested logit

Due to the underlying assumptions of the MNL model, and the proportional substitution behaviour implied by IIA, alternative forms have been sought that, to varying degrees, relax these assumptions. One of the most popular is the NL model which, as will become clearer later, essentially consists of a set of linked hierarchical multinomial models. In the NL model, choice alternatives that are a priori thought to have unobserved factors of utility that are correlated are grouped together into sets known as nests. The theoretical basis of the model is that each pair of alternatives in a nest has the same correlation of unobserved factors, but there is no correlation between pairs of alternatives in different nests (Train, 2009). Each nest exhibits the IIA property and proportional substitution behaviour, but IIA is relaxed between nests, so that the ratio of probabilities of two alternatives in different nests can vary. The NL model is therefore appropriate when the researcher can group alternatives in such a way that IIA holds for each nest but not across nests (Train, 2009). The nest structure is usually depicted using a tree diagram, where a branch represents a group of alternatives and each alternative is a leaf on a twig. For example, Figure 3.2 shows four travel-to-work modes grouped into two nests: public transport (bus or train) and car (drive alone or car share). This structure implies that if bus was removed as an alternative, train would be a

better substitute than either of the car modes, and the probability of train would increase proportionately more than that of car share or drive alone.

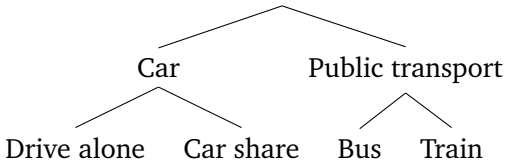


FIGURE 3.2: Mode choice in an NL model.

The degree to which the unobserved factors are correlated, and therefore the degree to which the alternatives in a nest are substitutes for one another, is represented in the model by the inclusive value (IV) parameter which is estimated during model calibration. This parameter determines the pattern of substitution and the extent to which a change in the probability of an alternative is 'passed on' to the alternatives in its nest rather than to alternatives in other nests. The lower the IV parameter the less independent and therefore more correlated the alternatives in a nest are, and the more they are substitutes for one another. With a lower IV parameter, a change in the utility of an alternative will have a proportionately greater effect on the probability of other alternatives within its nest rather than other nests. The IV parameter can be different for each nest but must be between 0 and 1 for the model to be fully consistent with RUM. If the IV parameter of a nest is 1, it indicates that there is no correlation and the alternatives do not need to be grouped — they can be connected as separate branches (direct to the root in a two-level model). If the IV parameter is 1 for *all nests* then all the alternatives can be connected directly to the root of the tree, and the model effectively collapses to a standard MNL model (Koppelman & Sethi, 2000).

It is important to note that the NL model does not impose any behavioural assumptions about the decision process, or the order in which an individual makes a decision. In the example shown in Figure 3.2, there is no assumption that an individual first decides between car and public transport and then decides between the applicable alternatives. The model is merely a mathematical construct to relax IIA and IIGD assumptions in a specific manner (see, for example Hensher, Rose, and Greene (2005); Hunt, Boots, and Kanaroglou (2004); Koppelman and Bhat (2006); Preston (1991b)).[8] It is not uncommon for researchers, including in the field of station choice modelling, to make the mistake of assuming that the model imposes behavioural assumptions. For example, Chakour and Eluru (2014) state that the NL model 'imposes a hierarchy that is very hard to validate in the dataset' and sought to overcome this apparent limitation by developing a 'behaviourally representative framework' where decision makers are initially split using a binary logit model component into one of two segments,

[8]However, note that in NL models that contain three or more levels the direction of change in the IV parameter values between levels of the tree can indicate whether the ordering of the levels is appropriate. Depending on whether the IV parameter at the top or bottom level of the tree is normalised to one, the IV parameters should monotonically decrease or increase respectively when moving from the top to the bottom of the tree (M. Ben-Akiva & Lerman, 1985). Williams (as cited in Boyce and Williams (2016)) concluded that this condition allowed for the order of levels to be empirically tested, with models that fail to meet the condition rejected on the basis of an inappropriate model structure.

where either station choice or access mode is decided first. This paper is discussed further in Section 3.3.2.3.

3.3.2.1 Nested logit probabilities

The probability of individual n choosing alternative i in nest B_k is given by the following formula[9]:

$$P_{ni} = \frac{e^{V_{ni}/\lambda_k}(\sum_{j \in B_k} e^{V_{nj}/\lambda_k})^{\lambda_k - 1}}{\sum_{l=1}^{K} (\sum_{j \in B_l} e^{V_{nj}/\lambda l})^{\lambda_l}}. \tag{3.10}$$

However, it is easier to interpret the NL model if it is thought of as two modelling steps[10]. At the lower level the model predicts a series of conditional probabilities for each alternative, conditional on the nest containing each alternative being chosen. Then, at the upper level, the model predicts the marginal probability of each branch. The probability of an alternative within a nest being chosen is then given by the product of the relevant marginal and conditional probabilities. The probabilities can be expressed in a simpler way than Equation 3.10 by using two logit equations, but first the representative utility of an individual n choosing alternative j in nest k, needs to be split into two components:

$$V_{nj} = W_{nk} + Y_{nj}, \tag{3.11}$$

where W_{nk} includes factors that relate to nest k and are constant for all alternatives in nest k (but vary between nests); and Y_{nj} includes factors that relate to alternative j and differ between alternatives in nest k. A mechanism is also needed to link the information from the lower (conditional) logit to the upper (marginal) logit. This is done by incorporating the expected maximum utility derived from all the alternatives in a nest as an explanatory variable in the upper model. The expected maximum utility is equal to the natural logarithm of the denominator of the lower model (i.e. the log of the summed exponentiated representative utilities for each alternative in the nest) and has several names, including IV, inclusive utility and logsum (Hensher et al., 2005; Train, 2009). The marginal probability of individual n choosing any alternative in nest B_k can now be expressed as:

$$P_{nB_k} = \frac{e^{W_{nk} + \lambda_k I_{nk}}}{\sum_{l=1}^{K} e^{W_{nl} + \lambda_l I_{nl}}}, \tag{3.12}$$

[9]The presentation of the nested logit probabilities provided here largely follows the notation of Train (2009)
[10]This relates to a two-level nested logit, more levels are possible

and the conditional probability of individual n choosing alternative i given that an alternative in nest B_k is chosen as:

$$P_{ni|B_k} = \frac{e^{Y_{ni}/\lambda_k}}{\sum_{j \in B_k} e^{Y_{nj}/\lambda_k}}, \tag{3.13}$$

where:

$$I_{nk} = \ln \sum_{j \in B_k} e^{Y_{nj}/\lambda_k}. \tag{3.14}$$

I_{nk} is the IV and λ_k is the IV parameter, referred to in the previous section, that is estimated by the model. The probability of individual n choosing alternative i in nest B_k is given by:

$$P_{ni} = P_{ni|B_k} \times P_{nB_k}. \tag{3.15}$$

3.3.2.2 Nested logit in station choice

Researchers who have used the NL model to analyse station choice have predominantly chosen a two-level model with access mode at the upper level and station choice at the lower level (Davidson & Yang, 1999; Debrezion et al., 2009; Fan et al., 1993; Givoni & Rietveld, 2014). For example, the nest structure adopted by Fan et al. (1993) is shown in 3.3, with only a single choice available for walk access as the distance between stations indicated that it was only plausible for a traveller to access their nearest station on foot. Some researchers have attempted to produce models with station choice at the upper level and mode choice at the lower level, but have rejected the approach as their models were not consistent with RUM, due to the IV parameter being outside of its required bounds. For example, Debrezion, Pels, and Rietveld (2007b) obtained an IV parameter of 2.02, while Fan et al. (1993) obtained an IV parameter of 7.97. An IV parameter greater than 1 indicates that the model is only consistent with RUM for some, but not all, possible values of the explanatory variables (Train, 2009). While Liou and Talvitie (1974) reported that station choice as the marginal logit was their preferred model, this work pre-dated the formal development of the NL model and did not correctly calculate IV to be consistent with RUM (the precise mechanism was first identified by M. E. Ben-Akiva (1973) in his PhD thesis).

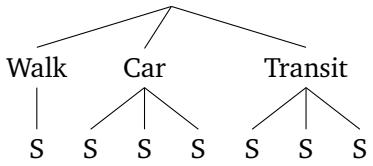


FIGURE 3.3: Nest structure used by Fan et al. (1993).

Very few alternative nesting structures have been implemented. Harata and Ohta (1986) investigated station choice before and after the opening of a new railway station and, on the basis that this could affect the choice of mode for the main journey, they used a three-level nested logit model with main mode at the upper level (bus or rail), followed by access

mode and station choice at the lower levels of the rail branch. And in research to develop an improved flow model to estimate demand for parkway stations, Lythgoe and Wardman (2002) introduced a station choice element by modelling travel choices from each zone to each destination using nested logit. In this model the choice to travel by rail or not by rail (or not at all) was in the upper level and the choice of station, conditional on a rail journey being made, was in the lower level (see Figure 3.4). A form of CNL model was subsequently developed to address spatial correlation between stations (Lythgoe et al., 2004), and this is discussed in Section 3.3.3.1. More information about the flow modelling approach can be found in Section 3.6. These examples aside, prior station choice research has usually assumed that the decision to travel by train has been made.

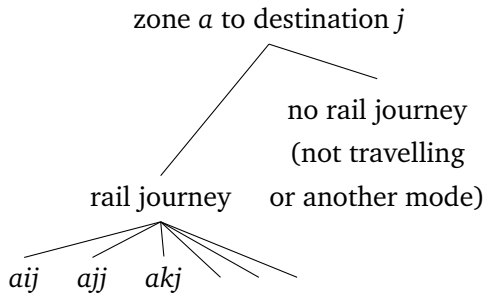


FIGURE 3.4: Nested logit structure used by Lythgoe and Wardman (2002).

In most research where NL has been used to model station choice, the same station alternatives appear in each nest for each of the access modes. This appears at odds with the primary advantage of the NL model — that it can potentially resolve substitution issues related to IIA. Train (2009) notes that the researcher should approach the grouping of alternatives into nests in terms of limiting the impact of IIA, by identifying alternatives where IIA either does or does not hold. He also states that each alternative should be a member of only one nest or subnest (for models with more than two layers). From a purely IIA perspective, there is nothing to be gained by placing the *same* stations under each access mode. Rather than creating distinctive groupings of alternatives, this approach creates multiple groupings composed of the same alternatives, and if proportional substitution is an issue for one nest it is likely to be an issue for all the nests. That said, grouping stations by access mode can be expected to result in improved models compared with multinomial logit, as correlations that exist between unobserved factors common to each access mode can be accounted for, resulting in better fitting models and less biased coefficients. This is confirmed by previous studies which have reported IV parameters that lie between 0 and 1 and that are significantly different from 0 or 1, indicating that correlation exists and the nest structure is appropriate and consistent with RUM. The degree of correlation between unobserved factors within a nest can be obtained by using the equation $1-\lambda$, and calculated correlation values[11] for reported IV parameters are shown in Table 3.2. They are in the moderate to low range.

[11] It is not quite a straightforward as this, but this is a good indication (Train, 2009).

Paper	IV Parameter	Correlation	Notes
Harata and Ohta (1986)	0.641	0.359	Before new station
Harata and Ohta (1986)	0.740	0.260	After new station
Fan et al. (1993)	0.414	0.586	
Debrezion et al. (2009)	0.614	0.386	
Givoni and Rietveld (2014)	0.546	0.454	

TABLE 3.2: Reported IV parameters and calculated correlation for nested logit station choice models.

While the nesting structure adopted in prior station choice modelling work has produced models that perform better than standard MNL, their inability to deal with inappropriate substitution behaviour, for example caused by location in space, suggests that an alternative approach might be warranted. It may be that addressing substitution behaviour is more critical when developing models that are to be used for planning purposes to predict demand at a new station, and abstraction from existing stations, than it is for studies that are primarily concerned with examining the influence of explanatory factors on station choice. The specific issue of spatial choice is discussed in Section 3.3.3.

3.3.2.3 A latent segmentation approach

As mentioned in Section 3.3.2, Chakour and Eluru (2014) incorrectly state that the NL model 'imposes a hierarchy that is very hard to validate in the dataset' and go on to suggest a latent segmentation approach to overcome this apparent limitation. In their proposed framework there are assumed to be two decision sequences taking place — either station choice first or access mode first. These two decision sequences are referred to as segments, and observations are split between the segments by the model during estimation based on a range of factors, including socio-economic variables. The proposed model framework consists of three model components, which are estimated simultaneously using maximum likelihood:

- The latent segmentation component — this is a binary logit model that determines the order of mode choice and station choice.
- Mode choice — an MNL model.
- Station choice — an MNL model.

As the latent segmentation component is a RUM-based binary logit model, the implied assumption is that individuals will make a choice, determined by utility maximisation, *of the order in which* they are going to make a choice of access mode and station choice. The authors describe this as a 'behaviourally representative framework' that is preferable to the NL model, but present no behavioural research to support this. They give a couple of possible

examples, such as a worker with primary access to a car who may choose the car as access mode and then choose a station based on car park availability; and someone living very close to a station, who may choose that station first and then choose to walk or go by public transport in poor weather. While these examples suggest an order in which the mode and station decisions might be made, that does not equate to an individual making a binary choice *of that order* based on utility maximisation.

This solution to the so-called 'imposed hierarchy' of the NL model is an approach that appears difficult to justify behaviourally. The authors have adopted a model form developed by Waddell et al. (2007) that was concerned with whether household residence is decided before or after the choice of workplace, without making a critical assessment of whether the behavioural assumption remains valid in an entirely different choice context. It is much clearer that an individual or household may indeed weigh up the utility arising from the order in which these two decisions are made. The authors report that the latent segmentation model had a lower Bayesian information criterion (BIC) (11,288.90) than separate sequential models for station choice conditional on mode choice (13,094.65) and mode choice conditional on station choice (12,437.51). However, no information on the nature of the sequential models used in these comparisons is provided. In particular, it is not clear that they were RUM-compliant nested logit models, which would have been the most appropriate benchmark.

3.3.3 The spatial choice problem

By the very nature of the human involvement in deciding where stations are located, it is extremely unlikely that they are distributed randomly in space or on the access network. Certain stations will be closer to some stations than others and, ignoring all other attributes of the stations, spatial correlation will be present. Spatial correlation might also occur between attributes of stations (whether they be observed or unobserved) as it is likely that they will be more similar when stations are closer together. If station A is closer to station B than station C it would be a reasonable expectation that station B is a better substitute for station A than station C. In the context of discrete choice models, if spatial correlation is present then the assumption of no correlation in unobserved utility between alternatives will not hold, unless it can be overcome with the model structure or represented in the observed utility.

Consider a scenario where there is a choice between three stations as shown in Figure 3.5(a), where the probability of an individual from origin O choosing one of two nearby stations A or B is 0.4 and of choosing more distant station C is 0.2. Assuming that A and B are near perfect substitutes for each other, if B was closed the probability of choosing A would be expected to increase to 0.8, with the probability of choosing C unchanged, as shown in Figure 3.5(b). However, an MNL model would allocate station B's probability proportionately between stations A and C resulting in the probability of choosing C rising from 0.2 to 0.33, as shown in Figure 3.5(c). In this example, the unobserved utilities of A and B are highly correlated due to their location in space; they exhibit spatial correlation. If an NL model was constructed to

limit the impact of IIA, a spatially-based grouping might be considered, with A and B in one nest and C in another, as shown in Figure 3.6.

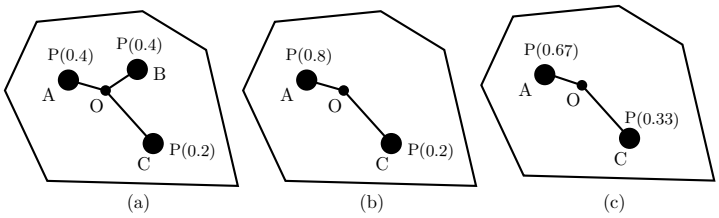


FIGURE 3.5: IIA substitution behaviour — the effect of spatial correlation.

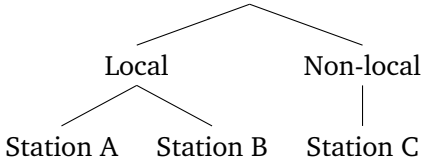


FIGURE 3.6: A possible method of nesting stations to address the impact of spatial correlation.

The issue of spatial correlation has been largely ignored in the station choice literature. Discrete choice models that are suitable when considering alternatives that do not have a spatial element (for example, mode choice) have been applied to railway stations that clearly are located in space; and studies that have used NL have placed the same stations in each access mode nest, which does nothing to address the spatial choice problem. While there are many examples of NL being applied to model spatial choice in a variety of other research fields, this does require the researcher to divide continuous space into discrete clusters of space. This is difficult to do in a manner that is justifiable and not arbitrary, particularly if the model is to be transferred to a different area from the one on which it was calibrated. Furthermore, the assumption of equal substitutability within each nest remains (Pellegrini & Fotheringham, 2002). Only two prior studies have considered the issue of spatial correlation in the context of station choice, and these both adopted alternative approaches. In the first, Lythgoe et al. (2004) incorporated a CNL station choice component into a direct demand model used to forecast the number of trips between station pairs. This method is discussed in section 3.3.3.1, following an introduction to the generalized nested logit (GNL) model of which CNL is a restricted case. In the second study, Weiss and Habib (2017) developed an ML model that specified a correlation between each pair of stations based on the distance between them, and this is considered in more detail in Section 3.3.4.

3.3.3.1 Generalized nested logit

The NL model is the simplest of the generalised extreme value (GEV) group of models which relax, to varying degrees, the assumption of no correlation between unobserved utility. This group also contains models that allow an alternative to be in more than one nest, enabling more flexible and complex substitution patterns to be represented. An example of why this

may be important is the mode choice NL model shown in Figure 3.2. While car share would be expected to have unobserved attributes in common with car, this may also be true of the public transport modes. For example, car share also suffers from lack of travel time flexibility. If car share could be placed in both the car *and* public transport nests the model could account, differently, for correlation with car and with bus and train (Train, 2009). A number of models with these overlapping nests have been separately specified, such as paired combinatorial logit (PCL) and CNL, but these can be considered restricted cases of the GNL model proposed by Wen and Koppelman (2001).

In the GNL model[12] an alternative can be present in a nest to varying degrees determined by an allocation parameter which has a value between zero (alternative not in nest at all) and one, where the sum of allocation parameters for each alternative must equal one. Each nest has a logsum (or dissimilarity) parameter that indicates the degree of independence between alternatives within a nest, where higher values indicate greater independence and lower correlation, as with the NL model. The logsum parameters should have a value between zero and one to be consistent with RUM. The correlation between alternatives within a nest, and the degree to which they are substitutes for one another, is a function of both the logsum and the allocation parameters, with correlation increasing as the logsum parameter reduces and the allocation parameter increases. The utility, logsum and allocation parameters are estimated simultaneously, and the probability of individual n choosing alternative i is given by:

$$P_{ni} = \sum_k \left(\frac{(\alpha_{ik} e^{V_{ni}})^{\frac{1}{\mu_k}}}{\sum_{j \in N_k} (\alpha_{jk} e^{V_{nj}})^{\frac{1}{\mu_k}}} \times \frac{\left(\sum_{j \in N_k} (\alpha_{jk} e^{V_{nj}})^{\frac{1}{\mu_k}} \right)^{\mu_k}}{\sum_k \left(\sum_{j \in N_k} (\alpha_{jk} e^{V_{nj}})^{\frac{1}{\mu_k}} \right)^{\mu_k}} \right), \tag{3.16}$$

where $j \in N_k$ is the set of alternatives that are members of nest k; α_{ik} is the allocation parameter which determines the portion of alternative i assigned to nest k; and μ_k is the logsum parameter for nest k. The first component of the product is the probability of alternative i being chosen from amongst all alternatives that are members of nest k, conditional on nest k being chosen ($P_{i|k}$). The second component of the product is the probability of nest k being chosen from amongst all nests (P_k). The probability of individual n choosing alternative i can therefore be re-written as:

$$P_{ni} = \sum_k P_{ni|k} \times P_k. \tag{3.17}$$

By imposing constraints on the parameters other GEV models can be specified using the GNL model. For example, in the CNL model the logsum parameters are constrained to be equal and in the PCL model the allocation parameters are constrained to be equal (to one).

As mentioned in the previous section, Lythgoe et al. (2004) used a form of CNL in the station choice component of a direct demand model. This model was calibrated on inter-urban rail

[12] This explanation of GNL and the notation used is based largely on Wen and Koppelman (2001).

journeys greater than 40 km between pairs of stations in Great Britain. A full discussion of the direct demand model is provided in Section 3.6, and only aspects of the model relevant to addressing spatial correlation are considered here. The CNL model replaced a NL model used in earlier research (Lythgoe & Wardman, 2002), specifically to address spatial correlation between stations. The model allows the proportion of journeys at a new station that are abstracted from existing stations, rather than newly generated, to be higher the closer the new station is to its competitor stations. In the direct demand model resident population is assigned to 16 polygonal zones generated around each origin station. A set of up to 15 competing stations is defined for each origin station, and this forms the choice set for each origin station zone. The dependent variable in the model is the number of journeys by rail between origin station i and destination station j. The population of each zone of origin station i is weighted by the probability of travelling from that zone (a) to destination station j (conditional on the decision to travel by rail from a to j) and then summed for all zones. In the station choice component each station i is nested with each of the other competitor stations k, with i apportioned across the nests by the allocation parameter α and with k fully allocated to the nest. The nest structure is illustrated in Figure 3.7.

The probability of travelling from zone a to destination station j via origin station i is given by the following equation:

$$P_{aij} = \sum_{k \neq i} \left(\frac{(\alpha_{ik} e^{V_{aij}})^{\frac{1}{\nu_{ik(a)}}}}{(\alpha_{ik} e^{V_{aij}})^{\frac{1}{\nu_{ik(a)}}} + (e^{V_{akj}})^{\frac{1}{\nu_{ik(a)}}}} \times \frac{\left(e^{V_{a\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right]j}}\right)^{\frac{1}{\mu}}}{\sum_{k \neq i} \left(e^{V_{a\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right]j}}\right)^{\frac{1}{\mu}}} \right), \tag{3.18}$$

where $V_{a\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right]j}$ is the utility of travelling from zone a using rail from either station i or station k (a 'choice pair') to station j; $\nu_{ik(a)}$ is the dissimilarity parameter between station i and station k (the choice pair) given that the journey starts at zone a (i.e. the degree to which i and k are substitutes given zone a); μ is the dissimilarity parameter (to be estimated) between choices of choice pairs (i.e. the degree to which choice pairs are substitutes); and α_{ik} is an allocation parameter to distribute the probability of station i to each of the choice pair nests. The first component of the product is the probability of station i being chosen from nest $\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right]$, conditional on that nest being chosen. The second component of the product is the probability of nest $\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right]$ being chosen from amongst all choice pairs $\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right] \forall k \neq i$.

The allocation and dissimilarity parameters were 'part calculated' prior to estimation of the direct demand model. A logit form was used for the allocation parameters:

$$\alpha_{ik} = \left(\frac{e^{\theta_{ik} L_{ik}}}{\sum_{k} e^{\theta_{ik} L_{ik}}} \right)^{\mu}, \tag{3.19}$$

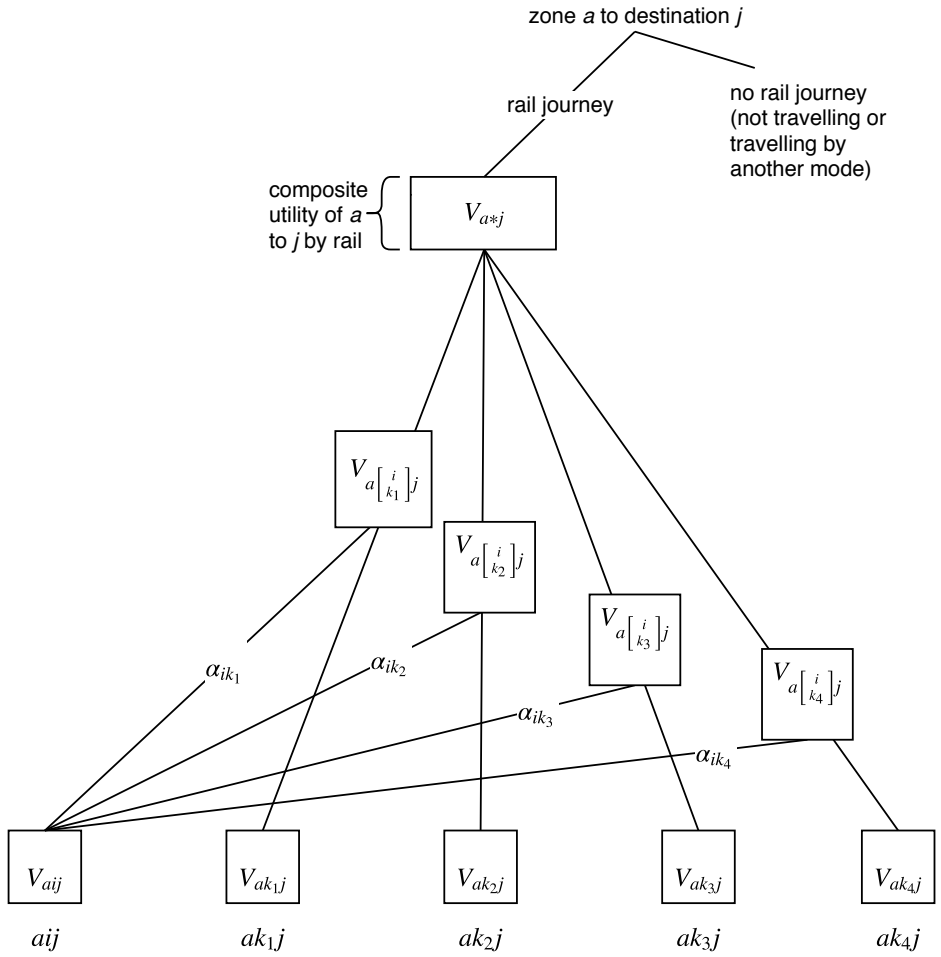


FIGURE 3.7: The cross-nested logit structure for origin station choice adopted by Lythgoe et al. (2004) showing utility notation. For clarity only the nests with respect to station i and four competing stations (k_1 to k_4) are included, with $\alpha_{ik_{1..4}}$ representing the allocation parameters.

where L_{ik} is the road distance between station i and station k and θ_{ik} is a parameter that was tested at different values. The following form was used for the dissimilarity parameter:

$$\nu_{ik(a)} = \left(\frac{2T_{ik}}{T_{ai} + T_{ak} + T_{ik}}\right)^{\phi} \mu, \tag{3.20}$$

where T_{ai} is the road journey time between zone a and station i; T_{ak} is the road journey time between zone a and station k; T_{ik} is the road journey time between station i and station k; and ϕ is a parameter to be estimated. If stations i and k were adjacent to each other then $\nu_{ik(a)}$ would be close to zero, indicating a high degree of spatial correlation and substitutability.

The direct demand model was estimated with the origin station choice component taking either the MNL or CNL form. The best fitting CNL model marginally improved model fit compared to MNL, with adjusted R^2 increasing from 0.6087 to 0.6108. This model specified

θ_{ik} as zero, which would be equivalent to defining the allocation parameter as[13]:

$$\alpha_{ik} = \left(\frac{1}{15}\right)^{\mu}, \tag{3.21}$$

which would allocate the same proportion of station i to each of the choice pair nests. This seems counter-intuitive as a negative θ_{ik} would be expected, causing a greater portion of station i to be allocated to the nests of nearer competitor stations than those further away, thus allowing greater competition between station pairs that are closer to one another. Indeed, Lythgoe (2004) notes that 'the allocation parameters should probably have been tuned to differentiate more effectively between the effects of different competing stations'.

The validation and testing of the CNL approach was limited to forecasting flows to London and Edinburgh from two hypothetical new stations located close to Leeds (identified as Leeds West and Leeds South) and examining the effect on existing stations (Lythgoe, 2004). This analysis found that while the proportion of journeys from the new stations that were abstracted from existing stations was broadly similar using both the MNL and CNL approaches, the proportion abstracted from the nearest competing station (Leeds in the case of Leeds West, and Wakefield in the case of Leeds South) was increased in the CNL model. However, these findings are of limited value as the forecasts and abstraction effects cannot be verified. It would have been more informative to apply the models to several recently opened stations that did not form part of the calibration dataset.

The approach adopted by Lythgoe et al. (2004) is unusual as the station choice model is not calibrated against observed choice. It forms one component of a direct demand model where the dependent variable is the number of journeys made between each station pair (Q_{ij}). This simplifies the model estimation as for each Q_{ij} only the probability of origin station i being chosen to travel to j for each origin zone is considered (this weights the population of each zone which is then summed for all zones). Furthermore, the choice set is defined at the origin station, and is therefore the same for each origin station zone. The difficulties in adopting this approach for a station choice model calibrated against observed choice where choice sets are defined at the individual level, and when it is to be incorporated into an aggregate demand model with zones defined at high spatial resolution, are discussed in Section 6.2.1.

3.3.3.2 Bespoke GEV models

It is possible for a researcher to develop new GEV models to meet specific research needs by following a generation process developed by McFadden (Train, 2009). An example of this in the realm of spatial choice is the Generalised Spatially Correlated Logit (GSCL) model developed by Sener, Pendyala, and Bhat (2011). In this model the degree of spatial correlation is represented by a function of a vector of attributes that defines the spatial relationship between all pairs of alternatives. They suggest a variety of variables that might be included

[13] Assuming 15 competitor stations in the choice set.

in the vector, although in a residential location choice study to test the approach only the distance between each pair of alternatives proved to be statistically significant. They found that the model was able to capture declining correlation effects as the distance between alternatives increased, indicating that a model of this form may be appropriate to model station choice.

3.3.3.3 Addition of an accessibility term

An alternative, and much simpler, approach to deal with spatial correlation that has been applied in other research fields is to include an accessibility term within the MNL model. This term is a measure of the accessibility of an alternative to all other alternatives within a choice set and can take a variety of forms. It is often a Hansen-type measure, where the distance between alternatives is weighted by a size-based attraction variable (e.g. population). As the term includes information from other alternatives the IIA property no longer holds and the model is able to capture competition (or agglomeration) effects. Probably the most enduring research of this nature was carried out by Fotheringham in the 1980s with the development of the competing destinations model (CDM), primarily based on studies of migration and consumer store choice (see Pellegrini and Fotheringham (2002) for a comprehensive review). More recent applications include incorporating two accessibility variables in destination choice models to account for agglomeration and spatial competition effects separately (Bernardin, Koppelman, & Boyce, 2009); and using accessibility terms to account for spatial competition in workplace choice models (Ho & Hensher, 2016).

The following form of the accessibility term is suggested by Fotheringham:

$$a_{ni} = \left(\frac{1}{M-1} \sum_{\substack{k \\ k \neq j}} \frac{W_k}{d_{jk}} \right)^{\theta}, \tag{3.22}$$

where M is the total number of k alternatives for individual n at origin i, W is a weight (usually size-based, for example, population), d is the distance from alternative j to alternative k, and θ a parameter to be estimated. A large value of a_{ni} indicates that an alternative is in close proximity to other alternatives, and vice versa. If $\theta < 0$ then alternatives that are more isolated will have a higher probability of being chosen and alternatives closer together will have a lower probability. Conversely, if $\theta > 0$ then more isolated alternatives will have a lower probability and alternatives closer together will have a higher probability[14]. If $\theta = 0$ then the model is the standard MNL. The accessibility term can be included directly in the utility function, and Fotheringham suggests a logarithmic transformation which would imply

[14]This relationship does, however, rely on the numerator (W) always being larger than the denominator (d) in the accessibility function shown in Equation 3.22.

that the utility gained from choosing alternative i increases or decreases (depending on the sign of θ) at a *decreasing rate* as a_{ni} increases[15].

As a result of introducing this term into the model, the IIA property is circumvented, *to an extent*, as the utility of an alternative is now dependent on a function that is determined by the location and 'size' of other alternatives in the choice set. Therefore, if a set of alternatives all have identical observed utilities (V_i), the probabilities are no longer bound to be equal. The behaviour of the model is illustrated in Figure 3.8, where the observed utility of alternatives A, B and C for an individual at origin O is assumed to be equal, the size weight is assumed to be constant and the distance between A and B is 1 unit and between B and C is 2 units. Figure 3.8(a) shows the probabilities for the standard MNL, 3.8(b) the probabilities for the CDM when $\theta = -0.5$, and 3.8(c) the probabilities of the CDM when $\theta = +0.5$. As C is more isolated, when θ is negative its probability increases (competition effect), and when θ is positive the probabilities of A and B increase (agglomeration effect).

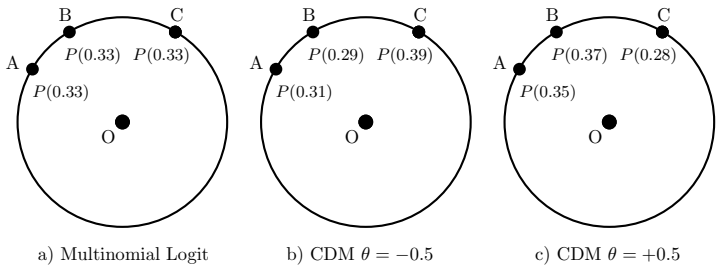


FIGURE 3.8: Effect of the CDM on choice probabilities.

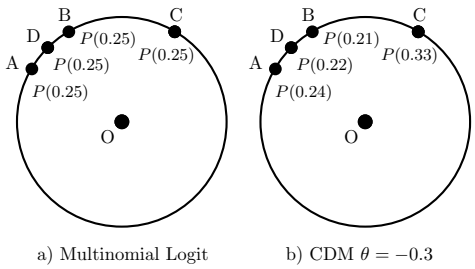


FIGURE 3.9: Behaviour of the CDM when a new alternative is added, compared to standard MNL.

However, it is important to note that the CDM is not entirely free of IIA and it does not result in fully flexible substitution patterns. If the V_i of an alternative *changes*, the model will still apportion the resultant change in probability proportionately across the other alternatives, as there has been no change in the accessibility term. However, the IIA property is now relaxed when an alternative is added to or removed from the choice set (and if the size weighting were to change), as the accessibility term will alter, and in these circumstances the ratio of probabilities of two alternatives is no longer constant. With a negative parameter, if a new alternative is added to the choice set the probability of alternatives with a high accessibility

[15]The requirement for a logarithmic transformation also has a theoretical basis arising from the term being considered a weight on the utility function in Fotheringham's hierarchical choice rationale for the CDM (see discussion later in this section).

term (in close proximity to other alternatives) will be reduced more than those with a low accessibility term (more isolated from other alternatives). This is illustrated in Figure 3.9, where a new alternative D is placed equidistant between A and B. In Figure 3.9(a), the standard MNL model, the probabilities of A, B and C are all reduced by the same proportion. However, in Figure 3.9(b), a CDM where θ is -0.3, the probability of C remains unchanged and the probabilities of A and B are reduced due to a competition effect. This suggests that incorporating an accessibility term of some form into an MNL station choice model might have the potential to forecast differential abstraction effects.

A potential concern with the CDM, in common with other models that include an explicit measure of dis(similarity) between alternatives in the utility function in order to circumvent the IIA property, is that it may not be consistent with the utility maximisation paradigm of consumer behaviour and the assumptions that underlie RUM models. These include the regularity condition, which requires that the probability of choosing an existing alternative should not increase if new alternatives are added to the choice set. As a consequence, in a RUM-compliant model the difference in utility between any two alternatives in a choice set should not be dependent upon the attributes of other alternatives or their existence (Hess, Daly, & Batley, 2018). Although attributes of alternatives ($j \neq i$) are not directly included in the utility function of i, because the accessibility term represents the average weighted distance of an alternative from all other alternatives it is necessarily dependent upon the spatial location and 'size' of the other alternatives in the choice set, and the regularity condition is therefore violated (a consequence acknowledged by Pellegrini and Fotheringham (2002)).

The implications of a random utility-based model not being consistent with RUM are difficult to determine as there is limited literature on the subject. The benefits of a RUM-compliant model are generally framed in the context of its foundation in economic theory supported by substantial empirical evidence (for example, see Hess, Beck, and Crastes dit Sourd (2017) and Hess et al. (2018)). Hess et al. (2017) emphasise the potential problems when non-compliant models are used to calculate economic measures such as willingness to pay or the value of time, and imply that the issue may be less important when models are used in forecasting; although Hess et al. (2018) note that RUM provides the justification for assuming that the observed (past) behaviour of individuals will continue in the future. Hess et al. (2018) also make an assessment of the theoretical RUM-consistency of a range of model types and their practicality for deriving economic measures and forecasting. They focus on models developed to address the perceived behavioural 'anomalies' of utility maximisation and do not specifically address the CDM, but they state that models which attempt to capture correlation effects through observed utility, for example to capture spatial overlap of links in models of route choice, are unlikely to be consistent with RUM. The CDM has similarities with these models, as well as the universal (or 'mother') logit model in which attributes of competing alternatives enter the utility function of each alternative through so-called cross-effects. McFadden and Train (2000) note that the 'mother' logit model 'is not guaranteed to

be consistent with RUM'.

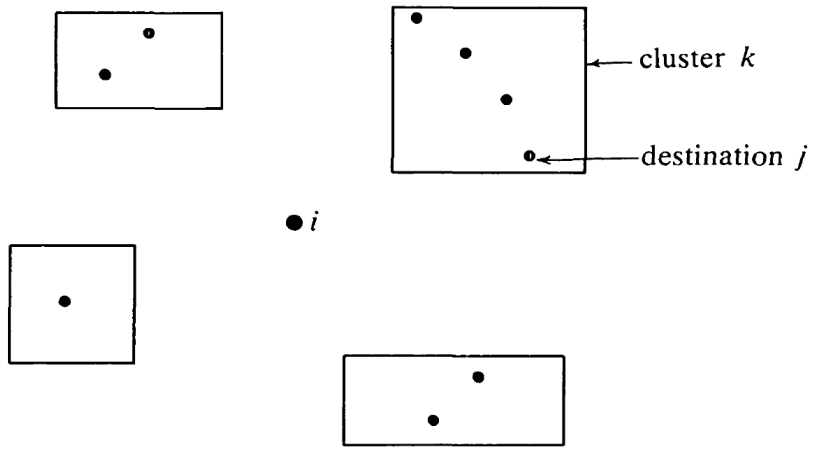


FIGURE 3.10: Demonstrating hierarchical destination choice. Note: Reprinted from 'Modelling hierarchical destination choice', by Fotheringham, A. S., 1986, Environment and Planning A, 18(3), 401-418. Image reproduced with permission of the rights holder, SAGE Publications.

Hunt et al. (2004) consider that researchers who use model adjustments of this type should 'either demonstrate that their model is consistent with random utility theory [or] describe the different behavioural assumptions associated with their model'. Fotheringham (1986) suggests that including the accessibility term in the utility function can be justified on the basis of utility maximisation theory in some circumstances. For example, in the case of a retail outlet that is not meeting the needs of a consumer, utility might be gained by having other stores nearby (in which case θ would be positive). Recognising that this would not always be a justifiable approach, Fotheringham also proposed a behavioural rationale that assumes individuals use a 'hierarchical information-processing strategy'. When faced with many spatial alternatives to choose between individuals will 'cognize' them in clusters which are evaluated first before an alternative is chosen (see Figure 3.10). In effect there is an unobserved nested hierarchy that avoids the need for the researcher to impose one. It is hypothesised that individuals will underestimate the size of large clusters and the number of alternatives within them, and therefore select them less often than expected. Consequently, alternatives within large clusters are less likely to be chosen. Destinations with relatively high accessibility (close to many others) are more likely to be in large clusters and therefore less likely to be chosen; while isolated destinations are more likely to be in small clusters and therefore more likely to be chosen. The parameter θ is therefore expected to be negative, with the likelihood of a destination being chosen reducing as its accessibility increases (hence the model is known as the *competing* destinations model). Pellegrini and Fotheringham (2002) describe the CDM as a generalisation of MNL where the utility function of each alternative is weighted to reflect the probability of that alternative being evaluated, with the model taking

the following general form:

$$P_{nij} = \frac{\exp(V_{nij})L_{ni}(j \in G)}{\sum_{k=1}^{M} \exp(V_{nik})L_{ni}(k \in G),} \tag{3.23}$$

where $L_{ni}(j \in G)$ is the likelihood that alternative j is in individual n's chosen cluster G. The likelihood function proposed by Fotheringham is the accessibility term shown in Equation 3.22, which can be added directly to the utility function as a logarithmic transformation:

$$P_{nij} = \frac{\exp(V_{nij}) \times a_{nij}^{\theta}}{\sum_{k=1}^{M} \exp(V_{nik}) \times a_{nik}^{\theta}} = \frac{\exp(V_{nij} + \ln a_{nij}^{\theta})}{\sum_{k=1}^{M} \exp(V_{nij} + \ln a_{nik}^{\theta})}. \tag{3.24}$$

3.3.4 More complex models

Alternative discrete choice models are available that, by making different assumptions about the distribution of unobserved utility, can represent any pattern of substitution and, unlike MNL and NL, account for random variation in taste. With reference to the covariance matrix shown in Equation 3.9, in these models the variances on the main diagonal are not assumed to be identical, and the off-diagonal covariances are no longer constrained to zero. However, this increased flexibility comes at a cost. The models are more complex to implement and interpret and the choice probabilities usually have to be approximated by simulation. There has been only limited application of these model types in prior station choice research.

3.3.4.1 Probit model

In the multinomial probit model, the random components of utility are assumed to follow a multivariate normal distribution with a mean vector of zero. The probability of individual n choosing alternative i from a choice set of J alternatives is an integral given by the following equation:

$$\begin{aligned} P_{ni} &= Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \quad \forall j \neq i) \\ &= \int I(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i)\phi(\varepsilon_n)d\varepsilon_n, \end{aligned} \tag{3.25}$$

where $I(.)$ is an indicator function of whether the bracketed statement is true, and the integral is over all values of ε_n, and ϕ is the cumulative distribution function of the standard normal distribution (Train, 2009). A derivation of the model is possible that enables coefficients to vary randomly by individual, allowing for taste variation. However, it is only suitable if the assumption that the random coefficients follow a normal distribution holds, which necessarily implies that the coefficients will be positive for some individuals and negative for others

(Train, 2009). This may not be an appropriate assumption for station choice models as it implies, for example, that some individuals will have a positive coefficient for access time and some will have a negative coefficient for service frequency. The multinomial probit model can also represent any substitution pattern, which can be accomplished through estimation of a full covariance matrix (although Train (2009, p. 109) notes that this 'renders the estimated parameters essentially uninterpretable'), or by the researcher imposing constraints on the covariance matrix to enable a desired substitution pattern (although this is far from being a straightforward procedure). Greene (2012, p. N-465) notes that the multinomial probit model is 'extremely difficult to estimate [and the] difficulty increases greatly with the number of alternatives'. This is especially the case if the covariance matrix is not constrained.

Desfor (1975) used a probit model but made the simplifying assumption that commuters only considered the two lowest cost stations for the census block where they resided, determined using a non-stochastic trip cost function consisting of distance, fare and parking cost. This allowed binary probit models to be estimated for the pair of lowest cost stations for each census block, with the difference in trip cost the only explanatory variable. While the models were reported to correctly predict the choices made by 88% of commuters, only those who actually chose one of the two highest utility stations (80% of the sample) could be included in the validation. This is likely to have enhanced model performance as the remaining 20% of cases were arguably the more difficult ones to predict. Furthermore, the model's usefulness for forecasting is limited, as it is impossible to make an a priori assessment of which travellers would choose one of the two lowest cost stations.

3.3.4.2 Mixed logit

It has been suggested that the mixed logit model has the 'flexibility of probit [while] keeping part of the simplicity of logit' (Munizaga & Alvarez-Daziano, 2001). As with probit, ML can represent any substitution pattern and can account for random variation in taste. The key feature of the model is that unobserved utility is represented by two components — one that is assumed to be IIGD (as in logit) and another that can follow any distribution and allows for correlation and heteroscedasticity (non-constant variance) across alternatives. This can be expressed in the following equation:

$$U_{nj} = V_{nj} + [\eta_{nj} + \varepsilon_{nj}], \tag{3.26}$$

where η_{nj} represents the additional random term which depends upon parameters and observed variables relating to alternative j and individual n, and the square brackets denote the stochastic (unobserved) portion of utility. The usual form of the ML probability is as

follows:

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}} \right) f(\beta) d\beta, \tag{3.27}$$

which is the weighted average of the logit formula at different values of β, with the weights provided by the density $f(\beta)$ (Train, 2009).

The ML model can be interpreted in a number of ways, which are equivalent but affect the way the model is specified. The 'random parameters' approach allows some or all of the parameters to vary by individual, from a distribution chosen by the researcher. Utility is specified in the same way as with the MNL model, except the vector of coefficients is now able to vary by individual. The 'error components' approach is useful when the researcher is seeking to achieve a certain substitution pattern and the primary aim of the model is prediction. In this case the utility for individual n of alternative j can be expressed as:

$$U_{nj} = \beta x_{nj} + [\mu_n z_{nj} + \varepsilon_{nj}], \tag{3.28}$$

where x_{nj} and z_{nj} are vectors of observed variables relating to individual n and alternative j, β is a vector of fixed coefficients and μ is a vector of random terms with zero mean that depend upon individual n. The pattern of correlation in unobserved utility, and thus the substitution behaviour of the model, is determined by the variables that are introduced into z_{nj}. If μ is zero for all n then this becomes the standard MNL model exhibiting the IIA property (Glasgow, 2001; Train, 2009).

There has been limited use of the ML approach in station choice modelling, with only three recently published examples found. Chen et al. (2014, 2015) and Pang and Khani (2018) developed models based on the random parameters interpretation, with the former investigating station choice under uncertainty (using a non-linear utility function); and the latter modelling the choice of park and ride lot by transit users. Weiss and Habib (2017) used the error components interpretation to obtain non-proportional substitution patterns by specifying correlation between pairs of stations based on the distance between them. These three approaches are considered in more detail below.

Chen et al. (2014) initially proposed a framework for modelling station choice of park and ride passengers under conditions of uncertainty, where the utility function is based on prospect theory (Kahneman & Tversky, 1979). They suggest that as well as assessing the relevant factors (outcomes) of each alternative an individual must also make a judgement about the likelihood of these outcomes occurring. Factors such as how long it takes to get to the station or how long it takes to find a parking space are not definite but uncertain as they can be affected by traffic congestion or decisions made by other passengers on any particular day. Therefore the choice made by a passenger will depend on their attitude to risk. They propose a ML model where the utility function is based on prospect theory and incorporates a risk

aversion component, representing risk attitude and degree of risk attitude, and that takes the following general form:

$$PT(U) = \sum_{m} [w(p_m)v(x_m)], \tag{3.29}$$

where $PT(U)$ is the utility of an uncertain factor, obtained by multiplying the utility value of each factor attribute (x_m) by its respective probability (p_m) and then summing the product for all the attributes. The probability is included as a weighted probability function, $w(p_m)$, that can take various forms and is known as the 'risk weighting function' and the attribute utility is included as a value function, $v(x_m)$. They propose the following uncertain factors and associated attributes:

- Travel time to station $PT(V_{TT})$ — average travel time and variance of travel time
- Parking search time $PT(V_{PST})$ — average and variance of search time for each of free parking, paid parking and on-street parking
- Crowding on trains $PT(V_{VC})$ — low level, average level and high level crowding.

Prospect theory-based utility functions are defined for each of these factors, following the form in Equation 3.29 and these are then summed to obtain the utility function of each station alternative:

$$V = PT(V_{TT}) + PT(V_{PST}) + PT(V_C) + C, \tag{3.30}$$

where C represents all the other — non-uncertain — factors such as parking cost, train fare and so forth. The utility function is incorporated into a ML model for estimation. The framework was subsequently applied to estimate station choice under parking search time uncertainty (Chen et al., 2015). In a comparator MNL model, parameters for variation in parking search time and availability of parking bays were found to be significant at the 1% level. However, in the ML model where these were treated as random parameters they were not significant and apparently not random. The ML model did indicate that survey respondents were risk averse to variation in parking search time, and this model had a lower Akaike information criterion (AIC) than the MNL model, but its validity is questionable.

Pang and Khani (2018) developed several random parameter ML models of parking lot choice, using data from an on-board survey of passengers travelling on commuter rail and bus services run by Capital Metro, the regional public transportation provider for Austin, USA. The standard deviations of several estimated parameters (car access time, transit in-vehicle time, number of transfers) were significant, indicating that they vary within the population. They also applied an extension to the model allowing for correlation between the random parameters. This revealed that the parameter for car access time was positively correlated with the parameters for number of transfers and walk time (within the transit leg), indicating that travellers 'motivated' by a short car access trip are more likely to be 'motivated' by

fewer transfers and a shorter walk time. The ML models performed substantially better than a comparator MNL model, with log-likelihood (LL) increasing from −1035 in the best MNL model to −751 in the ML correlated coefficients model. A potential weakness of this work is the assumption made that the random parameters follow a normal distribution, implying that a parameter can take both a negative and a positive value. This is likely to be counter-intuitive for many of the explanatory variables, and a log-normal distribution might have been more appropriate (with variables expected to have a negative parameter entered as negative values).

The work of Weiss and Habib (2017) is particularly interesting as it is seeking, primarily, to address the issue of spatial correlation in the context of station choice (see Section 3.3.3). The proposed model, which they call the spatially weighted error correlation (SWEC) model, specifies a correlation between each pair of stations (in each individual's choice set) based on a function of distance between them. The function used is the inverse of the square root of the distance, which has the effect of increasing the correlation for stations that are closer together. For each off-diagonal in the covariance matrix the parameter is specified as follows:

$$\frac{b}{\sqrt{d_{ij}}}, \tag{3.31}$$

where b is the estimated parameter, i is the matrix column index and j the row index, and d is the distance between alternatives i and j. Several models, including a comparator MNL model, were estimated using revealed preference data obtained from a telephone survey of five percent of households within the Greater Toronto and Hamilton area, Canada. The SWEC models performed better than the MNL model in terms of goodness of fit, although not dramatically so. The best performing SWEC model had an adjusted McFadden's R^2 of 0.455 and a LL of −3144, compared with 0.419 and −3357 for the MNL model. However, the research unfortunately reveals nothing about the effectiveness of this approach in terms of generating realistic non-proportional substitution patterns, nor whether the predictive performance of the SWEC model is an improvement over the standard MNL model.

3.3.5 Model validation and testing

The overall performance of a discrete choice model is often assessed using a likelihood ratio index, which measures how well the model with its estimated parameters performs compared to a base model:

$$\rho = 1 - \frac{LL(FULL)}{LL(NULL)}, \tag{3.32}$$

where $LL(FULL)$ is the maximum log-likelihood with variables and $LL(NULL)$ is the maximum log-likelihood of the base model. There are two forms of the base model commonly used. The first is estimated under the assumption that each alternative has an equal chance of being chosen (the 'no information model'); and the second is estimated with alternative specific constants entered only, which implies that the probability of choosing an alternative

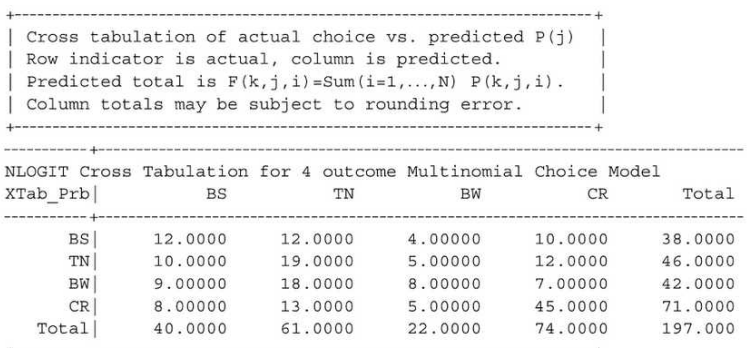
is the same as the actual market share of that alternative in the dataset, for each individual (see Hensher, Rose, and Greene (2016, pp. 446-456) for a fuller discussion). The most used likelihood ratio index is the adjusted McFadden's R-squared (rho-squared), which penalises for the number of predictor variables (k) included, especially if those variables do not sufficiently add to the explanatory power of the model:

$$R^2_{adj} = 1 - \frac{LL(FULL) - k}{LL(NULL)}. \tag{3.33}$$

However, it is important to note that it is only valid to compare models on the basis of their rho-squared if they have been estimated using identical samples and the same set of alternatives, i.e. when $LL(NULL)$ is the same for all the models (Train, 2009).

Another commonly used measure of model performance, adopted in a number of station choice studies (Blainey & Evens, 2011; Desfor, 1975; Fan et al., 1993; Harata & Ohta, 1986; Liou & Talvitie, 1974; Mahmoud et al., 2014) is predictive accuracy. For each individual, the alternative with the highest probability according to the model is identified and then compared with the choice that the individual actually made. Across all individuals, the percentage where both these match is referred to as the percent correctly predicted, and this might be used to compare the performance of different models. However, this approach is fundamentally flawed. The researcher is unable to say which alternative an individual will choose, as the true utility of each alternative is not known. That is why a probabilistic model was adopted in the first place. By definition, the choice with the highest probability will not always be chosen, it is just more likely to be chosen. Train (2009) gives the example of a model with two alternatives that have predicted choice probabilities of 0.75 and 0.25. This means that if 100 individuals were asked to choose between the two, 75 would be expected to choose one, and 25 the other. However, the percent accurately predicted procedure would assume that all 100 choose the one alternative with the highest probability. Train (2009) suggests that this performance measure should be avoided as 'the procedure misses the point of probabilities, gives obviously inaccurate market shares, and seems to imply that the researcher has perfect information'. A better measure is to compare the number of times an alternative was chosen with the sum of the predicted probabilities for that alternative across the sample, which can be presented as a contingency table (if the number of alternatives is not too large) as shown in Figure 3.11 (Hensher et al., 2016). This approach enables the predictive performance of models estimated on different samples to be compared.

Validating a predictive model against the sample used to calibrate it can lead to optimistic performance estimates. Additional validation can include testing the model on similar but independent data, for example by splitting the data into two parts and using one to develop the model and the other (the hold-out sample) to measure its performance, or by using advanced techniques such as cross-validation or bootstrapping (Steyerberg et al., 2001). The validation and testing methods used in prior station choice research are summarised in Table 3.3, and it is apparent how little testing has been carried out. The two earliest studies tested



Cross tabulation of actual choice vs. predicted P(j)
Row indicator is actual, column is predicted.
Predicted total is F(k,j,i)=Sum(i=1,...,N) P(k,j,i).
Column totals may be subject to rounding error.

NLOGIT Cross Tabulation for 4 outcome Multinomial Choice Model

XTab_Prb	BS	TN	BW	CR	Total
BS	12.0000	12.0000	4.00000	10.0000	38.0000
TN	10.0000	19.0000	5.00000	12.0000	46.0000
BW	9.00000	18.0000	8.00000	7.00000	42.0000
CR	8.00000	13.0000	5.00000	45.0000	71.0000
Total	40.0000	61.0000	22.0000	74.0000	197.000

FIGURE 3.11: Example of contingency table of predicted choice outcomes produced by NLOGIT. Note: Reprinted from '*Applied choice analysis*' (2nd ed.), by Hensher, D. A., Rose, J. M., & Greene, W. H., 2016, p. 501, Cambridge University Press. Image reproduced with permission of the rights holder, Cambridge University Press.

models against data from a new location (Liou & Talvitie, 1974) or an additional survey (Desfor, 1975), with the 'percent correctly predicted' measure suggesting they performed well; and Lythgoe and Wardman (2002, 2004) estimated demand for two new parkway stations, although the model substantially under-predicted demand. Only Sharma et al. (2017) have reported a rigorous process for testing model performance against a hold-out sample.

3.4 Obtaining and preparing choice data

3.4.1 Data sources

The choice data used in prior station choice research has usually been obtained from revealed preference (RP) OD passenger surveys carried out at stations or on trains. Two exceptions are Wardman and Whelan (1999) who combined data from 4,000 on-train and postal surveys with some 29,000 observations from a stated preference (SP) exercise; and Chen et al. (2015) who used a SP survey of 600 rail users at seven stations. The primary advantage of using RP data is that it reflects actual choices made by individuals; and in the case of station choice modelling a variety of data sources are available (in the UK at least) from which attributes that may explain those choices can be obtained. In contrast, SP data is based on what individuals *say* they would do under hypothetical choice situations, and this may differ from what they actually do (Train, 2009). However, such data is useful when information on actual choices is not available, such as a new product, or when attributes that explain actual choice are not readily available. For example, the attributes used by Chen et al. (2015) in their SP survey included: 'usual parking search time' and the 'probability that the worst parking search time occurs in one month'. For a detailed discussion of RP and SP and their relative merits see Train (2009, pp. 152-156) and Boyce and Williams (2016, pp. 219-229).

	Validation		Testing	
Author(s)	Fit (adjR2 unless stated)	Predictive accuracy (%)	Method	Predictive accuracy (%)
Liou and Talvitie (1974)		SC: 79-82; MC: 92	Tested against data from two other railroads.	SC: 96-98 MC: 86
Desfor (1975)		PnR: 91; KnR: 88	Tested against additional survey dataset.	PnR: 87; KnR: 90
Harata and Ohta (1986)		SC: 88-90; MC: 85-89		
Kastrenakes (1988)	0.325			
Fan et al. (1993)	0.904 (SC); 0.656 (MC); 0.614 (subway)	92.3 (SC); 78.6 (MC); 67.9 (subway)		
Wardman and Whelan (1999)	0.0837 (with respect to constants)			
Lythgoe and Wardman (2002, 2004)	0.532		Demand forecast for two new parkway stations - Warwick and East Midlands. Warwick flows to London under-forecast by 28%.	
Lythgoe et al. (2004)	0.611			
Debrezion et al. (2007a)	0.377 (linear); 0.274 (cross-effect); 0.410 (translog)			
Debrezion et al. (2009)	0.251			
Blainey and Evens (2011)	0.795 (NE); 0.632 (Wales)	83 (NE); 72.3 (Wales)		
Chakour and Eluru (2014)	BIC (11288.90)		Hold-out sample (no results)	
Mahmoud et al. (2014)	0.53 (rail and subway); 0.24 (rail); 0.32 (subway)	76.42 (rail and subway); 75.23 (rail); 79.18 (subway)		
Chen et al. (2015)	AIC (MNL: 1873.6; ML: 1644.2)			
Weiss and Habib (2017)	0.419 (MNL), 0.455 (ML*); AIC (MNL: 6732; ML: 6311*)			
Sharma et al. (2017)	0.779(RUM); 0.785(RRM*). RRM better fit (Ben-Akiva and Swait test).		30% hold-out sample (10% samples drawn 10 times from hold-out)	64.7 (RUM); 70.6 (RRM*)
Pang and Khani (2018)	LL (MNL*: -1035; ML*: -765)			

SC = station choice; MC = mode choice; PnR = Park and ride; KnR = Kiss and ride; BIC = Bayesian Information Criterian; AIC = Akaike Information Criterion; NE: North East England; RUM = random utility maximisation, RRM = random regret minimization, * = best model

TABLE 3.3: Summary of validation and testing of station choice models used in prior research (not exhaustive).

Passenger surveys may be at the national level, for example Blainey and Evens (2011) and MVA Consultancy (2011) used data collected in Britain by the NRTS during 2004–2005, or at the local or regional level, such as the survey of commuter rail lines carried out by New Jersey transit (Kastrenakes, 1988). An alternative approach was adopted by Desfor (1975) who collected licence plate numbers from cars that were parked or dropping off passengers at stations, and used the registered addresses of the vehicle owners as a proxy for trip origin. In contrast, the models developed to assess demand for stations on the planned high speed rail line between London and the West Midlands in the UK (HS2) were not based on any observed station choice data. Rather than calibrating a model to estimate parameters, GJTs were calculated using established elasticities from an existing multi-modal model (see Section 3.3.1 for more details). The approach adopted by Lythgoe and Wardman (2002, 2004) does not require data on ultimate trip origins or destinations as the dependent variable is not observed station choice but the number of rail trips on particular flows derived from ticket sales data (see Section 3.6 for more details). Table 3.1 includes information on the survey size and data type used in prior station choice studies.

3.4.2 Disaggregate vs. aggregate

Discrete choice models are often thought of as disaggregate-only models which are estimated using data at the individual level. However, the dependent variable can also be the observed share of each alternative at some unit of aggregation, and this approach has been adopted in some studies. For example, Debrezion et al. (2007a) used the observed proportion of the three most frequently chosen stations at postcode area level as the dependent variable in an MNL model, and Debrezion et al. (2009) estimated an NL model with the proportion of *joint* access mode and station choice for each postcode area as the dependent variable, with 12 choice combinations per postcode area (three alternatives per area and four access modes). In both cases, although the original data was disaggregate and obtained from an OD survey carried out by the Dutch Railway Company, it was supplied to the researchers in an aggregated form. In another study, Kastrenakes (1988) had access to disaggregate data from 26,000 responses to an OD survey of nine commuter lines, but chose to aggregate it at the minor civil division level, a decision that probably reflects the capabilities of the analytical software available at that time.

There are several consequences of aggregating data prior to model estimation: it is statistically inefficient as data from many individual observations is grouped into a relatively small number of zone-based observations; the model is unable to account for intra-zonal variability (for example, the access distance to a station is treated as being the same for an entire zone); and there is the potential for statistical bias, for example caused by the issue of 'ecological fallacy' (Ortuzar, 1980). An ecological fallacy occurs when results from a model estimated using zonal data are assumed to also apply to the individual observations that make up the zones. This would only be true if the zones were homogeneous, which is rarely the case,

and the degree of intra-zonal heterogeneity will determine the extent to which ecological fallacy is a problem. This could mean that a variable that is not significant in an aggregate model may in fact be a significant factor in choice at the individual level, and vice versa. Ecological fallacy is closely related to the modifiable areal unit problem (MAUP), which is a consequence of the arbitrary nature of zones, which can vary in size (the 'scale problem') or vary in composition (the 'aggregation problem') at the whim of the researcher. Different decisions regarding the size and composition of zones can result in different model results, for example as scale increases correlation coefficients tend to increase (Openshaw, 1984). Fotheringham and Wong (1991) examined the impact of MAUP on the calibration of a logit regression model and found it to be sensitive to both the scale and aggregation problems and 'to produce highly unreliable results'. They suggest three potential solutions: report results using different aggregation scales and zone structures; attempt to create 'optimal zoning systems' that maximise inter-zonal variation and minimise intra-zonal variation (though what is optimal might not be the same for all variables); or avoid using aggregated data. Using disaggregated data to calibrate models does, however, present a problem of its own. How can the results of these models be used in the aggregate models required to forecast station demand? This issue is considered further in Section 3.6.

3.4.3 Defining choice sets

A choice set must meet three conditions to be consistent with the discrete choice framework. First, the alternatives must be mutually exclusive; second, the number of alternatives must be finite; and third, the choice set should include all possible alternatives (Train, 2009). A passenger can only depart from and arrive at a single railway station, and there are clearly a finite number of stations in any choice set, so the first two requirements are met. The third is more problematic, as the researcher usually only knows what choice was ultimately made (unless data is from an SP survey). The choice set will depend on the stations which are feasibly available based on a passenger's origin and destination, but will also vary on an individual basis, influenced by socio-demographic characteristics, level of knowledge, attitudes and perceptions (Basar & Bhat, 2004). The choice set might also be constrained in certain circumstances. For example, if an individual can only walk to a station, then there must be a cut-off distance at which a station is no longer considered feasible. A feature of logit models is that an alternative can never have a probability of zero, and if an alternative has no realistic prospect of being chosen it can be excluded from the choice set (Train, 2009). However, setting a threshold is fraught with difficulties, and often a fuzzy concept. How, for example, can the appropriate cut-off distance for walk access to a station be set, when it will surely vary on an individual basis?

Castro, Martinez, and Munizaga (2009) highlight the potential for 'serious problems' with model predictions if the choice set is poorly specified and argue that while in some circumstances it might be plausible to exogenously define feasible alternatives, for example in the

case of travel mode choice, in other situations, such as when modelling spatial alternatives, it becomes very complex or arbitrary. A potential solution is to use a probability-based approach, for example the two-stage MNL model developed by Basar and Bhat (2004) to study airport choice, where the probability of an alternative being in an individual's choice set is modelled first.

A range of methods with varying degrees of complexity have been adopted for defining choice sets in the field of station choice modelling, and these are summarised in Table 3.4. Most methods can be split into one of three groups, based on distance, observed choice, and catchments. In the distance-based method each individual has their own choice set determined by the closest x stations to their origin, with the aim of maximising the number of observed choices accounted for, while keeping the number of alternatives to a reasonable number (Blainey & Evens, 2011; Fan et al., 1993; Mahmoud et al., 2014; Weiss & Habib, 2017). In the observed choice method, the choice set is defined at the area level, for example the stations chosen by passengers living in a particular locality (Kastrenakes, 1988) or the most frequently chosen stations in a postcode area (Debrezion et al., 2009). The catchment-based method assigns a catchment of a certain radius to each station, and this determines whether an alternative is within either an individual or area-based choice set (Adcock, 1997; Lythgoe & Wardman, 2004). This method is of some concern, especially for models that aim to improve demand prediction, as the main advantage of modelling station choice is to *overcome* the inadequacies of defining station catchments in this way, as discussed in Section 2.4.1. Unusually, Adcock (1997) used alternative rail legs from trip origin to destination as the choices, rather than stations, reflecting that the entire door-to-door trip was modelled.

An interesting alternative method was adopted by Chakour and Eluru (2014), based on the concept of the maximum distance passengers are willing to travel relative to their nearest station (D):

$$D = \frac{\text{Distance to chosen station - Distance to closest station}}{\text{Distance to closest station}}. \tag{3.34}$$

This ratio was calculated for every individual in the dataset and the 95th percentile was taken as the threshold value. For each individual a D ratio was then calculated for all stations in the study area (replacing the chosen station in the ratio), and only those stations with a D ratio less than the threshold were included in the individual's choice set. They found that using a single ratio to determine the threshold was problematic, as someone living very close to their nearest station is likely to be willing to travel much further relative to that distance than someone whose nearest station is a much greater distance from their home. To address this they calculated a separate threshold value for five 'distance to nearest station' bands. The resultant choice sets varied in size from 1 to 18 alternatives, with 91% containing between 1 and 5 stations.

Pang and Khani (2018) and Sharma et al. (2017) both adopted particular strategies to select a manageable number of alternatives for each individual from large universal choice sets (188 and 418 respectively). Both these studies are concerned with park and ride lot choice,

and were not restricted to modelling station or subway choice as they also included bus services. Pang and Khani selected the alternatives based on 'thresholds' of 15 minutes and 50 minutes which were applied to the shortest possible access leg and the total trip time (origin to destination) respectively. The thresholds were identified from a sensitivity analysis and were set to account for 90% of observed choice. Sharma et al. first removed any alternatives where the access leg time would be greater than the total time from origin to destination, and then selected 19 at random from the remainder (plus the chosen alternative).

3.5 Measuring representative utility

3.5.1 How do passengers choose a station?

Ideally there would be a body of behavioural research exploring the station choice decision process which could be drawn upon to inform model development. This might answer questions such as: 'how many stations do passengers consider?', 'what information do passengers evaluate in making their choice?', 'what information sources are used?' and 'how much effort and time do passengers put into weighing-up the pros and cons of alternative stations?'. Unfortunately there does not appear to be any research of this nature, so it is necessary to draw upon other sources of information to guide model development.

Stated preference surveys can give useful insights into factors that are important to passengers. For example, Adcock (1997) carried out a review of stated preferences surveys that had been commissioned to assess proposals for station development, and identified the following factors as particularly important to passengers: generalised journey time (consisting of actual journey time, transfer penalties, and service frequency penalties); fare; access and egress distances; ease of car parking; ease of road access; level of car ownership; and journey purpose. Due to data availability, only the first three were included in the models subsequently developed. More recent stated preference surveys can give further insights, for example a study into customer priorities for released capacity on the West Coast Mainline identified crowding on trains and interchange as the two factors that most influence the quality of the rail journey experience for existing passengers. More specifically, the value of rail falls significantly as soon as a passenger does not have a seat, and passengers want direct services — the waiting time between trains is of little relevance as they would rather have no change at all (Passenger Focus, 2012).

Chakour and Eluru (2014) approached the problem by including a broad range of variables in their station choice models — relating to individual and household socio-demographics, the trip, levels of service, station, land-use and the built environment. During model calibration statistically insignificant variables were systematically removed in a process 'guided by intuition and findings from earlier literature'. Kastrenakes (1988) also tested a range of variables in different combinations and found many of them to be 'noncontrolling of rail riders'

Study	Choice set definition - station or route	Level defined	Additional constraints
Liou and Talvitie (1974)	Alternatives 'chosen on the basis of data and were usually near the chosen station'	Unknown	
Desfor (1975)	Least cost and second least cost (determined by a cost function) for each Census Block.	Area	
Harata and Ohta (1986)	Four alternate routes (origin to a common station).	Individual	Walk access mode not available if station >3km from origin; bus access mode not available if no stop within 800m of origin.
Kastrenakes (1988)	Observed station choice for each municipality. Each municipality has own choice set.	Area	
Fan et al. (1993)	Commuter rail: Five closest stations measured by straight-line distance to the passenger's home (accounts for 98% of observed choice). Subway: The two closest stations on the two closest lines measured by straight-line distance from the passenger's home (accounts for 95 of observed choices on a station basis and 99 percent on a line basis).	Individual	Only closest station available for walk access.
Adcock (1997)	Up to ten alternative rail legs (origin to destination) for each rail trip in the dataset were selected, with a catchment of 15km radius assumed for most stations, 35km for large stations (15km was found to account for $\geq$ 90% of passengers).	Individual	
Wardman and Whelan (1999)	Two stations (no details of selection method provided)	Individual	
Lythgoe et al. (2004)	Each station assumed to have a 20km catchment divided into zones. Potential competing stations had to be within 20km of at least one zone of another station and were then ranked by criteria, with the top 15 making up the choice set.	Area	
Debrezion et al. (2007a)	The three most frequently chosen stations for each postcode area	Area	
Debrezion et al. (2009)	Three most used stations for each postcode area.	Area	
Blainey and Evens (2011)	Nearest 10 stations to trip end, measured by network distance.	Individual	
Fox et al. (2011)	'iterative process' to select the 5 'best' alternatives, where the most attractive stations determined from an earlier model form the choice set for the next model.	Individual	
Chakour and Eluru (2014)	Based on the concept of the maximum distance passengers are willing to travel relative to their closest station.	Individual	Individual Transit not available as access mode choice if station 'very close' to origin, or if no transit stop within 37 minutes walk.
Givoni and Rietveld (2014)	All 11 stations in the Amsterdam area.	Universal	
Mahmoud et al. (2014)	5 closest stations for commuter rail (accounts for 98% of trips) and 3 closest stations for subway model (accounts for 80% of trips) .	Individual	
Sharma et al. (2017)	For each observation: universal choice set (of 418) reduced by removing those where access leg time > origin to destination time; then a random sample of 19 selected, plus the chosen station.	Individual	
Weiss and Habib (2017)	4 closest stations by drive time, plus the chosen station if not included	Individual	
Pang and Khani (2018)	Alternatives selected from universal choice set (188) based on 'thresholds' of 15 minutes and 50 minutes applied to the shortest possible access leg and the total trip time (origin to destination) respectively. Chosen alternative added if not included.	Individual	

TABLE 3.4: Summary of choice set specifications used for station choice models.

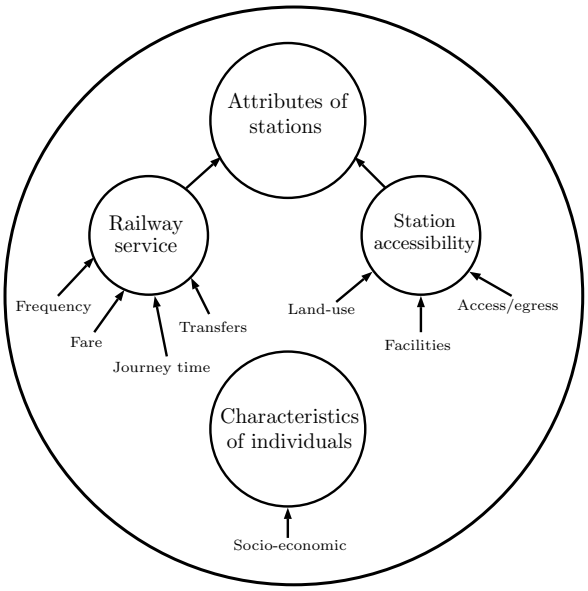


FIGURE 3.12: The type of factors that influence the decision to choose one station over another.

station choice.' This reflects the general approach of the prior research, which concentrates primarily on model structures and gives less attention to selecting and defining attributes. There is research in closely related fields which could inform the selection and definition of attributes which has rarely been drawn upon. For example, there is a sizeable body of work relating to station accessibility, covering themes such as walking and cycling (Park, Kang, & Choi, 2014; Puello & Geurs, 2015; Zhao et al., 2003), access mode (Cervero et al., 1995; Guan et al., 2007), accessibility for the elderly (Lin et al., 2014), the access journey (Givoni & Rietveld, 2007; Keijer & Rietveld, 2000; Passenger Focus, 2007, 2011), the door-to-door journey (Brons & Rietveld, 2009), and the role of the built-environment (Cervero et al., 1995; Jiang, Zegras, & Mehndiratta, 2012). There has also been research into the potential to increase demand by improving access to stations (Brons, Givoni, & Rietveld, 2009; Giannopoulos & Boulougaris, 1989; Wardman & Tyler, 2000) and the effect of station enhancement on rail demand (Hagen & Heiligers, 2011; Preston et al., 2008).

It is useful to group the factors that influence the decision to choose one alternative over another into two groups, one containing attributes of the alternatives and one containing characteristics of the decision makers (Ortúzar & Willumsen, 2011). In terms of station choice, the attributes of the alternatives can be further grouped into those relating to station accessibility, such as distance to the station, and those relating to the railway service provided from a station, such as frequency of service (Givoni & Rietveld, 2007). Choices also depend upon the prejudices and tastes of individuals, and it may be possible to represent some of these characteristics in models by introducing variables based on socio-economic data (Ortúzar & Willumsen, 2011). The interplay of the type of factors involved is illustrated in Figure 3.12.

3.5.2 Accessibility attributes

3.5.2.1 Access and egress

The most common variable included in previous research is access distance from the trip origin[16] to the departure station, with increasing distance expected to have a negative effect on station choice. Most studies have used the straight line measure for access distance (Adcock, 1997; Debrezion et al., 2007b, 2009; Desfor, 1975; Mahmoud et al., 2014), which is unlikely to reflect the true distance travelled by any chosen access mode. This can be improved upon by measuring distance via the road network or cycle path network (Blainey & Evens, 2011; Fan et al., 1993; Givoni & Rietveld, 2014; Sharma et al., 2017). Distance is normally included as a continuous variable, although Debrezion et al. (2007a) created a series of distance bands which were entered into the model as dummy variables. This approach allows a separate coefficient to be estimated for each band and for the changing effect of distance on utility to be represented. They found the coefficient was positive for all bands, relative to the furthest band ($>$ 10,000 m) which was excluded as the reference, with higher coefficients for lower distances and a smooth decline as distance increases. The utility of distance was seven times higher at 250 m that at $>$ 10,000 m.

An alternative to access distance is estimated travel time or in-vehicle time for the access trip, which again is expected to have a negative effect on station choice. This may simply be distance converted into time (Kastrenakes, 1988), or a more accurate reflection of journey time by access mode, for example public transport (Debrezion et al., 2009; Givoni & Rietveld, 2014) or car (Chen et al., 2015; Fox, 2005; Pang & Khani, 2018; Weiss & Habib, 2017). Travel time is intuitively a more appropriate measure, as when a passenger is weighing up the relative utility of two stations it will be the length of time it takes to get to/from a station or the total journey time that is the important factor to them, rather than the actual distance travelled which is unlikely to be known in most instances. Clearly there will be a correlation between distance and time, but travel time can be influenced by a range of factors other than distance such as the class of road and flow conditions.

If only a single parameter is estimated for access distance this will represent an average effect on utility across the different access modes. However, this effect would be expected to vary, with a larger negative coefficient for non-motorised access modes compared with motorised. In a NL model with access mode at the upper level, this can be accommodated by specifying a different parameter for distance or time in the utility function of each nest, which was the approach adopted by Debrezion et al. (2009). However, in the MNL model developed by Blainey and Evens (2011), only a single coefficient is estimated for access distance. This model could potentially be improved by using dummy variables representing each access mode interacted with access distance. For example, suppose there are three access modes

[16] The trip origin is commonly the address postcode (at varying degrees of spatial resolution), though some studies have used geocoded home addresses (Fan et al., 1993; Mahmoud et al., 2014; Pang & Khani, 2018; Weiss & Habib, 2017), and Desfor (1975) used the census block centroid.

(car, bus, walk), the access distance component of the utility function could be modified from $V_{ik} = \beta Ad_k$ to:

$$V_{ik} = \sum_{m=1}^{3} \beta_{mdist}(Dmode_{km} \times Ad_k), \tag{3.35}$$

where $Dmode_{km}$ is a dummy variable with value 1 if individual i uses access mode m, and zero otherwise; Ad_k is access distance; and β_{mdist} is the parameter for mode m.

There has been relatively little attention given to the egress journey, with Adcock (1997) including egress distance, and Chakour and Eluru (2014) including a dummy variable to identify whether the egress mode was public transport. Adcock found that passengers were willing to accept longer access journeys than egress journeys and suggests this could be due to the availability of a car or better knowledge of public transport options at the home end. It is possible that the availability of a bicycle or better knowledge of cycling or walking routes at the home end might also result in acceptance of longer access journeys. Cervero et al. (1995) found that walking was the predominant egress mode for a greater distance than it was the predominant access mode. This issue of 'asymmetry' of private access/egress transport modes has been identified by Keijer and Rietveld (2000) who found that 35% of passengers used a bicycle to access the station, but only 10% cycled from the egress station to their final destination. They also found that public transport was more important for the egress leg than the access leg, but as their study only included trips where the destination was *not* the home this may reflect the higher availability of public transport in city centre destinations compared with residential origins.

In some studies factors relating to the access journey are incorporated into a composite measure of generalised cost or generalised journey time (Lythgoe & Wardman, 2004; MVA Consultancy, 2011), and other less often used variables include the cost of the access journey, such as car cost or bus fare (Fox, 2005; Liou & Talvitie, 1974; Wardman & Whelan, 1999), and the frequency of public transport (Debrezion et al., 2009; Wardman & Whelan, 1999).

Kastrenakes (1988) used a 'local to users' dummy variable to indicate whether a station was located in a particular minor civil division and therefore considered the local station to passengers living in that minor civil division. Interestingly, this variable was not correlated with access time and Kastrenakes suggests that it may be capturing 'intangibles' such as a greater awareness of services and parking within a passenger's home town. Due to the aggregate nature of the study, access time was the average from the residential centre of each minor civil division to each of the alternative stations, and this could have masked a potential correlation at the level of the individual. However, similar variables have been included in several disaggregate studies. Fan et al. (1993) used a dummy variable to indicate whether a station was the closest to home (trip origin) of the choice alternatives, and found that including it resulted in a better model and exerted a 'strong bias effect'. Adcock (1997) included a 'nearest station used' dummy, in addition to access distance, and found that this was a particularly important factor for season ticket holders, perhaps reflecting the prior choices that these passengers have made about where to live.

Several studies have used a variable to capture any potential preference amongst commuters for an origin station that is broadly in the same direction of travel from their home as their workplace. Both Mahmoud et al. (2014) and Weiss and Habib (2017) calculated the angle in degrees between a straight line from the origin to the workplace and a straight line from the origin to the station. Mahmoud et al. entered the measure as a continuous variable, while Weiss and Habib used a dummy variable to indicate if the angle was greater than 90 degrees. A negative effect on utility was reported in both cases, indicating that passengers prefer a station to be in a similar direction of travel as their destination. An improved measure could be derived that is based on travel on the access network, rather than using a straight line that may not reflect the network routes available. It is also possible that the size of this effect depends on the distance to the origin station, as the direction of travel might be of little consequence for very short access journeys. This could be explored by calculating a separate parameter for different access journey length bands. A related variable used by Chakour and Eluru (2014) was the distance by rail from each station to the central business district, which was found to have a negative coefficient. As the central business district was the assumed destination on the commuter lines studied, this indicates that passengers prefer stations that are in the direction of the destination. There is a potential relationship between these variables and overcrowding, as it is possible a passenger would choose an earlier station on the line, in order to guarantee that they got a seat, and thus their access trip would be in the opposite direction to their destination.

Pang and Khani (2018), included the number of intersections on the car access route, but did not report it in their final models, presumably as its effect was not significant. A similar measure that has not been considered in previous station choice research, but is common in accessibility studies, is the directness of the route. This can be calculated as the ratio of the network distance to the straight line distance. A more direct route might indicate that access is easier, though this may vary by access mode and could be more relevant to passengers who walk or cycle to a station (Lin et al., 2014).

A summary of the full range of factors related to the access and egress journey used in prior station choice research is provided in Table 3.5.

3.5.2.2 Facilities

Car parking is the dominant station facility attribute considered in prior studies, and has been represented in a variety of forms, such as a dummy variable indicating the presence of a car park (Debrezion et al., 2007a, 2009; Liou & Talvitie, 1974), the number of parking spaces (Blainey & Evens, 2011; Chakour & Eluru, 2014; Chen et al., 2015; Fan et al., 1993; Fox, 2005; Mahmoud et al., 2014; Pang & Khani, 2018; Weiss & Habib, 2017), the availability of free spaces (Kastrenakes, 1988), and parking cost/fee (Chen et al., 2015; Desfor, 1975; Kastrenakes, 1988; Mahmoud et al., 2014; MVA Consultancy, 2011). In most cases the presence of a car park and the number of parking spaces has a positive effect on station

Author(s)	Factors considered
Liou and Talvitie (1974)	Out-of-vehicle time, in-car time, on-bus time, car operating cost, out-of-pocket cost (parking/bus fare), total cost (operating plus out-of-pocket)
Desfor (1975)	Distance to station (straight line, part of cost function)
Harata and Ohta (1986)	Walk time (walk, bus), wait time (bus), in-vehicle time (bus)
Kastrenakes (1988)	Local station (dummy), access time (shortest route)
Fan et al. (1993)	Access time plus rail in-vehicle time (transit and car), transit fare, closest station (dummy for car access mode), walk distance (walk mode)
Adcock (1997)	Access and egress distance (straight line), nearest station used (dummy)
Wardman and Whelan (1999)	Access time, access cost (by journey reason: commute, business, leisure), bus headway
Lythgoe and Wardman (2004)	Time and cost (by car) of accessing origin station (used within a GC)
Fox (2005)	Driving cost, in-car time
Debrezion et al. (2007a)	Distance to station (range of categories as dummies, straight line)
Debrezion et al. (2009)	Distance to station (straight line), travel time by PT, PT frequency (services per hour)
Blainey and Evens (2011)	Distance to station (road network)
MVA Consultancy (2011)	Access time (part of GJT)
Chakour and Eluru (2014)	Time to closest station, average time to viable stations, time to chosen station, egress mode is transit, distance from station to CBD
Givoni and Rietveld (2014)	Car distance, public transport travel time, taxi distance, bicycle distance, walking distance, other distance (all distances by network)
Mahmoud et al. (2014)	Access distance (straight line), direction of station in degrees from home relative to regular work place
Chen et al. (2015)	Access time
Weiss and Habib (2017)	Access time (drive), drive cost (part of total trip cost), direction of station from home relative to work place > 90 degrees (dummy)
Sharma et al. (2017)	Access distance (network), parking lot within 1km of a freeway, parking lot within CBD
Pang and Khani (2018)	Access time (car); number of intersections; proportion of access route on highway (rather than local streets)

TABLE 3.5: Summary of access and egress factors used to construct utility functions in station choice models.

choice, although there have been conflicting results and counter-intuitive coefficient signs in some cases. For example, Fan et al. (1993) developed two models for the Greater Toronto area, one for commuter rail and one for subway, and while the coefficient for parking availability was positive and significant in the commuter rail model, it was 'not useful' in the subway model. They raise an interesting point that if a passenger arrives at their first choice station to find no spaces available they have no option but to try another station, but as far as the model is concerned their revealed choice will be the second station. As subway stations are closer together and have a more frequent service, a passenger can drive on to another station comfortable in the knowledge that they will have a short wait for the next train, possibly explaining why the subway model is insensitive to parking capacity. Kastrenakes (1988) attempted to include parking fee and parking availability variables, but both resulted in significant but counter-intuitive coefficients, implying that passengers are more likely to choose a station as fees increase or as parking availability decreases, and they were excluded from the final model. However, there are likely to be endogeneity issues at play here and, as Kastrenakes notes, a high parking fee and lack of availability could both result from a station being very popular. In addition, a positive coefficient for the number of parking spaces may

not indicate that more spaces attract passengers but that more passengers lead operators to provide more spaces (Chakour & Eluru, 2014). Adcock (1997) notes that the size of the station car park might not be a good measure of parking availability, as other car parks or on-street parking may be local to the station. He proposes a car park valuation exercise as a promising alternative, given that a small-scale survey indicated a good correlation between highly valued car parks and the proportion of railheaders. However, this may be impractical for large area studies, and he suggests the calibration of station-specific 'attractiveness' parameters based on revenue data from MOIRA that could capture the effects of car parking and other station facilities.

Another potential difficulty is that parking-related variables will only be relevant to decision-makers who drive to (and potentially those dropped off at) a station. While this can be accounted for in NL models by specifying mode-specific utility functions, it is a problem for MNL models. For example, Blainey and Evens (2011) included a variable for the number of car parking spaces in an MNL model, but a descriptive analysis of the data reveals that car is a minority access mode. There is potential to improve such a model by using a dummy variable representing car as access mode interacted with the car parking spaces variable, with the car parking spaces component of the utility function being modified from $V_{ik} = \varphi Ps_k$ to:

$$V_{ik} = \varphi_{Ps}(Dcar_k \times Ps_k), \tag{3.36}$$

where $Dcar_k$ is a dummy variable with value 1 if individual i uses the car as access mode, and zero otherwise; Ps_k is the number of parking spaces and φ_{Ps} is the parameter for the parking spaces variable (which will only be estimated against those observations where the access mode was car).

A summary of the full range of factors related to station facilities (and land-use) used in prior station choice research is provided in Table 3.6.

3.5.2.3 Land-use

Only Chakour and Eluru (2014) have incorporated land-use variables in models of station choice. They identified six characteristics of Montreal traffic analysis zones using principal component analysis, such as high density/high walkability, commercial, or government/institutional. However, neither choice of access mode or choice of station was found to be elastic with respect to these variables, with a 15% uplift resulting in a change of less than 1% in access mode share or station choice, leading them to conclude that access mode and station choice 'do not react to land-use changes'. However, research in related areas suggests that land-use might play an important role in station choice. For example, Cervero et al. (1995) found that residential density and land-use mix influence how passengers access stations and the size of access catchments.

Type of factor	Author(s)	Factors considered
Facilities	Liou and Talvitie (1974)	Available parking space
	Desfor (1975)	Parking costs (part of cost function)
	Kastrenakes (1988)	Parking availability and fee (both had counter-intuitive signs and excluded from models)
	Fan et al. (1993)	Number of parking spaces (natural logarithm, for car access mode)
	Wardman and Whelan (1999)	Facilities at station, parking availability
	Fox (2005)	Number of park and ride spaces
	Debrezion et al. (2007a)	Park and ride facility (dummy)
	Debrezion et al. (2009)	Parking area (dummy), bike stands (dummy)
	Blainey and Evens (2011)	Number of car parking spaces
	MVA Consultancy (2011)	Car park cost (part of GJT)
	Chakour and Eluru (2014)	Size of parking lot (range of categories as dummies)
	Givoni and Rietveld (2014)	Quality of parking space, quality of guarded bike parking facility
	Mahmoud et al. (2014)	Park-and-ride lot capacity, parking cost at morning peak, refreshment kiosk (dummy), washroom (dummy), reserved parking (dummy)
	Chen et al. (2015)	Parking capacity, parking fee, parking fine and control frequency (around station), various attributes related to parking search time
	Weiss and Habib (2017)	On subway (dummy), lot capacity (natural logarithm), washroom (dummy),
	Sharma et al. (2017)	Served by trains (dummy), formal parking
	Pang and Khani (2018)	Designated PnR (dummy), has a rail service (dummy), has express bus service (dummy), no. of parking bays
Land-use	Chakour and Eluru (2014)	Government and institutional areas (at origin or at station), commercial area

TABLE 3.6: Summary of facility and land-use related factors used to construct utility functions in station choice models.

3.5.3 Railway service attributes

Attributes used to represent railway service quality include measures of train frequency, such as trains per hour, per day or at peak periods (Blainey & Evens, 2011; Debrezion et al., 2007a; Fan et al., 1993; Kastrenakes, 1988; Pang & Khani, 2018); rail journey time (Fox, 2005; Givoni & Rietveld, 2014; Harata & Ohta, 1986; Liou & Talvitie, 1974; Pang & Khani, 2018; Weiss & Habib, 2017); journey distance (Blainey & Evens, 2011); fare (Adcock, 1997; Fox, 2005; Harata & Ohta, 1986; Sharma et al., 2017; Weiss & Habib, 2017); and number of transfers (Fox, 2005; Harata & Ohta, 1986; Pang & Khani, 2018). In some cases a single measure of GJT, derived from several railway service attributes, has been used (Adcock, 1997; Atkins Limited, 2011; Kastrenakes, 1988; Lythgoe & Wardman, 2004; MVA Consultancy, 2011). The aforementioned measures have the intuitively expected effect on utility in all studies, with the exception of Blainey and Evens (2011), where a positive coefficient for journey distance was obtained for South Wales. Distance may not be a good proxy for time, as a longer route could be faster dependent on the line running speed, stopping patterns and whether the service is direct, and this may have resulted in a misspecified model.

To explore the effect of train frequency on utility, Debrezion et al. (2007b) used two alternative utility function forms, one with a cross-effect linear additive function and the other with a transcendental logarithmic (translog) function. In the cross-effect function access distance categorical dummy variables were cross multiplied with station frequency of service, allowing the model to show the effect of service frequency on utility at different distance categories:

$$V_j = \sum_{c=1}^{21} \beta_{cfreq}(Dcateg_{jc} \times freq_j) + ..., \tag{3.37}$$

where $Dcateg_{jc}$ is 1 if station j is in distance category c, and zero otherwise; and $freq_j$ is number of trains per day. Results show that the positive effect of frequency on utility is greater for passengers living closer to a station ($\beta = 0.0717$ at 250 m), compared with those living further away ($\beta = 0.0016$ at 9,500–10,000 m). In the translog model, access distance and frequency are included individually as their natural logs and their squared natural logs, and as the product of their natural logs:

$$V_j = \beta_{dist} ln(dist_j) + \beta_{distsq}(ln(dist_j))^2 + \beta_{freq} ln(freq) + \beta_{freqsq}(ln(freq))^2 + \beta_{distfreq}(ln(dist) \times ln(freq)) + ..., \tag{3.38}$$

where *dist* is access distance as a continuous measure. This model is used to better understand how the train frequency effect changes with access distance. Results of this model show that utility declines smoothly as access distance increases for all frequency levels, but the curve is flatter for stations with higher service frequency, indicating that a station's catchment (or market area) is larger when frequency is higher. However, the size of this effect diminishes as frequency increases.

An interesting alternative approach to capturing the rail service attributes is the rail service quality index (RSQI) developed by Debrezion et al. (2009), where three determinants of rail service quality — frequency of trains (represented by waiting time); quality of connectivity to other stations (represented by transfer time); and relative position of the station on the network (represented by in-vehicle time between station pairs) — are combined into a single quality index. A doubly-constrained spatial interaction (flow) model, containing trip data from 365 stations in the Netherlands, was used to estimate coefficients that were then used to calculate an RSQI for each station. The flow model is similar in approach to a trip distribution model, with balancing factors that enforce the constraint, estimated for each origin:destination pair:

$$T_{ij} = A_i O_i B_j D_j f(GJT_{ij}) g(GJT_{ij}/d_{ij}) \exp(\varepsilon_{ij}), \tag{3.39}$$

where T_{ij} is the number of trips between stations i and j, O_i is the total number of trips originating at station i, D_j is the total number of trips attracted by station j, A_i and B_j are the balancing factors, $f(GJT_{ij})g(GJT_{ij}/d_{ij})$ is a two-part impedance function (where GJT contains waiting time, in-vehicle time, and transfer time), and ε is an error component. An

RSQI is calculated for each origin:destination station pair using estimated coefficients from the flow model, and an aggregate RSQI is then calculated for each station i by summing all the departure:destination RSQIs for that station:

$$RSQI_i = \sum_j \hat{B}_j D_j \hat{f}(GJT_{ij}) \hat{g}\left(\frac{GJT_{ij}}{d_{ij}}\right). \tag{3.40}$$

In a NL model with access mode at the upper level, the RSQI had a significant and positive effect on station choice. However, it should be noted that the RSQI approach was only necessary as trip destination data was not available in the passenger survey data used in this study. If it had been, then the attributes could have been used directly in the station choice model.

A summary of the full range of factors related to station facilities (and land-use) used in prior station choice research is provided in Table 3.7.

Author(s)	Factors considered
Liou and Talvitie (1974)	On-train travel time difference
Desfor (1975)	Return fare (part of cost function)
Harata and Ohta (1986)	In-vehicle time (rail), out-of-vehicle time (rail), cost (rail), number of transfers
Kastrenakes (1988)	Trains per hour, GJT
Fan et al. (1993)	Number of morning peak trains
Adcock (1997)	GJT (consisting of actual journey time, interchange and frequency penalties), fare, mileage travelled on London Underground
Wardman and Whelan (1999)	Journey time, journey headway, journey cost (by journey reason: commute/business/leisure)
Lythgoe and Wardman (2004)	Fare (part of GC), GJT (part of GC, consisting of rail travel time, no. of interchanges and headway between trains).
Fox (2005)	Fare, in-vehicle time, wait time, number of transfers, interchange walk time
Debrezion et al. (2007a)	Frequency (trains per day), intercity status (dummy for each region)
Debrezion et al. (2009)	Rail service quality index (constructed using a direct demand model)
Blainey and Evens (2011)	Train frequency, total distance from origin to destination station
MVA Consultancy (2011)	In-vehicle time, frequency penalty, interchange penalty, fare (part of GJT)
Atkins Limited (2011)	GJT (in-train time, waiting time, boarding penalty)
Chakour and Eluru (2014)	Train frequency, trip is in direction of central business district
Givoni and Rietveld (2014)	Rail journey time
Mahmoud et al. (2014)	Station has a connection to local or regional services, station is a regional transit station
Weiss and Habib 2017)	Fare (part of total trip cost), journey time (station to destination)
Sharma et al. (2017)	Transit fare, in-vehicle time (transit), wait time (transit leg)
Pang and Khani (2018)	Number of transfers; in-vehicle time (continuous and banded); walk-time (in transit leg); total transit travel time; frequency/hour (natural logarithm)

TABLE 3.7: Summary of railway service related factors used to construct utility functions in station choice models.

3.5.4 Socio-economic attributes

Some of the models developed in previous studies have included socio-economic attributes, mostly relating to age (Chakour & Eluru, 2014; Fan et al., 1993; Fox, 2005; Pang & Khani, 2018), sex (Chakour & Eluru, 2014; Fan et al., 1993; Fox, 2005; Pang & Khani, 2018), income (Fan et al., 1993; Liou & Talvitie, 1974; Pang & Khani, 2018) and car ownership (Chakour & Eluru, 2014; Debrezion et al., 2009; Fox, 2005; Pang & Khani, 2018).

A particularly important feature of discrete choice model theory is that only the difference in utility between alternatives is relevant to the decision maker. The absolute value of utility is irrelevant. For example, adding a constant to the utility of every alternative will not change the alternative with the highest utility, neither will it change the alternative chosen by the individual and, from the researcher's perspective, neither will it change the choice probabilities. As a consequence, only parameters 'that capture differences across alternatives' can be estimated (Train, 2009). This has important implications when socio-economic variables are included in a model. Socio-economic variables, such as income or car ownership, are constant for all alternatives in a decision maker's choice set, as they are a characteristic of the individual and not the alternative. Adding a socio-economic variable to the utility function of all the alternatives would simply add a constant to each alternative and would not create a difference in utility between them. This problem can be handled either by excluding the variable from the utility function of one of the alternatives, or by interacting the socio-economic variable with a variable that does differ between alternatives. For example, the fare for a train journey (a variable that differs between alternatives) could be divided by the income of the decision maker (a variable that is constant between alternatives) (Train, 2009).

The variable interaction approach was adopted by Pang and Khani (2018), who interacted (multiplied) both access time and frequency with income. Negative parameters were estimated for both interacted variables, indicating that as income increases the negative utility associated with access time increases and the positive effect of frequency on utility is reduced. According to Pang and Khani this shows that those on higher incomes are 'more motivated' by shorter access distances and 'less motivated' by higher service frequency. However, there are alternative potential explanations, other than the behavioural ones suggested. For example, perhaps property prices are higher closer to park and ride lots (the focus of this study) and so higher income travellers live on average closer to them (and thus chosen lots will have shorter access journeys).

Adcock (1997) notes that research shows that a passenger's propensity to railhead increases as the number of cars per household increases, with the effect most marked in moving from one to two car households. This may be because in a two-car household there is still a car available to use while the other is parked at a station all day, or may reflect increased affluence. This suggests that car ownership may influence station choice. Debrezion et al. (2009) estimated a parameter for car ownership for each access mode (excluded from the

walk mode utility function for reasons discussed above) in a nested logit model with access mode at the upper level. The parameter was positive for car or bicycle as access mode, but with P-values of 0.483 and 0.720 respectively, these were not statistically significant findings. An increase in car ownership would intuitively be expected to result in an increased probability of using the car as access mode, and it might be that the use of an aggregate measure of car ownership (cars per head for each postcode area) has resulted in a Type II error. A significant negative parameter was estimated for the public transport access mode, indicating that public transport is less likely to be used to access a station as car ownership increases. A negative effect of car ownership on public transport (transit) use was also found by Chakour and Eluru (2014).

A summary of the socio-economic factors used in prior station choice research is provided in Table 3.8.

Author(s)	Factors considered
Liou and Talvitie (1974)	Ratio of total cost to median income
Harata and Ohta (1986)	Student (dummy)
Fan et al. (1993)	Age (car and transit modes), sex (car mode), annual income > $50,000 (car mode)
Fox (2005)	Car driver (male, 16-19, 20-24, one car), car passenger (male, 35-44, zero cars, one car), rail-only pass
Debrezion et al. (2009)	Car ownership (per head for postcode area)
Chakour and Eluru (2014)	25 years old and younger, male, car ownership, reside in zone with high vehicle ownership and high percentage of larger vehicles
Pang and Khani (2018)	Age, income, sex, number of vehicles, household size, race, years living in Austin

TABLE 3.8: Summary of socio-economic related factors used to construct utility functions in station choice models.

3.5.5 Alternative-specific constants

An alternative specific constant (ASC) for each alternative can be included in its utility function to capture 'the average effect on utility of all factors not included in the model' (Train, 2009). Its role is analogous to the constant in a linear regression model. As only differences in utility matter (as discussed in Section 3.5.4 above), it is necessary to normalise the constants, which is usually achieved by normalising one of them to zero (i.e. excluding an ASC from the utility function of one of the alternatives). The other ASCs are then interpreted relative to the excluded one (Train, 2009).

In prior station choice research, ASCs are not always included in the utility functions. Blainey and Evens (2011) found that incorporating ASCs resulted in a better fitting MNL model, and in a NL model with access mode at the upper level and station choice at the lower level, Givoni and Rietveld (2014) included *only* ASCs in the access mode utility functions. In a NL model with station choice at the upper level, which collapsed to the MNL form, Wardman and Whelan (1999) interacted the ASCs for each access mode with access distance (ASC

× distance), so that the model could account for the affect of access distance on choice of access mode. The estimated parameters for the interacted ASCs, relative to the reference access mode (car), were negative for walk and cycle (−0.08) and negative, but less so, for bus (−0.016), as would be expected intuitively.

3.6 Station choice models in station demand forecasting

While from a transport planning point of view it might be expected that a key aim of station choice modelling would be to predict the impact of changes to station and service provision, few of the studies discussed in this chapter have addressed this issue, instead focussing on developing models to better understand the factors that influence station choice. There are several examples of local applications, for example Harata and Ohta (1986) used their model to estimate aggregate passenger flows at their study station, and Kastrenakes (1988) examined the effect of introducing a hypothetical commuter line by using predicted station shares to weight variables in a mode choice model. However, there has been limited progress toward integrating a station choice element into the aggregate models, such as trip end and flow models, that are typically used to predict demand for new stations or services (as discussed in Chapter 2).

Wardman and Whelan (1999) attempted to define station catchments based on their station and access mode choice model. However, they excluded the access mode choice element due to the amount of time required to derive the access mode variables for each zone to each competing station, and instead used a simpler distance to station measure in the model. To define a station's catchment they used the model to apportion the population of each zone (postal sector) to one of five competing stations allocated to that zone, before entering the data into a direct demand summation model. However, due to time and computer resource constraints, they had to use a subset of the data and this resulted in the summation model failing to converge. This approach does not appear to have been revisited since, despite the substantial advances in computational capability.

Lythgoe and Wardman (2002, 2004) proposed an alternative approach for incorporating a station choice element into a direct demand model, specifically to forecast demand for new parkway stations. The dependent variable in the model is the number of journeys between a parkway station and destination stations, obtained from ticket sales data, and there are no observed choice probabilities. The theoretical approach is described below, but Lythgoe and Wardman (2002) and Lythgoe and Wardman (2004) should be consulted for greater detail and full model derivations.

A parkway station's generation potential is represented by the population within 40 km of the station (obtained from the 1991 census data), which is allocated to a grid of 16 polygonal zones. The population of each zone is allocated to a point that represents the zonal centre of

population. The following direct demand (flow) model can now be formulated:

$$V_{aij} = n \times P_a \times Pr(rail_{aij}), \tag{3.41}$$

where V_{aij} is the number of trips from zone a to destination j via parkway station i, n is the unknown average number of decisions to travel (by any mode or not to travel) in one year, and P_a is the population of zone a. $Pr(rail_{aij})$ is the probability of an individual in zone a choosing to travel to destination j using parkway station i, which is obtained from a NL model with a choice between rail and no rail at the upper level and choice of station at the lower level (See Figure 3.4). The choice set available to each parkway station zone is composed of the parkway station itself and 10 other (non-parkway) stations that are within 40 km of at least one zonal centre of population and considered to be the most competitive[17]. $Pr(rail_{aij})$ is the product of a conditional and marginal probability:

$$Pr(rail_{aij}) = Pr(rail_{aij}|rail_{aj}) \times Pr(rail_{aj}), \tag{3.42}$$

where $Pr(rail_{aij}|rail_{aj})$ is the conditional probability of using station i to get to j given that a choice to use rail to get from a to j has been made; and $Pr(rail_{aj})$ is the marginal probability of using rail to get to j (rather than another mode of transport or not travelling at all). V_{aij} is unknown as there is no data on trips at zone level. However, the total number of journeys from parkway station i to destination j is known from ticketing data. Thus, using a summation model:

$$\begin{aligned} V_{ij} &= \sum_a n \times P_a \times Pr(rail_{aij}) \\ V_{ij} &= \sum_a V_{aij}. \end{aligned} \tag{3.43}$$

The model is estimated after a log transformation, and based on certain assumptions with regard to n, using non-linear least squares regression. A limitation of the model identified by the authors is that while the proportion of journeys abstracted from each competing station can vary dependent upon relative utility values, the ratio of journeys abstracted from competing stations to new journeys generated at the parkway station is constant (0.5 in their model)[18]. Later work by the authors enhanced the approach by developing a form of CNL model that allows the ratio to vary depending upon the proximity of a station to its competitor stations (see Section 3.3.3.1 for a detailed discussion).

In other work, Blainey and Evens (2011) developed a method that utilised a station choice model to forecast demand abstraction by new stations, and tested this by forecasting the

[17] Each parkway station is assumed to have a 40 km catchment divided into zones. A competing station must lie within 40 km of at least one zonal centroid, thus this model is only suitable for journeys that are greater than 80 km. Competing stations were ranked based on population and revenue weighted by distance for each origin station, with the top 10 available as choices in the NL model.

[18] For example, in the reported case studies, the East Midlands Parkway model predicts 880 abstracted journeys which is 0.5 of predicted new journeys; and in the Warwick Parkway case study, the model predicts 41,651 abstracted journeys from competing stations, which is 0.49 of predicted new journeys.

probability of a passenger in a particular zone (census output area) using Aber station in South Wales before and after the opening of a proposed new station at Energlyn. They also outlined a potential method for converting these probabilities into the number of trips abstracted from Aber by using the station choice model to allocate estimated annual trips for each output area to the two stations, but did not propose a method to account for new trips that might be generated by the new station. There are also a few examples of a limited station choice element being introduced into regional strategic (four-stage type) models. Fox (2005) developed a park-and-ride station choice model, where station choice is modelled for car access mode only, that was incorporated into the Policy Responsive Integrated Strategy Model (PRISM), a disaggregate demand model for the West Midlands region of the UK. A similar model was later developed for the Sydney Strategic Travel Model (Fox et al., 2011).

3.7 Conclusions

Following a brief history of station choice modelling research in Section 3.2, Section 3.3 examined the theoretical underpinnings of both closed-form and simulation-based discrete choice models, alongside a critical review of their application in research published over the past 40 years. The vast majority of previous studies have adopted either MNL or NL models, although recent work has begun to consider more complex approaches, such as the ML model. However, there has been little recognition in this body of work that railway stations are located in space, and that the use of standard choice models that are a-spatial in nature might not be appropriate. It is a reasonable assumption that stations that are closer to each other in space will be better substitutes for one another than stations that are further apart, following Tobler's (1970) first law of geography, that 'everything is related to everything else, but near things are more related than distant things'. It is therefore concerning that the NL model, which is intended to address inappropriate patterns of substitution that occur in MNL models, has been implemented in station choice models by including every station in the choice set of every nest. This fails to address the substitution issue. If station choice models can be developed that have more realistic substitution behaviour, they may be more accurate and have greater transferability. While recent research is beginning to tackle this issue, solutions applied in other fields, such as including an accessibility term or applying a specialist GEV model, should also be explored in the station choice context. However, before developing ever more complex explanatory models it is important that the predictive performance of the simpler approaches is more rigorously assessed using measures consistent with probabilistic choice models; and when more complex models are developed, it is essential that their predictive performance is compared with simpler models so that an informed assessment of the trade-off between complexity and performance can be made.

In Section 3.5 attention turned to the attributes that can help explain station choice behaviour. It is clear that the direction effect of a range of factors related to accessibility and services has been consistently reported across many studies. The evidence indicates that station

utility decreases as the access journey becomes further or longer, as the rail leg journey time increases, when the journey involves more transfers or has a higher fare, and when service frequency is reduced. Establishing the effect of station facilities, such as car parking, is more problematic, potentially due to endogeneity issues. While a number of important explanatory variables have been identified, there is still potential to identify new ones. For example, researchers have paid scant attention to the impact of land-use factors, and spatially detailed land-use datasets, such as the Ordnance Survey's 'Points of Interest' data in Britain, are an untapped resource. There may also be gains in predictive performance with improved measurement of the variables that have the greatest explanatory power, such as access journey and rail service factors. This could be through better measurement of access journey time by mode using route planning software, incorporating road speed information that could identify congestion prone stations, or better alignment of survey trip data with train schedule information so that the service available to each individual is better represented.

Issues surrounding the data used in station choice modelling was the subject of Section 3.4. While aggregate data has in some cases been used to model station choice, this has been dictated by limitations of available data, rather than modelling needs. The preferred option for future research is individual trip data where the ultimate origin (and destination if required) is at a spatial resolution sufficient for the variability in explanatory factors between decision makers to be revealed. In the UK, the unit postcode area boundary is probably the maximum spatial unit of address aggregation appropriate. A definitive mechanism for defining choice sets has not been established, and the methods adopted have been fairly simplistic and not evidence-based. It is not clear, for example, what the implications are of seeking to maximise the number of observed choices that are accounted for, when this may add alternatives to the choice set that would never realistically be considered. Research is needed to evaluate the different methods for generating choice sets for station choice models, including an assessment of their impact on predictive performance.

The lack of integration with demand forecasting is a significant limitation of previous station choice research that was highlighted in Section 3.6. This is important as the absence of a choice-modelling methodology which can adequately capture patterns of abstraction and competition between railway stations may have contributed to the limited accuracy of recent demand forecasts for new stations in the UK (as discussed in Section 2.5). There has been a very limited amount of work to explore the incorporation of probability-based catchments into the traditional aggregate models, with the majority of previous studies narrowly focussed on identifying the explanatory factors affecting station choice. The models proposed by Lythgoe and Wardman (2004) and Lythgoe et al. (2004) are only suitable for forecasting journeys of 40km or greater, and there has been limited work testing the transferability of these models, an issue shared with much of the published research. Wardman and Whelan (1999) were intending to incorporate probabilistic catchments into their direct demand summation model, but faced issues caused by limited computer processing capability. Given the substantial increase in computing power that has occurred since 1999, their general

approach to defining probabilistic station catchments will be revisited as part of the research presented in this thesis.

Chapter 2 established that the trip end model is the most commonly used model in the UK to assess proposals for new local railway stations, and that the discrete and deterministic station catchments defined in this type of model do not represent the complex reality of station catchments revealed by empirical evidence. An alternative approach was then proposed that would use models of station choice to define probability-based catchments. This chapter has shown that while there is a substantial body of research related to station choice, this has overwhelmingly focused on developing explanatory models relevant to specific local contexts. There has been far less attention given to how these models perform in a predictive capacity and how they might be used to improve the models, such as the trip end model, that forecast passenger demand for new stations. As set out in the Introduction to this thesis, the aim of this research is to develop a transferable station choice model that has the potential to adequately predict station choice in most local situations in the UK, and to use that model to incorporate probabilistic catchments into aggregate models of rail demand.

The next two chapters will describe the work carried out to obtain and prepare the data necessary to build and estimate the station choice models. Chapter 4 is concerned with the data that reveals actual station choices made by rail passengers, and Chapter 5 is concerned with the data that can help explain those choices.

Chapter 4

Observed station choice data

4.1 Introduction

This chapter is concerned with the survey data that reveals observed station choice. This is a key requirement for the calibration of discrete choice models, as it will represent the chosen alternative in the dependent (choice indicator) variable.

The first part of the chapter deals with obtaining, preparing and validating suitable data. Section 4.2 considers the type and nature of the data required and the sources that were selected for this study; Section 4.3 describes procedures that were developed to maximise the usefulness of the data by matching incomplete textual addresses to unit-level postcodes and estimating the coordinates of origins/destinations known to be located on a specific street; Section 4.4 considers how the data was checked and cleaned, and provides a detailed breakdown of the adjustments that were made; and Section 4.5 explains the automated process that was developed to ensure, as far as reasonably practicable, the validity of the reported trips.

The second part of the chapter, Section 4.6, provides a descriptive analysis of the cleaned and validated datasets, considering the nature of the access and egress journeys, with a particular focus on transport mode used (Section 4.6.1); exploring the extent to which passengers choose their nearest station (Section 4.6.2); and finally an analysis and visualisation of observed station catchments revealed by the data (Section 4.6.3). The findings of the descriptive analysis are discussed with reference to the methods adopted to define station catchments in the aggregate demand models typically used to forecast demand for new stations (as covered in Section 2.4.1). The chapter closes with a summary of the information presented and its implications (Section 4.7).

4.2 Passenger survey data

In order to develop disaggregate models of station choice, information is required about individual trips via the rail network. To model choice of access station, this data needs to include, at a minimum, the ultimate origin of the trip, such as home or work address, and the station where a train was first boarded. If the models need to account for variables related to the train leg, such as travel time or number of transfers, then the final egress station is also needed, and possibly the ultimate destination of the trip. The data must also be at a spatial resolution that is sufficient for the variability in explanatory factors between individual decision makers, such as access distance, to be revealed. For UK-based research, the unit postcode area boundary is probably the maximum spatial unit of address aggregation appropriate for this type of analysis. The unit postcode is the most detailed spatial unit available from postcode data in the UK, and for small postal users (i.e. not business addresses) it typically represents around 15 addresses, although it is possible for it to contain up to 100 addresses in densely populated areas.

4.2.1 Data sources

Data from a series of on-train passenger surveys were obtained from the Welsh Government (WG) and Transport Scotland's Land-Use and Transport Integration in Scotland (LATIS) service. These two datasets were chosen to increase the robustness of the models by maximising the number of observed choice data points; to allow the predictive performance of models calibrated using data from different regions to be compared; and to enable model transferability to be tested. The Welsh surveys were conducted in Spring 2015 and primarily covered stations in South East Wales (Cardiff, Newport and the South Wales valleys) and Swansea. The LATIS surveys were carried out in 2013 (a small survey), 2014 and 2015. While covering stations throughout Scotland, they were concentrated in the Central Belt. In both cases the survey questionnaires focused on the 'current train', asking for the boarding/alighting station and the access/egress mode, along with questions about the ultimate trip origin/destination and reasons for travelling. There were some supplementary socio-demographic questions, including sex, age (WG only), and household car ownership (LATIS) or car availability (WG). Prior to subsequent processing and validation the WG and LATIS surveys contained some 7,000 and 52,000 responses respectively, and were supplied in Microsoft Excel spreadsheet format.

4.3 Address matching and estimation — LATIS data

The WG data had been through some data processing before it was supplied, and nearly all observations included valid origin and destination unit-level postcodes. This was not the

case for the LATIS data, where addresses had not been validated and many observations had missing, incorrect or incomplete postcodes. For example, less than 50% of the origin addresses included a valid unit-level postcode. Survey respondents are likely to know the origin or destination postcode for some types of trip, such as those beginning or ending at their home address, but not for others. In order to ensure that the dataset used in model calibration was representative of a broader range of trip types, a procedure was developed to match the incomplete address information to postcodes using the Ordnance Survey (OS) AddressBase product which contains over 28 million UK addresses from Royal Mail's postal address file (PAF). The aim of this procedure was to either identify a specific postcode from the provided address information or, failing that, to approximate the geographic location of an address.

4.3.1 Survey data preparation

In order to conform with a privacy impact assessment agreed with Transport Scotland, a procedure was adopted to ensure that at no point during address matching and data analysis would the working dataset contain both detailed address information and other survey response information. The survey data was saved to an encrypted partition on physical media and the individual spreadsheets were merged to create a single CSV file for each year (from now on referred to as the 'complete' CSV files). A unique ID was then assigned to each entry in each of the complete CSV files. The origin and destination address fields only, along with the unique ID, were then extracted to a separate 'address matching' CSV file for each year. These files were used in the address matching process. Once the addresses had been matched to postcodes or locations estimated, the other address fields were removed from these files. All fields other than the address fields were then extracted from the complete CSV files, and the matched postcodes or coordinates of estimated locations were merged from the address matching CSV files based on the unique ID field, creating new 'working' CSV files that were used in subsequent data cleaning and validation. This procedure is illustrated in Figure 4.1.

4.3.2 AddressBase database preparation

Using an R script, the 29 individual CSV files that formed the supplied AddressBase dataset (dated 7 April 2016) were read into R and then appended to a PostgreSQL database table. The resultant AddressBase table consisted of some 28 million rows, and 15 relevant fields which were retained. The fields are listed in Table 4.1 along with an additional explanation of their purpose where this is not clear from the field name (information on the AddressBase structure was obtained from the technical specification document (Ordnance Survey, 2015)).

Several new fields, required for the address matching process, were then created using SQL queries:

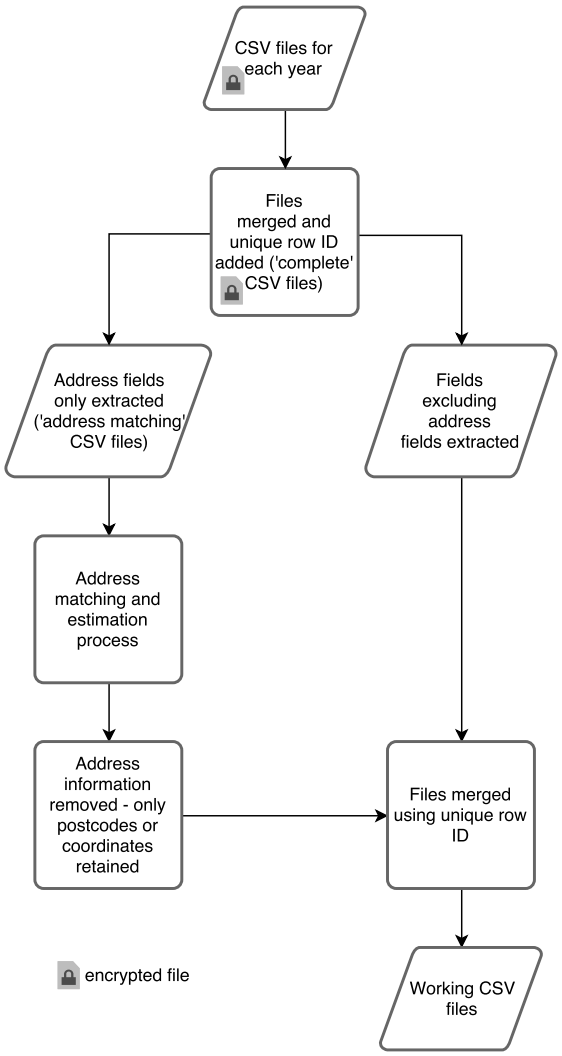


FIGURE 4.1: Procedure followed to ensure compliance with the privacy impact assessment during the address matching process.

- `full_text_address` and `address_short` — these were formed by concatenating certain existing AddressBase fields, as indicated in Table 4.1.
- `postcode_count` — the number of unique postcodes for each distinct combination of `POST_TOWN`, `DEPENDENT_THOROUGHFARE`, `THOROUGHFARE`, `DEPENDENT_LOCALITY`, and `DOUBLE_DEPENDENT_LOCALITY` (from now on referred to as a 'unique thoroughfare'). The dependent thoroughfare and locality fields were included to limit the problem of duplicate thoroughfare names within the same postal town. This may not deal with identically named thoroughfares located within different postal districts within the same city. In these cases the addresses are normally distinguished by the very fact they exist in different postcode districts rather than by using dependent locality or thoroughfare fields. For example, there are many instances of 'College Road, London', that would be treated as the same street using this method of obtaining the postcode count. On the very few occasions that this issue impacted streets within the survey data it was detected during the address matching process, as these streets display an

Field name	Explanation	Full text address	Short address
UDPRN	Royal Mail's unique delivery point reference number	n	n
ORGANISATION_NAME		y	n
DEPARTMENT_NAME		y	n
PO_BOX_NUMBER		n	n
SUB_BUILDING_NAME	Property subdivision, e.g. Flat 10	y	y
BUILDING_NAME		y	y
BUILDING_NUMBER		y	y
DEPENDENT_THOROUGHFARE	Name of an adjoining road used to distinguish thoroughfares with the same name in a postal town	y	y
THOROUGHFARE	A road with delivery points	y	y
POST_TOWN		y	y
DEPENDENT_LOCALITY	Used to distinguish thoroughfares with the same name in a postal town, where no dependent thoroughfare	y	y
DOUBLE_DEPENDENT_LOCALITY	Used to distinguish thoroughfares with the same name and in the same locality within a postal town	y	y
POSTCODE		n	n
POSTCODE_TYPE	Idenitifes if the postcode belongs to a large or small user (as defined by Royal Mail)	n	n
COUNTRY		n	n

TABLE 4.1: AddressBase fields retained in the PostgreSQL table, with explanation of their purpose (where not obvious). Also shows which fields were concatenated to create the 'full text address' and 'short address' fields.

unusually large maximum distance to the centroid of street postcodes (see `max_d_2ct`, below). The query to generate this field is shown in PostgreSQL code segment B.1.1 in Appendix B.

- `stpc_cent_geom` — this is the centroid of all the individual postcode centroids belonging to each unique thoroughfare (from now on referred to as the 'calculated centroid'). The SQL query first identifies the set of postcodes for each unique thoroughfare, as for the `postcode_count` field above, then 'collects' together the point geometries for the centroids of these postcodes from a database table containing the ONS Postcode Directory (ONSPD), and finally calculates the centroid of those centroids. The query to generate this field is shown in PostgreSQL code segment B.1.2 in Appendix B.

- `max_d_2ct` — this is the maximum Euclidean distance from any of the individual postcode centroids belonging to each unique thoroughfare to the calculated centroid for that thoroughfare. The query collects together the point geometries for the centroids of the set of unit postcodes for each unique thoroughfare, and then finds the maximum Euclidean distance from any of these to the calculated centroid. The query to generate this field is shown in PostgreSQL code segment B.1.3 in Appendix B.

The relationship between the `postcode_count`, `stpc_cent_geom` and `max_d_2ct` fields is illustrated in Figure 4.2. If the calculated centroid is used to represent the location of an

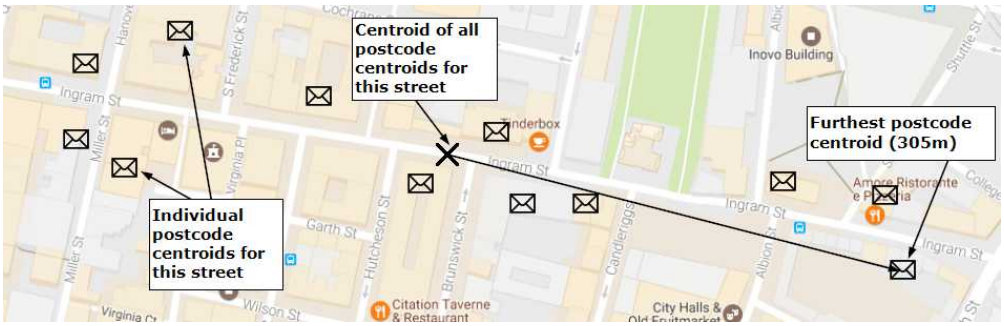


FIGURE 4.2: Postcode centroids for Ingram Street, Glasgow, showing calculated centroid and maximum distance from calculated centroid to any postcode centroid.

origin or destination on a street, the maximum Euclidean distance indicates how far the 'real' address postcode centroid could be from that location.

4.3.3 Address matching process

The matching process was performed separately for each survey year and separately for origin and destination addresses. The address matching CSV files for each year contained a postcode field and three address fields for both the origin and the destination. The following initial steps were carried out:

1. The CSV file for each year was imported into an R data frame.
2. The postcode field was checked against the ONSPD database table. Records with a valid postcode were identified and filtered from the data frame.
3. Any records where the three address fields were empty were filtered from the data frame.
4. For the remaining records an attempt was made to identify the postal town of the provided address, by looking for an *exact* match within a list of distinct postal towns obtained from the AddressBase table. Each of the three address fields was checked in turn, with the *last* match recorded.
5. For those records where the postal town could not be matched, an attempt was made to match the postal sector identifier in the postcode field[1] to a list of unique sector postcodes and their respective postal towns pulled from AddressBase. This was achieved by stripping off everything including and after the first space in any string in the postcode field. This approach is not 100% accurate, as it is possible for some sector identifiers to cover more than one postal town (for example, KY11).
6. A full address field was generated by concatenating the three address fields.

[1]These will be incomplete or invalid postcodes which were not filtered out in step 2.

7. Common abbreviations were amended or expanded to match the format used in AddressBase. For example, 'ST ANDREWS' was amended to 'ST. ANDREWS' and ' GDNS ' was replaced with ' GARDENS '.

8. Any full address which exactly matched a postal town (i.e. consisted only of a postal town), was filtered from the data frame.

Due to the size of the AddressBase table it was not practical to perform a search of all addresses for each record to be checked, as tests confirmed that each query could take several minutes. Performing some 26,000 queries, at two minutes each, would have taken nearly 900 hours, or over a month, to complete. An alternative approach was therefore adopted which limited the scope of the search. In the first instance the search was restricted to the identified postal town of the record (obtained using the steps outlined above), and then for any remaining unmatched addresses the search was restricted to addresses within Scotland.

For the postal town restricted search, for each unique postal town present in the data frame, the following steps were performed (see R code segment A.4 in Appendix A):

1. A temporary database table of records from AddressBase with that postal town was created.

2. A GIN index, a feature of the PostgreSQL `pg_trgm` module, was created for both the full text and short address fields. These indexes allow fast similarity searches to be performed.

3. For each record in the data frame with this postal town, a similarity search query was performed using the `pg_trgm` module. This uses a trigram algorithm where the number of trigrams (groups of three consecutive characters) that two strings share are counted and a similarity index is returned that can range between 0 (strings completely dissimilar) and 1 (strings match exactly) (The PostgreSQL Global Development Group, 2017). Using a union query, the top four results from two similarity searches on the full and short address fields, ordered by the similarity index (descending), were retrieved. Each of the results was written to the data frame along with other required database fields[2].

For all records where the postal town was unknown, or where an address was not found using the similarity search described above[3], the process was repeated but instead of creating

[2]AddressBase addresses may have organisation and department name fields populated, but this level of detail may not have been provided by the survey respondents. This would have an impact on the accuracy of the matching process. For example, if the respondent had provided: '180 Vincent Street Edinburgh', but in AddressBase this address was recorded as: 'Company Name 180 Vincent Street Edinburgh' it may not be returned in the top matches. The short address field attempts to deal with this issue by excluding the company name and department fields. This approach will find the top matches based on two queries that use different versions of the AddressBase address field.

[3]The minimum similarity index was set at its default value of 0.3. Any address with a similarity index less than this would not be returned.

	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Origin.Post	Origin	Origin.posttown	Origin_full_address	pt_check	status	chk	M1.s	M1.add	M1.pc	M1.pcnt	M1.maxd2ct	M1.stpcc
2	EH12	N	EDINBURGH	67 SOUTH GYLE EDINBURGH	N	n		0.83	67 SOUTH GYLE PARK EDINBURGH	EH12 9EW	2	23	55.93433,-3.30143
3		N	EDINBURGH	WAVERLEY STATION EDINBURGH	N	y		0.77	WAVERLEY RAILWAY STATION EDINBURGH	EH1 1BB	NA	NA	NA,NA
4	EH3	N	EDINBURGH	CORNWALLIS PLACE EDINBURGH	N	y		0.93	6 CORNWALLIS PLACE EDINBURGH	EH3 6NG	1	0	55.96108,-3.19477
5	EH9	N	EDINBURGH	9 MONCREIFF TERRACE EDINBURGH	N	y		0.76	9/9 MONCRIEFF TERRACE EDINBURGH	EH9 1NB	3	24	55.93907,-3.18391
6	G81	N	CLYDEBANK	DALMUIR	N	n		0.31	9C DALMUIR COURT CLYDEBANK	G81 4AB	2	0	55.91034,-4.42674
7		N	unknown	WALDORT ASTONA EH1	N	EH1 2AB		NA	NA	NA	NA	NA	NA
8		N	unknown	SOUTH GYLE	N	n		0.39	5 SOUTH GYLE LOAN EDINBURGH	EH12 9EN	1	0	55.935,-3.30119
9	KY4	N	COWDENBEATH	33 NETHERTON ROAD COWDENBEATH	N	y		1	33 NETHERTON ROAD COWDENBEATH	KY4 9BF	2	154	56.11924,-3.35454
10		N	EDINBURGH	EBLENHEIM PLACE EDINBURGH	N	y		0.79	6 BLENHEIM PLACE EDINBURGH	EH7 5JH	1	0	55.95753,-3.18364
11	EH3	N	EDINBURGH	CANONGATE EDINBURGH	N	n		0.91	3 CANONGATE EDINBURGH	EH8 8BX	17	313	55.95153,-3.17962
12	HU8	N	HULL	24 LABURNUM AVENUE HULL	N	n		0.63	2 THE GROVE LABURNUM AVENUE HULL	HU8 8PQ	1	0	53.76416,-0.31599
13	EH14	N	EDINBURGH	NHS ROBBS LOAN EDINBURGH	N	y		0.64	NHS LOTHIAN 47 ROBB'S LOAN EDINBURGH	EH14 1AB	7	155	55.93099,-3.2473
14		N	unknown	HAYMARKET	N	n		0.5	HAYMARKET EDINBURGH	EH12 5EY	1	0	55.94564,-3.2183
15		N	DUNFERMLINE	JAMES STREET DUNFERMLINE	N	p		0.89	19 JAMES STREET DUNFERMLINE	KY12 7QE	2	63	56.07203,-3.45626
16		N	EDINBURGH	GORGIE ROAD EDINBURGH	N	n		0.92	1 GORGIE ROAD EDINBURGH	EH11 2LA	40	1175	55.93562,-3.23978
17		N	EDINBURGH	NATIONWIDE GEORGE STREET EDINBURGH	N	y		0.76	NATIONWIDE BLDG SOC 71 GEORGE STREET EDINBURGH	EH2 3EE	26	393	55.95306,-3.19998
18	EH7	N	EDINBURGH	EAST EDINBURGH	N	n		0.71	EDINBURGH	EH12 1EF	NA	NA	NA,NA
19	EH12	N	EDINBURGH	HUDSON HOUSE ALBANY STREET EDINBURGH	N	y		0.95	HUDSON HOUSE 8 ALBANY STREET EDINBURGH	EH1 3QB	6	113	55.95724,-3.19118
20	KY70GZ	N	KIRKCALDY	24 SCOTT AVENUE KIRKCALDY	N	n		0.65	24 HAIG AVENUE KIRKCALDY	KY1 2JE	2	67	56.12835,-3.14505
21	EH8	N	EDINBURGH	GEORGE SQUARE EDINBURGH	N	p		0.92	7 GEORGE SQUARE EDINBURGH	EH8 9JZ	6	127	55.94346,-3.18845
22		N	unknown	UNIVERSITY OF EDINBURGH BUCCLEUCH STREET	N	y		0.68	UNIVERSITY OF EDINBURGH 15 BUCCLEUCH PLACE EDINI	EH8 9LN	3	86	55.94287,-3.18603
23		N	EDINBURGH	SCOTTISH GOVERNMENT REGENT ROAD EDINBURGH	N	EH1 3DG		0.63	SCOTTISH GOVERNMENT 47 ROBB'S LOAN EDINBURGH	EH14 1TY	7	155	55.93099,-3.2473
24		N	unknown	HOLROYD PARK QUEENS DRIVE EDIMBURGH	N	y		0.4	HOLYROOD PARK EDUCATION CENTRE 1 QUEEN'S DRIVE	EH8 8HG	1	0	55.95369,-3.16769
25	EH6	N	EDINBURGH	2F1/1 CONNELLY BANK ROAD EDINBURGH	N	y		0.64	2F1 91 COMELY BANK ROAD EDINBURGH	EH4 1BJ	11	362	55.95896,-3.22064
26	EH1	N	EDINBURGH	PRINCES STREET EDINBURGH	N	n		1	PRINCES STREET EDINBURGH	EH2 2EJ	23	654	55.95221,-3.19678
27		N	EDINBURGH	WGH CREWE ROAD EDINBURGH	N	p		0.73	2/2 CREWE ROAD WEST EDINBURGH	EH5 2PB	7	228	55.97455,-3.2396

FIGURE 4.3: Extract from address matching spreadsheet used for manual review of addresses with the highest similarity index from AddressBase. This example only shows the top matching address for each row (column 'M1.add').

a temporary postal town table, a table of all addresses in Scotland was used for the similarity search. When both searches were complete, those records in the data frame with potential matching addresses were exported from R to a CSV file. The CSV file was then imported into Microsoft Excel and an automated colour-coding system was used to aid a manual review process, using the following key criteria:

1. Correctly matched postcode accepted where possible.
2. If street name matched but house number/business name not matched:
   (a) if street has a single postcode: postcode accepted.
   (b) if street has more than one postcode: if `max_d_2ct` is <= 250 m, use the co-ordinates of the calculated street centroid (`stpc_cent_geom`) as the origin or destination location.

An example section of the matching spreadsheet is shown in Figure 4.3. Some manual look-ups using Google search and/or Google Maps were carried out for common addresses that could not be matched but were unambiguous, for example where just the name of a shop or hotel was provided. After the manual check was complete, the spreadsheet was ordered by `max_d_2ct` (descending). As mentioned earlier, a limitation in the process of identifying the postcodes that relate to a specific street can result in postcodes from multiple streets within the same postal town that have the same name being grouped together. This is especially common in London, where roads with the same name would normally be differentiated for postal purposes by postcode district. When this occurs, it produces an excessively large `max_d_2ct` value. This effect proved useful in identifying a few instances where the street name provided by the survey respondent was unknowingly ambiguous (because it occurs more than once in a town), and in these cases the match was rejected.

Once the address matching was completed, post-processing was carried out in R before the matched postcodes and estimated location coordinates were merged into the original survey data (as described in Section 4.3.1 and illustrated in Figure 4.1.). Figures 4.4 and 4.5 summarise the address matching process for trip origins in 2014 and 2015 respectively, while Figures 4.6 and 4.7 summarise the address matching process for trip destinations in 2014 and 2015 respectively. It was not necessary to apply the address matching process to the small 2013 survey. In total, the address matching process resulted in a 31% increase in the number of validated trip origins, and a 58% increase in the number of validated destinations. The LATIS survey data that was taken forward to the data cleaning and validation stages consisted of those records with both a valid origin *and* destination — a total of 19,951 records.

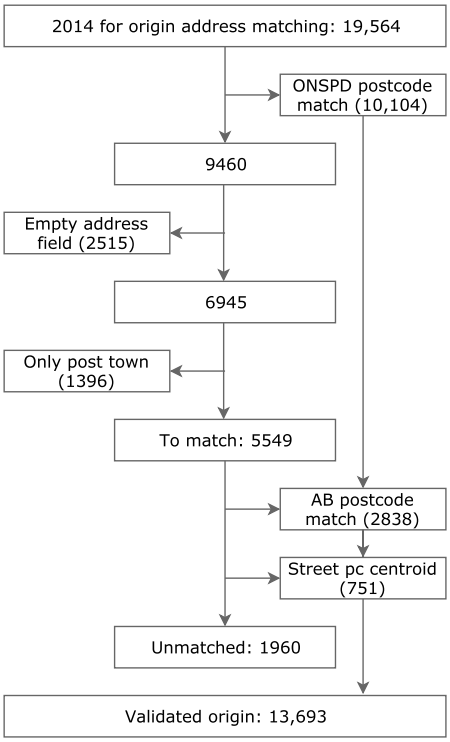


FIGURE 4.4: Address matching — LATIS 2014 Origins.

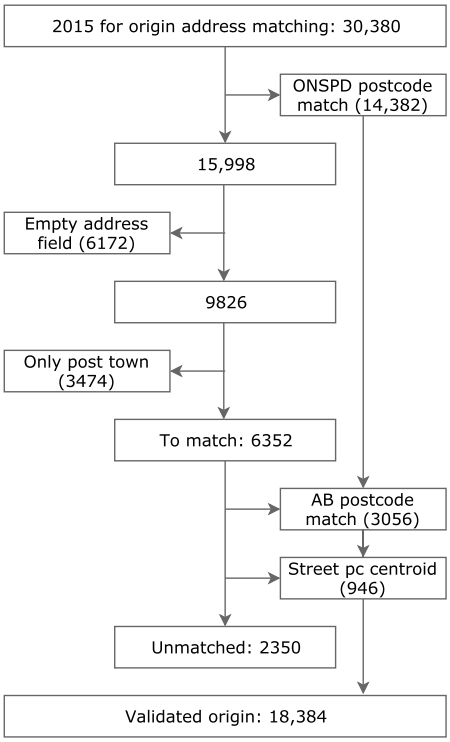


FIGURE 4.5: Address matching — LATIS 2015 Origins.

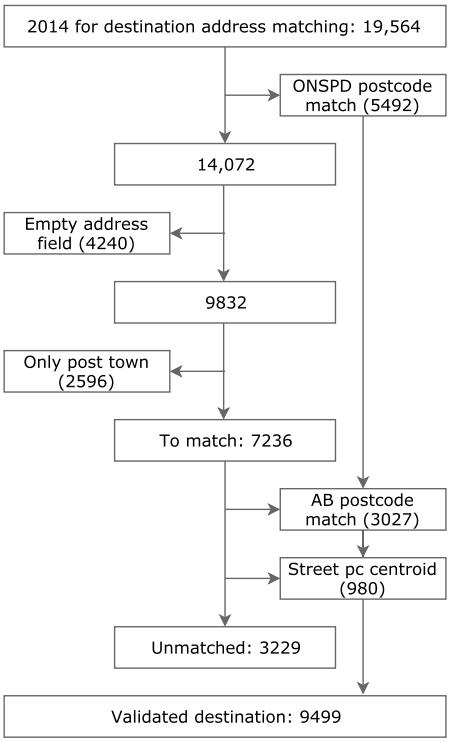


FIGURE 4.6: Address matching — LATIS 2014 Destinations.

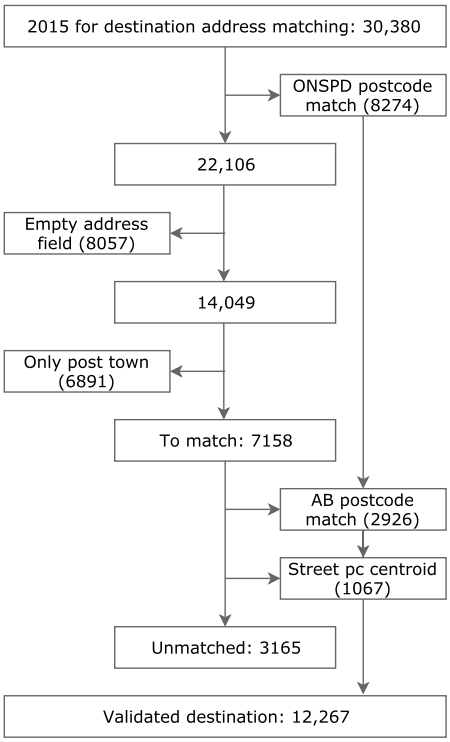


FIGURE 4.7: Address matching — LATIS 2015 Destinations.

4.4 Data cleaning

It was necessary to clean both the WG and LATIS survey data. This process was automated using R scripts. Data was checked and manipulated either using R data types and tools or, once the data had been written to a PostgreSQL database table, by running appropriate SQL queries from within R, using the RPostgreSQL package (Conway et al., 2016). This approach means that the entire cleaning process was transparent and reproducible and the data for model inputs can be generated again from the raw data.

4.4.1 WG data

The WG survey data had already been through some data processing prior to being supplied. Each observation had an apparently valid ultimate origin and destination unit-level postcode, and each spreadsheet contained separate sheets labelled: 'clean', 'illogical' and 'reversed'. However, when the 'illogical' and 'reversed' sheets were reviewed it was not always apparent why a trip was considered 'reversed' or 'illogical'. In addition, some trips on 'clean' sheets were found to be illogical. All the trips on 'illogical' and 'reversed' sheets were manually reviewed and where it was not obvious why they had been excluded, they were copied to the 'clean' sheet. All the 'clean' sheets were saved in CSV file format and then combined into a single CSV file which was imported into R.

A number of criteria were applied to check the supplied data, resulting in either amendments to, or the removal of, observations from the dataset. Origin and destination station names were matched against station names in the National Public Transport Access Nodes (NaPTAN) database. A list of unique station names that could not be automatically matched was manually reviewed, and where the intended station name was unambiguous the correct name was recorded in a look-up table which was then used to correct station names in the dataset. Those observations with origin or destination station names that could not be matched were removed. A number of observations were removed because the origin or destination postcode was not located on the mainland and it would not be possible to derive access and egress variables for these using the trip planner (discussed in Section 5.2)[4]. To limit the amount of public transit schedule data that needed to be incorporated into the trip planner, observations where the origin postcode was not in Wales were also removed. Any observations where the access or egress mode was given as 'another train' (respondents were asked for their access and egress mode in respect of the 'current train' they were travelling on when questioned) were removed, as it was not possible to determine the initial boarding and alighting station for these trips. The full range of adjustments made to the dataset during cleaning are detailed in Figure 4.8.

[4]The location of each postcode in respect to a range of administrative boundaries is recorded in the ONSPD.

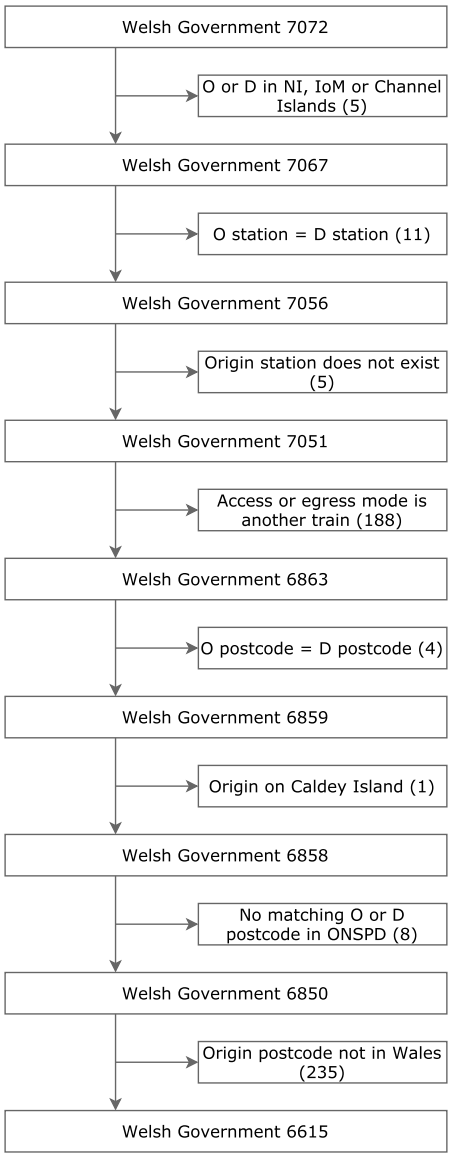


FIGURE 4.8: Adjustments made to the WG survey data during cleaning.

4.4.2 LATIS data

A number of criteria were applied to check the supplied data, resulting in either amendments to, or the removal of, observations from the dataset. Origin and destination station names were validated using the procedure described for the WG dataset. A variety of other data checks were carried out, including removing observations where the access or egress mode was not provided and where the origin station was the same as the destination station. In some cases multiple access or egress modes were recorded. This was a particular issue in the 2015 survey, as respondents were not asked to provide the 'main' mode used. When two modes were provided the following rules were followed to assign the main mode used:

- Where the two modes were motorised and non-motorised, the motorised mode was assumed to be the main mode.

- Where the two modes were walk and cycle, cycle was assumed to be the main mode.
- Where the two modes included 'other' and a non-motorised mode, 'other' was assumed to be the main mode.

Any remaining observations with multiple access or egress modes were removed from the dataset. To limit the amount of public transit schedule data that needed to be incorporated into the trip planner, only those observations where the origin was located in Scotland were retained. In addition, any observations with origins or destinations located on islands without road access to the mainland were removed, as it would not be possible to generate access and egress variables for these using the trip planner. The full range of adjustments to the LATIS dataset during cleaning are detailed in Figure 4.9.

4.5 Automated trip validation

Due to the large number of survey observations in the WG and LATIS datasets it was not practical to manually check each one to ensure the reported trip was sensible. An alternative strategy was adopted that generated information inherent in the reported trip and used that to automatically validate the trip. This approach was used to identify excessively long station access and egress legs, and unrealistic trips, as detailed below.

4.5.1 Excessive access or egress legs

4.5.1.1 Walk time

For each observation in the cleaned WG and LATIS datasets, a trip planner (see Sections 5.2 and 5.3) was used to obtain the walk time in minutes from the ultimate trip origin to the origin (boarding) station; and the walk time in minutes from the destination (alighting) station to the ultimate destination. A histogram and kernel density plot was then produced for access time (Figure 4.10) and egress time (Figure 4.11) using 5-minute bins. Based on the observed distribution, any observation with walk-mode access and/or egress time in excess of 60 minutes was removed from both datasets. This cut-off point felt intuitively appropriate, in addition to being supported by the data. Access or egress times in excess of 60 minutes will largely be due to errors in the survey data. It is possible that some long journeys are genuine, for example if a passenger travels by train to begin a day's walk via the public footpath network to their destination. However, the models being developed in this project will not be able to predict station choice for this type of trip and their exclusion will have a positive rather than negative effect on model performance.

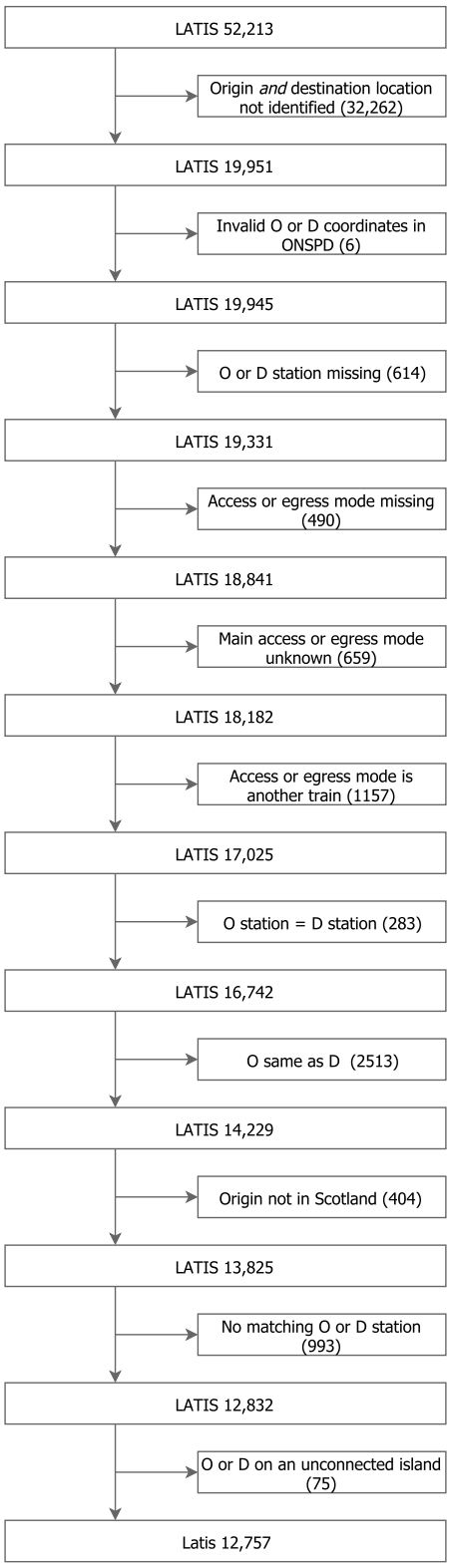
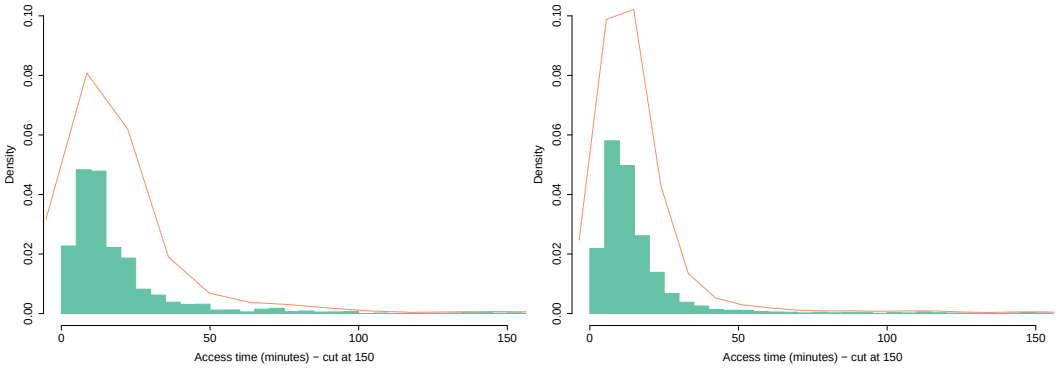
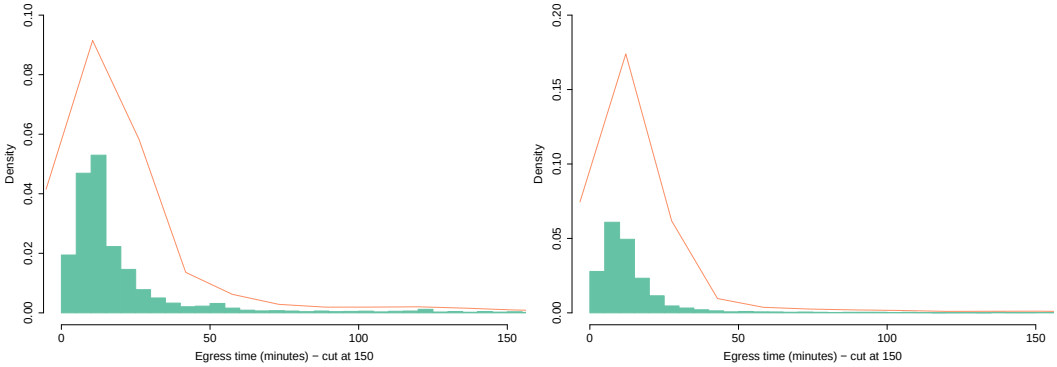


FIGURE 4.9: Adjustments made to the LATIS survey data during cleaning.



(A) WG (B) LATIS

FIGURE 4.10: Histogram of station access time for walk mode with kernel density plot.



(A) WG (B) LATIS

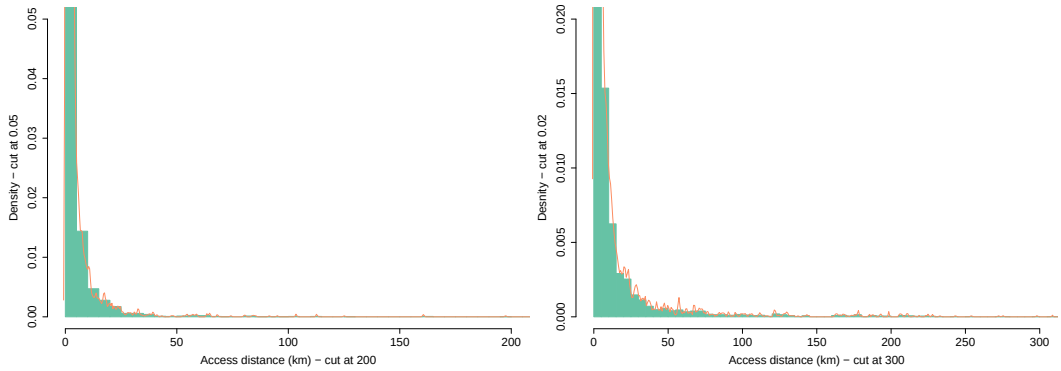
FIGURE 4.11: Histogram of station egress time for walk mode with kernel density plot.

4.5.1.2 Distance

Once the observations with walk access or egress time in excess of 60 minutes had been removed form both datasets, histogram and kernel density plots were produced for station access distance (Figure 4.12) and egress distance (Figure 4.13). Based on the observed distributions, trips with access or egress legs in excess of 70km were removed from the WG dataset, and those in excess of 200km were removed from the LATIS dataset. The distribution of access and egress distance is skewed further to the right in the LATIS dataset. A random review of some observations with access and egress legs of this magnitude, indicated that these could be valid trips. For example, someone travelling from a remote part of the Highlands and Islands might choose to drive into Inverness to begin their rail journey.

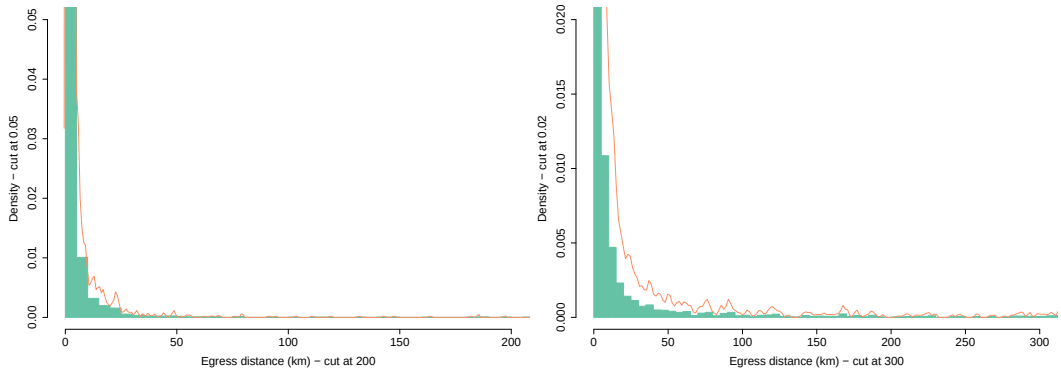
4.5.2 Illogical trips

There are two main types of illogical trips that are observed in this type of data. The first is the so-called 'reversed trip' where the origin station is located close to the ultimate destination, and the destination station is located close to the ultimate origin. The second occurs when



(A) WG (B) LATIS

FIGURE 4.12: Histograms of station access distance with kernel density plot.



(A) WG (B) LATIS

FIGURE 4.13: Histograms of station egress distance with kernel density plot.

there is a substantial 'back-track' from the reported destination station towards the trip origin. A range of ratios were tested on the WG dataset, using measures of components of the trip generated by the trip planner, that might reliably identify these illogical trips. Two ratios were found to be particularly effective.

The first, the RV ratio, captures the 'reversed trip' effect and is the distance from origin postcode to destination station over the distance from origin postcode to origin station, as shown in the following equation:

$$RV = \frac{D(op, ds)}{D(op, os)}, \tag{4.1}$$

where D is distance in km, op is origin postcode, ds is destination station, and os is origin station. The closer the ratio is to zero, the more pronounced the reversal effect becomes (see the illustrative example in Figure 4.14).

The RV ratio was calculated for each observation in the WG (clean) dataset and for ratios < 0.5, where the distance from the origin postcode to origin station is more than double the distance from the origin postcode to the destination station, the trips were visualised in

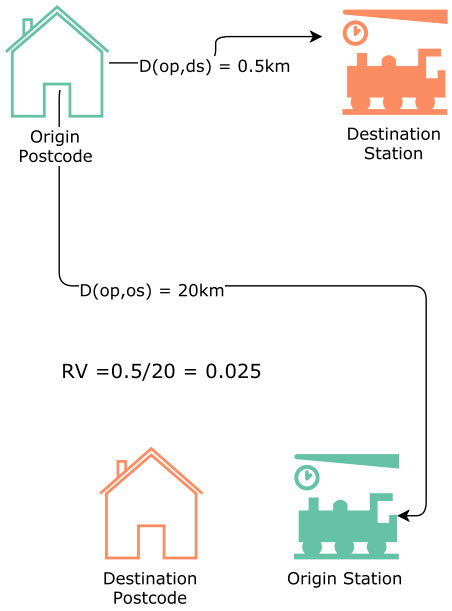


FIGURE 4.14: Illustrative example of the RV ratio.

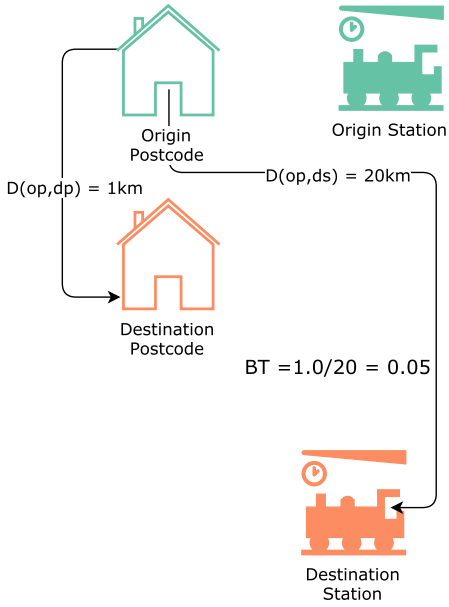
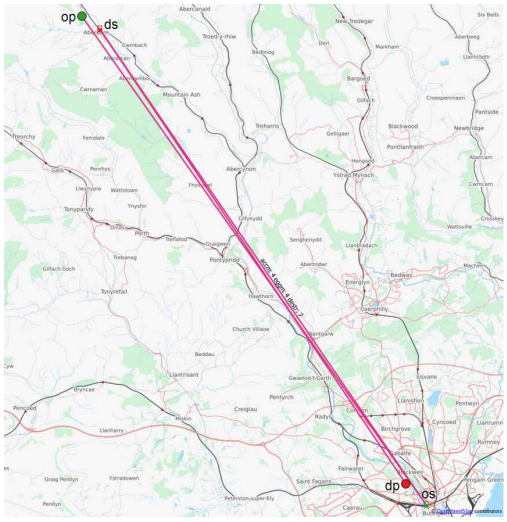
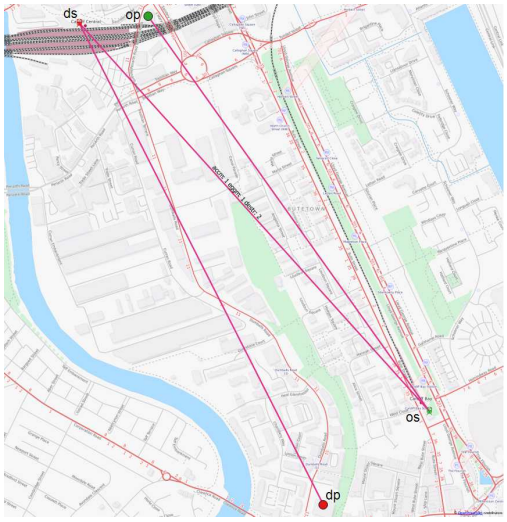


FIGURE 4.15: Illustrative example of the BT ratio.

QGIS. Figure 4.16 shows example trips with RV ratios of 0.04 and 0.41. One observation, with an RV ratio of 0.49, was considered a plausible trip, while the remaining observations with an RV ratio < 0.5 were removed from the WG dataset (a total of 20).



(A) RV ratio of 0.04



(B) RV ratio of 0.41

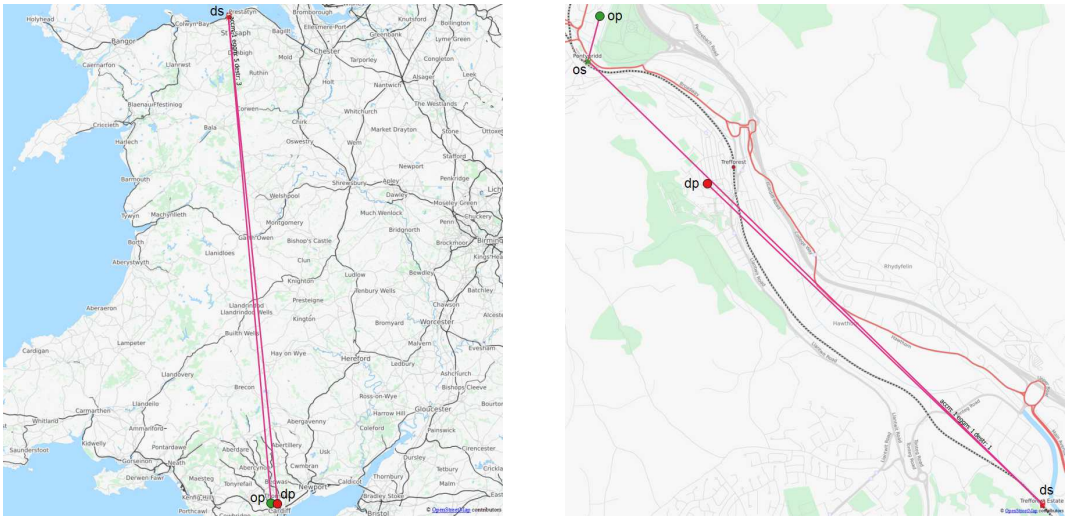
FIGURE 4.16: Example trips with stated RV ratios.

The second, the BT ratio, captures the 'back-track' effect and is the distance from origin postcode to destination postcode over the distance from origin postcode to destination station, as expressed in the following equation:

$$BT = \frac{D(op,dp)}{D(op,ds)}, \tag{4.2}$$

where *dp* is destination postcode. The closer the ratio is to zero, the more pronounced the back-track effect becomes (see the illustrative example in Figure 4.15).

The BT ratio was calculated for each remaining observation in the dataset (after the RV ratio validation), and for ratios < 0.5, where the distance from origin postcode to destination postcode is less than half the distance from origin postcode to destination station, the trips were visualised in QGIS. Figure 4.17 shows example trips with BT ratios of 0.01 and 0.27. Two observations, with BT ratios of 0.41 and 0.45, were considered plausible trips, while the remaining observations with an RV ratio < 0.5 were removed from the WG dataset (a total of 30).



(A) BT ratio of 0.01 (B) BT ratio of 0.27

FIGURE 4.17: Example trips with stated BT ratios.

For both the RV and BT ratios, the distance measures were obtained from the trip planner for walk mode. This was found to give more consistent results than using drive mode, primarily because the latter can produce longer circuitous routes caused by one-way systems that mask the relative geographical positioning of origins and destinations that the ratios are intended to detect. To establish the effectiveness of the steps taken to remove illogical trips, 100 random observations were selected from the WG dataset (after removal of trips as determined by the RV and BT ratios) and their reported trips were individually visualised in QGIS. All 100 trips were considered logical. Based on the findings from working with the WG dataset, all trips with RV and BT ratios < 0.5 were automatically removed from the LATIS dataset. Due to the larger size of this dataset it was not considered practical to individually verify these trips by visualising them in QGIS.

Figures 4.18 and 4.19 detail the adjustments made to the WG and LATIS datasets as a result of the automated trip validation process. In the case of the LATIS dataset the observations with illogical trips were removed prior to removing those with excessive access or egress legs. When the methodology was first developed and tested using the WG dataset, the illogical

trips were removed *after* observations with excess walk-time legs had been removed. It was considered a more robust approach to produce the kernel density plots used to identify the appropriate cut-off points for excessive access and egress legs after the illogical trips had been identified and removed. In addition, observations from questionnaires completed on or after 6 September 2015, when the new Borders Railway line opened towards the end of the 2015 survey period, were removed from the LATIS dataset.

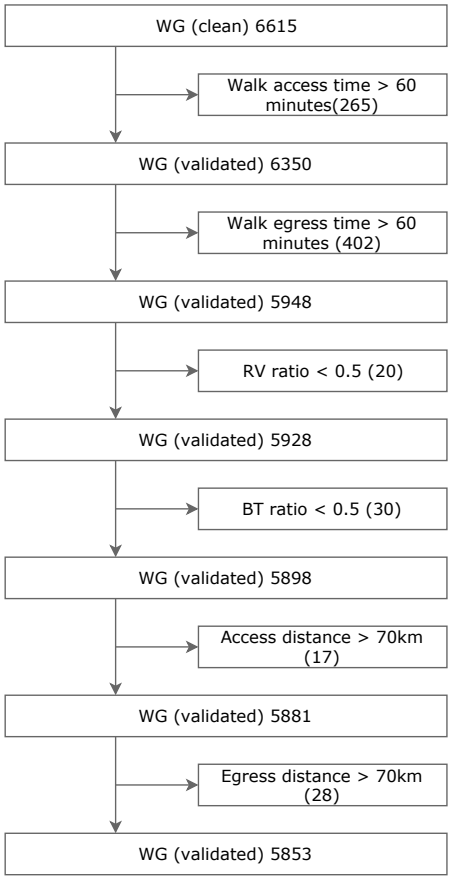


FIGURE 4.18: Trip validation adjustments made to the WG dataset.

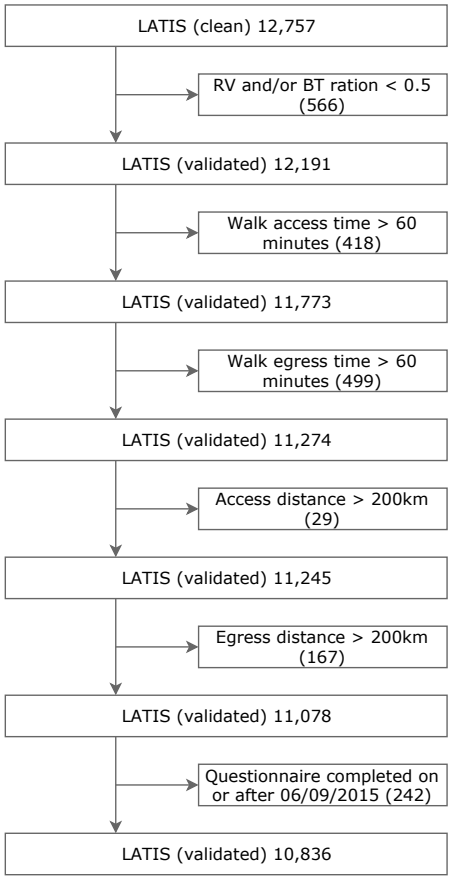


FIGURE 4.19: Trip validation adjustments made to the LATIS dataset.

4.6 Descriptive analysis

4.6.1 The access and egress journey

The observations in the cleaned and validated datasets were disaggregated by access and egress mode. The breakdown of observations by mode, along with average access and egress distances for each mode are presented in Tables 4.2 and 4.3, while histograms of access and egress mode are shown in Figures 4.20 and 4.21.

Walk is by far the dominant means of station access and egress, followed by car and then public transport. A somewhat higher proportion walked to or from the station in the WG

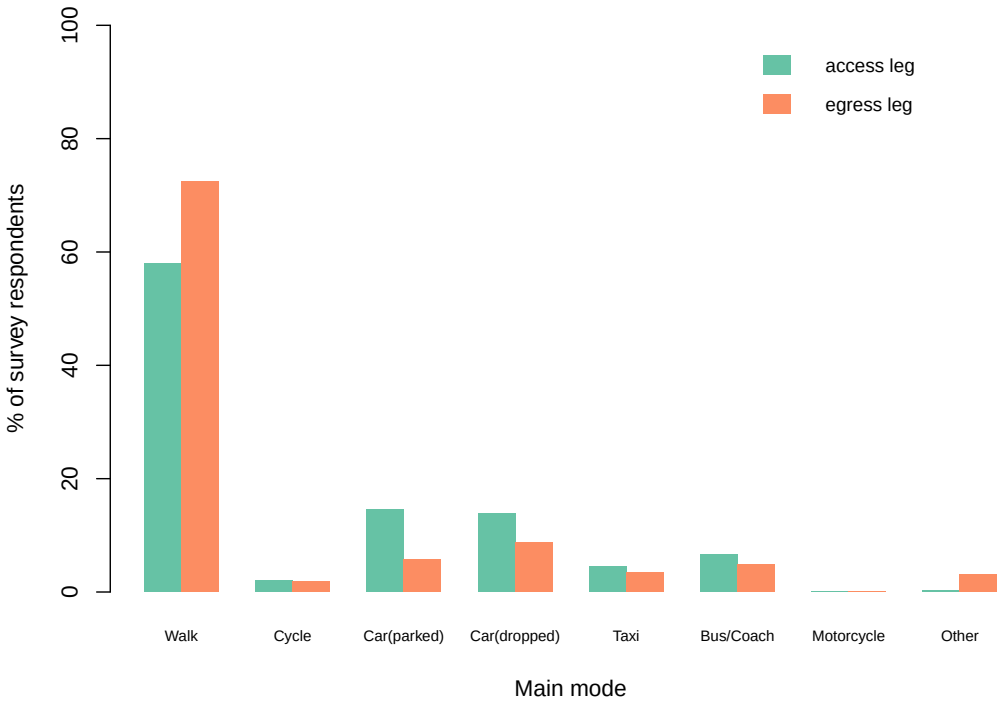


FIGURE 4.20: Responses disaggregated by main access or egress mode — WG dataset.

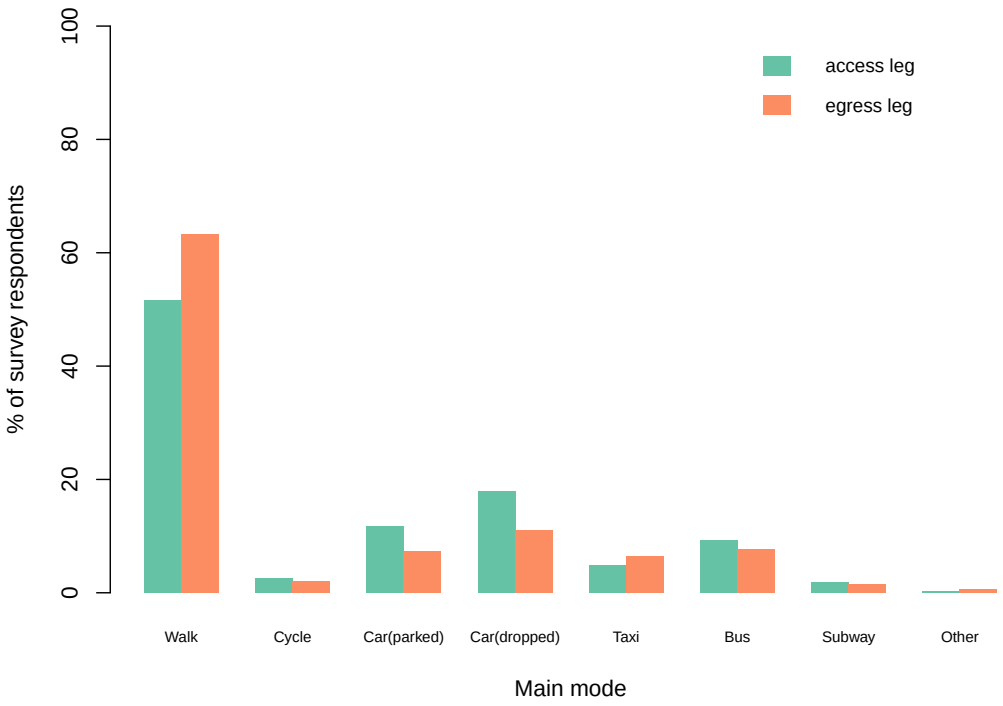


FIGURE 4.21: Responses disaggregated by main access or egress mode — LATIS dataset.

	Access leg			Egress leg		
Mode	No.	%	Avg. dist. (km)[1]	No.	%	Avg. dist. (km)[1]
Walk	3393	58.0	1.13	4235	72.4	1.14
Cycle	119	2.0	3.19	104	1.8	2.85
Car (parked)	849	14.5	6.04	332	5.7	5.69
Car (dropped)	815	13.9	5.79	516	8.8	6.71
Taxi	263	4.5	3.57	200	3.4	4.53
Bus/Coach	388	6.6	6.03	281	4.8	6.41
Motorcycle	8	0.1	10.85	6	0.1	4.4
Other	18	0.3	5.67	79	3.1	6.19

[1] In all cases street distance measured using walk mode

TABLE 4.2: Observed trips disaggregated by access and egress mode — WG dataset.

	Access leg			Egress leg		
Mode	No.	%	Avg. dist. (km)[1]	No.	%	Avg. dist. (km)[1]
Walk	5596	51.7	1.06	6858	63.3	0.97
Cycle	269	2.5	6.18	220	2.0	5.68
Car (parked)	1271	11.7	11.59	802	7.4	16.76
Car (dropped)	1933	17.9	9.43	1191	11.0	10.63
Taxi	521	4.8	8.22	690	6.4	13.28
Bus/Coach	1006	9.3	10.92	839	7.8	11.76
Subway	201	1.9	4.93	169	1.6	6.31
Other	34	0.3	37.34	62	0.6	28.89

[1] In all cases street distance measured using walk mode

TABLE 4.3: Observed trips disaggregated by access and egress mode — LATIS dataset.

dataset (58% and 72%) compared to the LATIS dataset (52% and 63%). While the same proportion drove to, or were dropped-off at, the station in both cases (around 28%), public transport access (bus, coach or subway) is more important in the LATIS dataset (11.2% compared to 6.6%). A similar difference is seen for the egress leg, where public transport accounts for 9.4% of LATIS journeys, but only 4.8% of WG journeys. Although both datasets have a similar average access and egress distance for walk mode (around 1 km), the average distances for other modes are noticeably higher in the LATIS dataset. This is as expected, given that the distribution of non-walk access and egress distances is skewed further to the right in the LATIS dataset, as discussed in Section 4.5.1. The average access and egress distances are particularly high for the 'other' category in the LATIS dataset, and this appears to be largely the result of several respondents using ferry or boat to travel from islands to the mainland, resulting in particularly long journeys when measured on the road network (via an alternative road bridge).

It is interesting to note that a higher proportion walked and a lower proportion used the car for the egress journey compared to the access journey in both datasets, with the effect more pronounced in the WG dataset. It should be borne in mind that the survey data does not consist of a uniform type of trip. For example, it is not limited to trips originating at the respondent's home address. For some trips the access journey will be from home to a station, while for other trips the access journey will be from a place of work to a station (and similarly for the egress journey). To explore whether the mix of trip types within the datasets could help to explain some of the findings discussed above, the observations were disaggregated by the nature of the origin and destination[5] (see Table 4.4). The analysis reveals that 62% of respondents in both datasets began their journey at home, and only 30% (WG) and 41% (LATIS) were travelling home. This indicates that the datasets do not contain a balanced set of trips (more people are leaving home than returning home). This may explain why a higher proportion of respondents used walk mode, and a lower proportion used one of the car modes, for the egress journey compared to the access journey, as this would be intuitively expected. A car is much more likely to be available at the home end of a journey and travellers are more likely to be reliant on walking from the egress station when the destination is not their home.

To further investigate the potential reason for an imbalance in the trip types, frequency histograms were produced for the time of travel[6] (see figures 4.22 and 4.23). The histograms reveal that more surveys were carried out in the morning peak for both datasets, which would explain the higher proportion of trips with home as the origin. After the morning peak, the WG surveys were fairly evenly spread throughout the remainder of the day, while their were fewer LATIS surveys during the rest of the morning and the afternoon, before another large peak in the evening. This probably accounts for the greater proportion of workplace origins in the LATIS dataset (31%) compared to the WG dataset (18%), the correspondingly higher percentage of respondents returning home (41% compared to 30%), and fewer shopping and leisure-related origins ($< 4\%$ compared to $> 14\%$). Indeed the LATIS dataset is dominated by home and work origins, which account for 93% of trips, compared with 80% in the WG dataset.

To establish how representative the mix of access and egress modes observed in the survey data is of rail trips in general across the UK, it was compared with data collected by the National Rail Passenger Survey (NRPS) during the first quarter of 2015 (Transport Focus, 2015a). Although the primary focus of the NRPS is to assess customer satisfaction with rail services and facilities, it also asks respondents what methods of transport they used to get to and from the ultimate origin and destination station of their journey. The Spring 2015 NRPS covered all of GB and consisted of some 30,000 responses (for more background information about the survey see Transport Focus (2015b)). The data is presented as a bar graph in

[5]The WG questionnaires asked the 'reason for being at the origin or destination' and the LATIS questionnaires asked 'where have you come from?' and 'where are you travelling to?'. The response options varied slightly and have been grouped into wider categories in the summary table.

[6]This is taken as the interview time for the WG dataset and the service start-time for the LATIS dataset.

	Origin				Destination			
	WG		LATIS		WG		LATIS	
Reason	No.	%	No.	%	No.	%	No.	%
Home or other accommodation	3652	62.40	6715	62.00	1775	30.33	4445	41.04
Usual workplace (or work-related)	1027	17.55	3326	30.71	2253	38.49	4361	40.26
Education	348	5.95	277	2.56	335	5.72	484	4.47
Shopping	215	3.67	49	0.45	367	6.27	184	1.70
Other (e.g. leisure, tourism, personal)	611	10.44	325	3.00	1123	19.19	1218	11.25
Unknown	0	0.00	139	1.28	0	0.00	139	1.28
Total	5853		10831		5853		10831	

TABLE 4.4: Reason for respondent being at trip origin or going to trip destination.

Figure 4.24. Unfortunately it is not possible to directly compare the mode breakdown shown in this figure with that shown for the survey data in Figures 4.20 and 4.21. This is because the NRPS includes overground (national rail) services as one of the access/egress mode options, and these were removed from the WG and LATIS datasets during data cleaning (as discussed in Section 4.4); and because NRPS respondents were not restricted to specifying only the main mode used (i.e. the sum of the mode percentages in Figure 4.24 exceeds 100). However, it does suggest that the mode split observed in the revealed preference data is broadly consistent with the mode split observed across the country. The NRPS is dominated by respondents using trains in London and the South East (63% of respondents), and this will account for the higher mode share for subway compared with the study datasets.

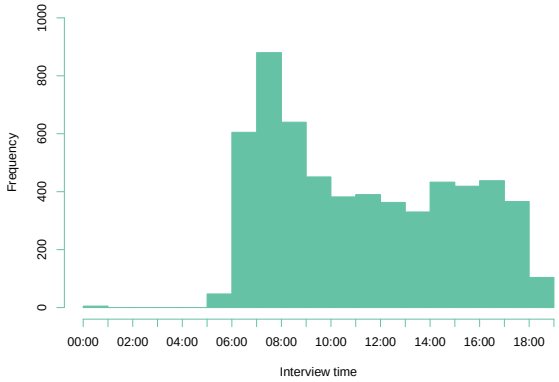


FIGURE 4.22: Histogram of interview time for WG observations.

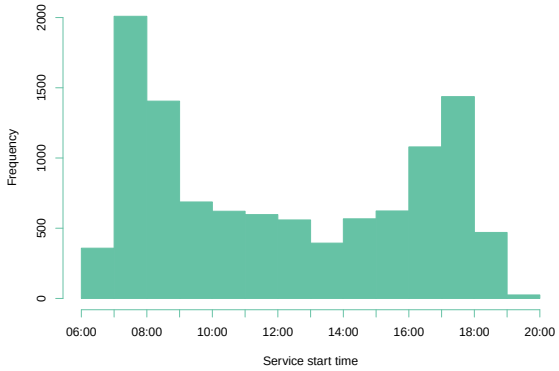


FIGURE 4.23: Histogram of service time for LATIS observations.

4.6.2 Rank of chosen station

It is an assumption of the catchment definitions used in the aggregate demand models discussed in Section 2.4.1, that rail passengers choose their nearest station. To explore the extent to which this reflects reality, the 30 closest stations to each survey origin were identified and then ranked by drive distance. The process for identifying the candidate stations and measuring the distances is described in detail in Section 6.3.

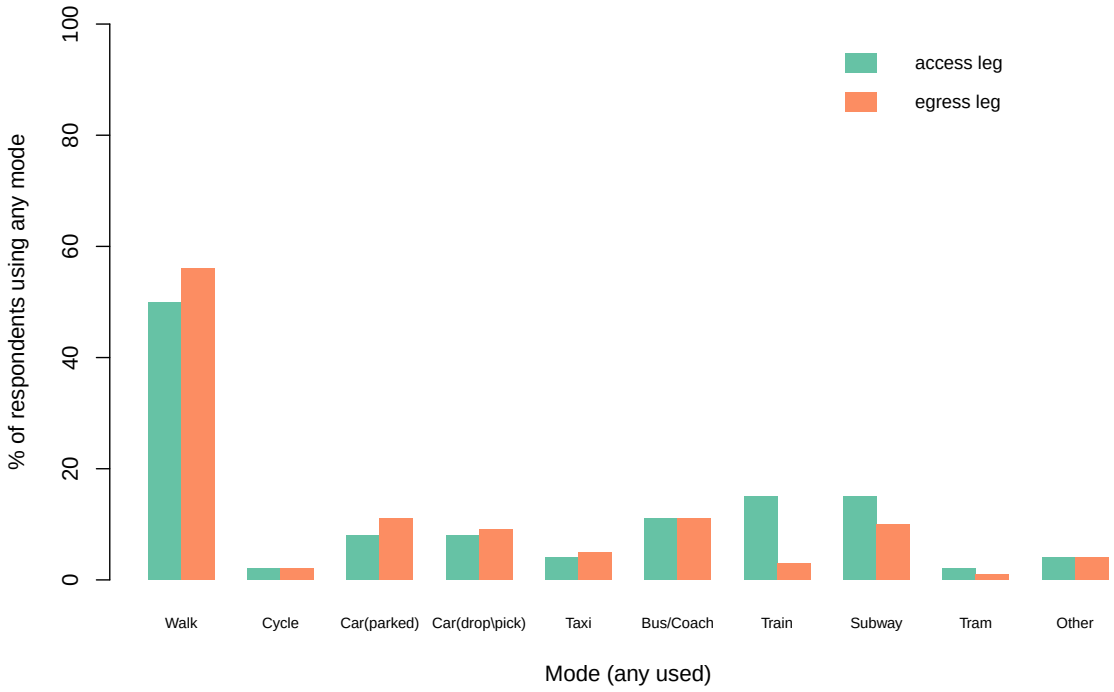


FIGURE 4.24: Reported modes used to access and egress stations (GB), from National Rail Passenger Survey, Spring 2015. Note: As respondents can select more than one mode, the sum of percentages exceeds 100. (Transport Focus, 2015a)

The percentage of observations choosing a station of each distance rank, for all access modes and disaggregated by the predominant access modes (walk, car and bus), is shown in Figures 4.25 and 4.26 for the WG and LATIS datasets respectively. Considering all modes, 69% of WG and 63% of LATIS respondents boarded the train at their closest station (as measured by drive distance). However, this overall figure disguises substantial differences between access modes. A far greater percentage of those who walked to the station chose their nearest one (WG: 81% and LATIS: 75%), while only around half of those driving or being dropped by car (WG: 52% and LATIS: 55%) or using the bus (WG: 48% and LATIS: 45%) selected their nearest station. While 95% of travellers who accessed the station by foot chose a station ranked below 5th (WG) or 6th (LATIS), for car and bus users the rank of station by which 95% of traveller's choice was accounted for was much higher, as indicated by the shallower cumulative percent curves. It is also interesting that a small (but not insignificant) proportion of bus and car passengers chose a station ranked below 20, for example 15% of those taking the bus in the LATIS dataset. One possible explanation is that a search in all directions from the origin picks up many small and medium sized stations which are being ignored by the traveller in preference to a more distant large inter-city station. This has potential implications for the selection of stations for an individual's choice set, suggesting that the definition may need to be more nuanced than one simply based on x nearest stations.

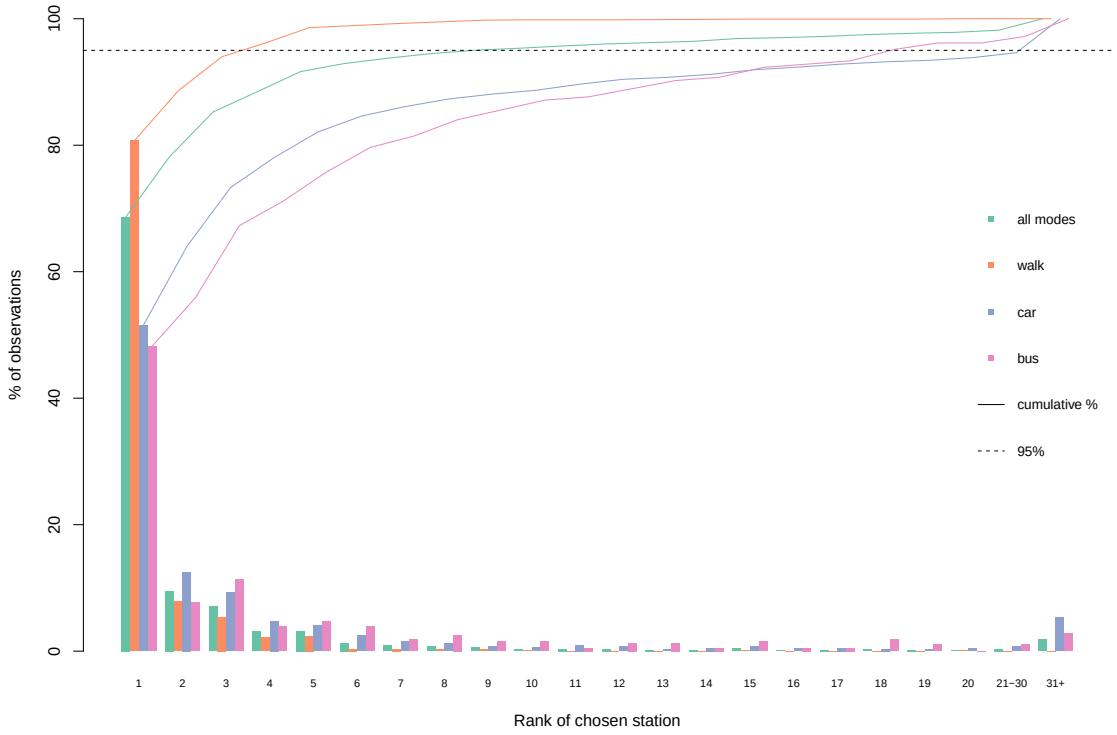


FIGURE 4.25: Rank of chosen station disaggregated by key modes (all ranks based on drive distance) — WG dataset.

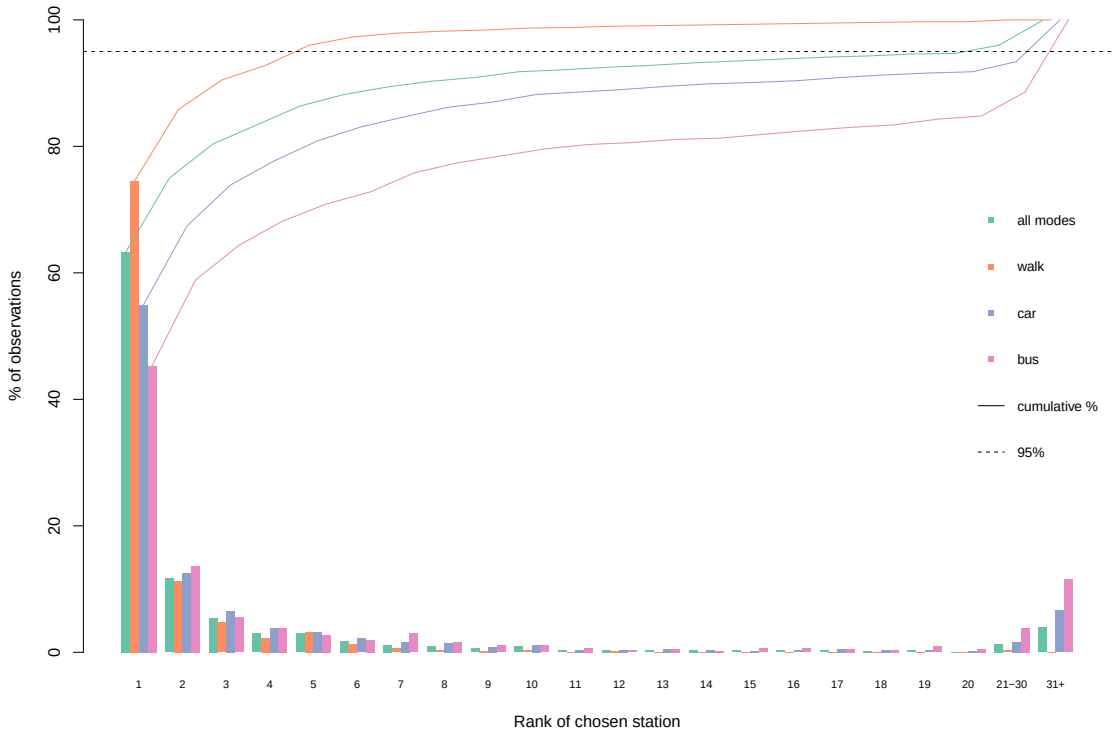
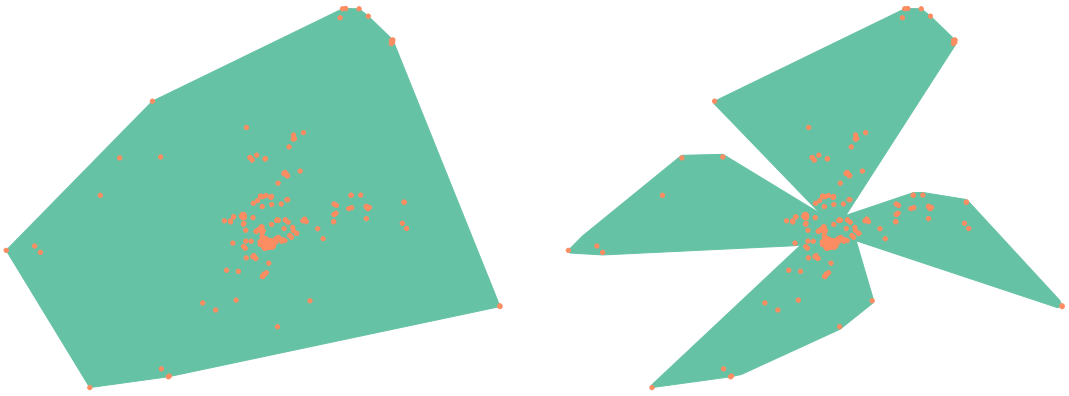


FIGURE 4.26: Rank of chosen station disaggregated by key modes (all ranks based on drive distance) — LATIS dataset.

4.6.3 Observed station catchments

As discussed in Section 2.4.1, the catchment definitions used in aggregate demand models assume that station choice is a deterministic process, that there is only one station that everyone within a zone will choose, and that station catchments are therefore discrete entities that do not overlap with one another. The large number of observations available in the datasets used in this study made it possible to create approximate representations of actual station catchments to test the validity of these assumptions. Given that the validated datasets consisted of over 300 distinct origin stations, an automated process was developed to generate the catchments using an R script and associated spatial database queries. The main steps performed for both the WG and LATIS datasets separately are summarised below (see R code segment A.5 in Appendix A):

1. For each distinct origin station, a temporary database table was created to hold the origin coordinates for all observations with that station as the origin station.
2. A polygon around the set of origins was created using the `ST_ConcaveHull` function (The PostGIS Development Group, 2017). This function is often described as placing a shrink wrapping that encloses the set of points, with the amount of 'vacuum sealing' controlled by the target percent parameter. A target of 0.99 was specified, after comparing the results obtained using a target of 0.99 and 0.98 on the polygon for Inverness station (see Figure 4.27). It was considered that the 0.98 catchment, although correcting for the 0.99 catchment extending over the sea in the North East, otherwise produced an odd shape catchment with gaps over land that are more likely to reflect the limitations of survey size than areas that are outside of the station's true catchment.
3. The polygon was written to a database table storing all the station catchment polygons for the dataset.



(A) Target set to 0.99 (B) Target set to 0.98

FIGURE 4.27: Polygons encompassing trip origins with Inverness as origin station using `ST_ConcaveHull` function and stated target values.

	WG		LATIS	
No. of catchments	No.	%	No.	%
1	9287	18	16964	11
2	10214	20	9596	6
3	9409	18	16331	11
4	6663	13	15373	10
5	3834	8	12215	8
6	3255	6	11662	8
7	2528	5	13738	9
8	2512	5	16257	11
9	1478	3	15443	10
10	1173	2	11133	7
11	647	1	8208	5
12	43	0	1823	1
13	1	0	758	1
14	0	0	261	0
15	0	0	4	0
Total	51044	100	149766	100

TABLE 4.5: The number of unit postcode polygons that are intersected by x (1–15) station catchments for the WG and LATIS datasets.

After creating the catchment polygons, to obtain further insight into the potential heterogeneity of station choice within zones, a spatial analysis was carried out to identify the number of station catchments that each postcode falls within. This involved identifying the set of distinct postcode polygons (from the OS Code-Point with Polygons dataset) that intersect any of the station catchments (as produced above) and then counting the number of unique catchments intersected by those postcode polygons (see R code segment A.6 in Appendix A). The outline of the station catchments and the postcode polygon catchment counts were then visualised using QGIS and a choropleth map was produced for each (See Figures 4.28 and 4.29). The breakdown of postcode polygons by the number of station catchments in which they fall is shown in Table 4.5. The complex interaction of the catchments is clearly apparent and there is little evidence to support the notion of stations having discrete non-competing catchments. Even for areas that appear to be only within a single catchment (for example parts of the Scottish Highlands and the west of Wales), this is due to the limited scope of the passenger surveys. For example, passengers choosing stations on the Inverness to Thurso and Wick line, the Inverness to Kyle of Lochalsh line, stations west of Swansea (apart from Carmarthen) and stations on the Heart of Wales line, are not represented. Even with these limitations, 62% (WG) and 83% (LATIS) of postcode polygons are within 3 or more station catchments, and 4% (WG) and 15% (LATIS) are within 10 or more catchments.

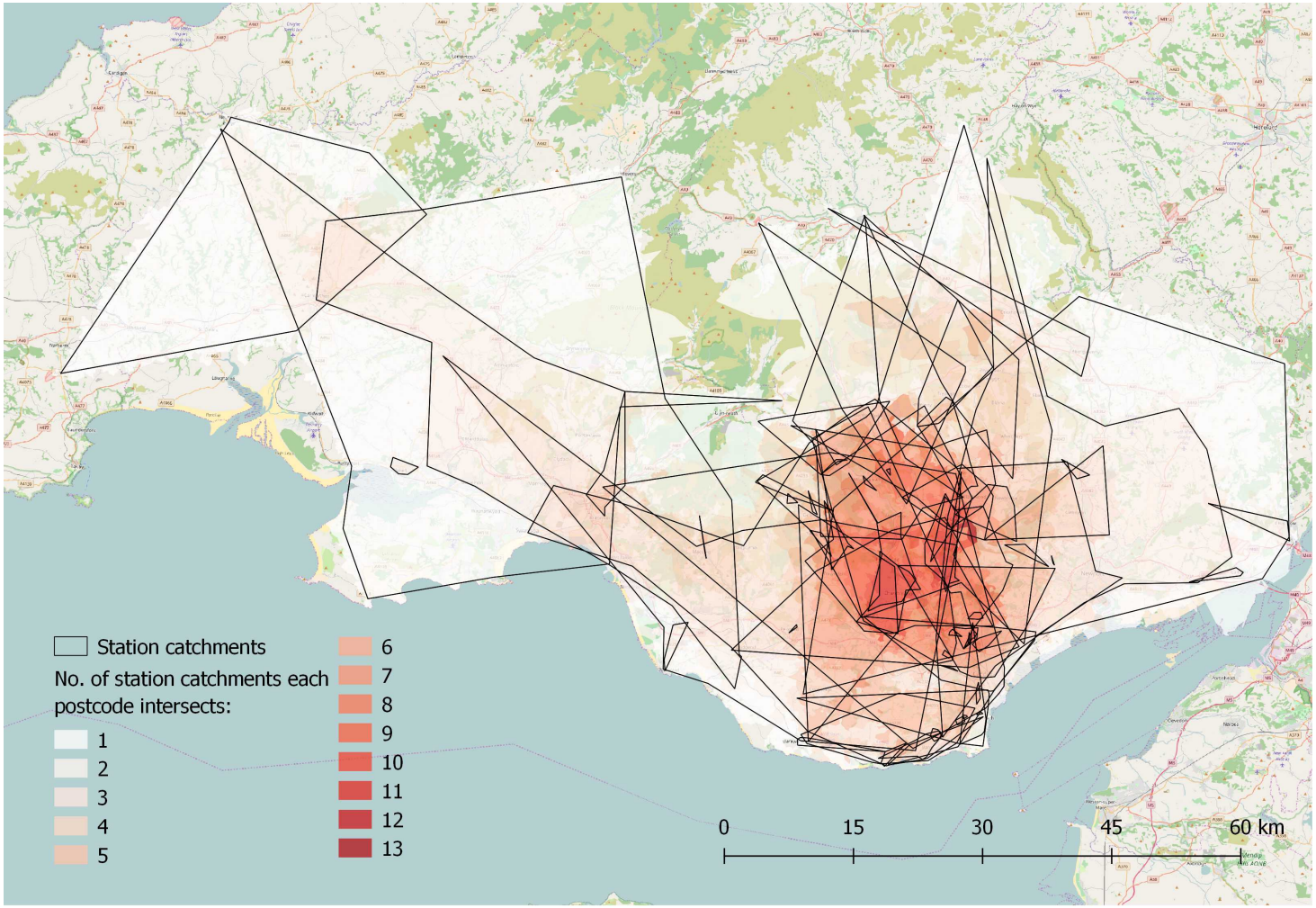


FIGURE 4.28: Approximate observed station catchments generated for the WG validated dataset, with each postcode classified to show the number of station catchments that it intersects.

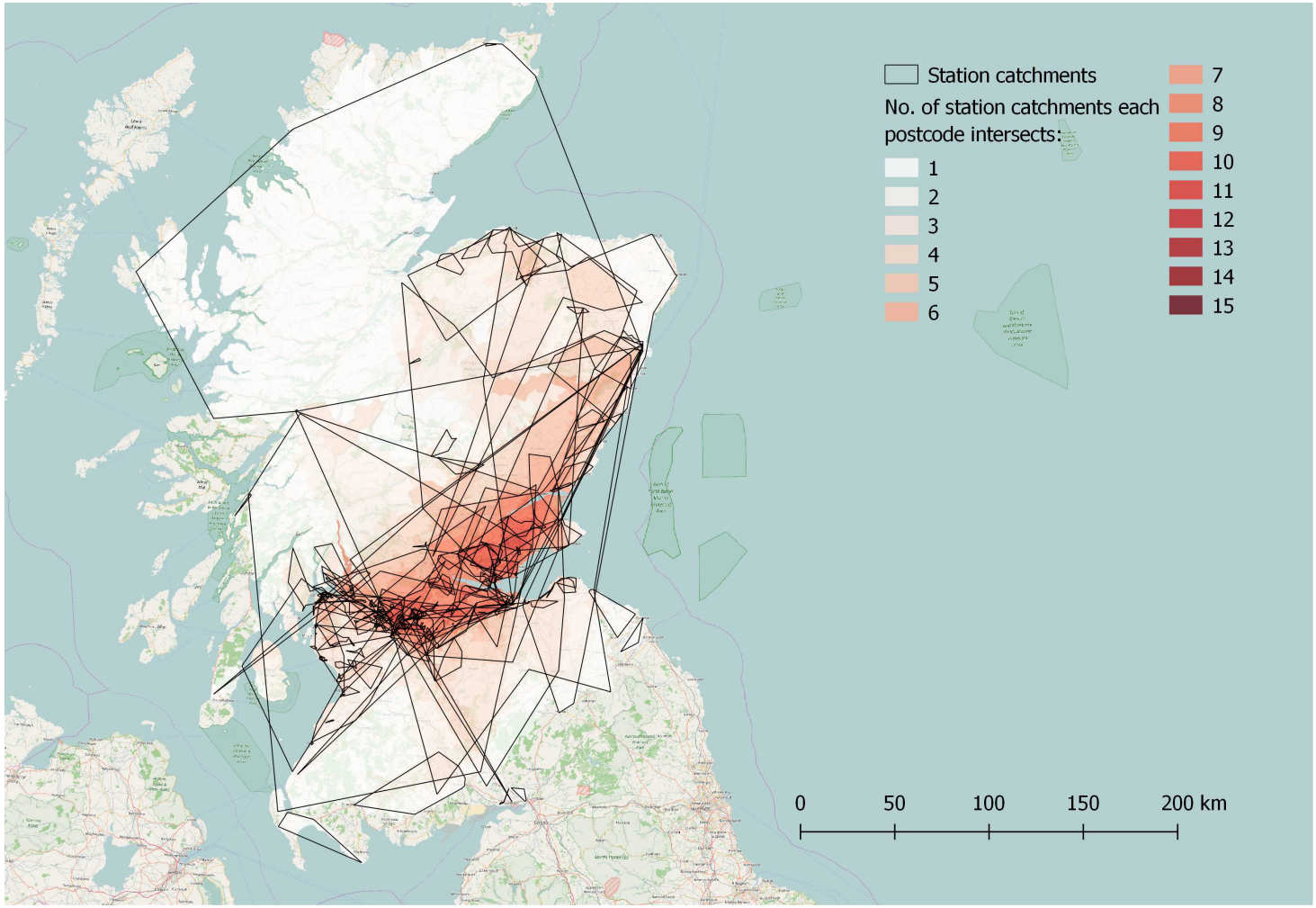


FIGURE 4.29: Approximate observed station catchments generated for the LATIS validated dataset, with each postcode classified to show the number of station catchments that it intersects.

4.7 Conclusions

The first part of this chapter described the sources of observed station choice data used for this study and the procedures that were applied to prepare, clean and validate them. As part of this process several potentially novel approaches were developed that may have wider applicability to other researchers conducting research using similar survey data. These are the matching of incomplete textual address information to unit-level postcodes; the estimation of coordinates of an origin or destination known to be located on a particular street, based on the spatial relationship of the set of postcode centroids for that street; and an automated system to identify the two types of illogical trip that are common in this type of survey data, using information inherent to the reported trip. These approaches maximise the usefulness of data that is very expensive to gather (observations are not discarded due to missing postcodes); ensure a broader range of trip types are represented (they are not limited to the type of trip where the respondent is likely to know the address postcode); and efficiently identify errors in the self-reported trips (which would otherwise be overlooked or subject to a very costly manual review process).

The second part of this chapter used descriptive analysis techniques to examine some key aspects of observed station choice revealed by the survey data. The mix of access and egress modes present in the survey data was found to be broadly consistent with the national picture, suggesting that, from this perspective, it might be suitable for calibrating models that can be usefully applied beyond the study areas. However, both datasets were found to have an imbalance of trip types, with almost two-thirds originating at home. An examination of the distance-rank of the chosen station revealed that although in the majority of cases the nearest station was chosen, a substantial proportion of respondents chose to board their train at a more distant station, especially those accessing the station using a motorised mode. This was followed by a spatial analysis that constructed approximate observed station catchments for the stations present in the survey data. This revealed that a substantial proportion of postcode polygons are located within more than one estimated station catchment, with many in considerably more. These findings undermine the simple catchment definitions that are used in the aggregate demand models typically applied to forecast demand for new local stations in the UK, and support the objective of this study to develop a more sophisticated methodology that better represents the complex nature of station catchments.

This chapter has focussed on one of the key inputs required for any discrete choice model, the observed choice. The next chapter is concerned with the attributes that might help explain this observed choice behaviour, and how these were derived from a range of disparate open data sources.

Chapter 5

Station choice predictor variables

5.1 Introduction

This chapter describes the potential predictor variables that were chosen to be tested during the subsequent calibration of the station choice models. It explains the range of data sources that were utilised, and how the variables were obtained from them. The chapter begins by describing the implementation of a bespoke multi-modal route planner, identified in the project objectives as a key requirement to enable a realistic representation of the station access and train leg components of the reported and available alternative trips (Section 5.2). An automated framework that was developed to enable the predictor variables to be efficiently generated from disparate open transport data sources is then outlined in Section 5.3, followed by a detailed explanation of how each predictor variable was derived (Section 5.4). The chapter then closes by drawing some key conclusions.

5.2 Implementing a multi-modal route planner

As the access journey is such an important factor influencing station choice, a key objective of this research was to generate a realistic representation of these journeys, for the station chosen by each survey respondent and the alternative stations available to them, taking into account the actual access mode used. In addition, the station choice models suitable for incorporating into flow demand models would need to include predictor variables able to describe the characteristics of the available train legs, such as on-train time, waiting time and the number of transfers. These requirements necessitated a trip planning tool that was able to generate routes for a range of motorised and non-motorised transport modes.

5.2.1 Identifying a suitable routing tool

A review of commercial and open source tools was conducted, and three potential solutions were identified: Google Maps API, Visography TRACC and OpenTripPlanner (OTP) (OpenTripPlanner, 2018).

5.2.1.1 Google Maps API

Although Google Maps is able to route using the UK public transport timetables for rail, coach and bus, access to the API is heavily restricted, both by limitations on the number of API calls from an IP address (typically 2,500 calls per day) and by restrictive usage conditions[1]. A further limitation is that it is based on current published timetables. It would not be possible to load historic timetable data to match the date that the on-train passenger surveys were carried out, nor to add new public transport routes, adjust service frequencies or add station stops. The ability to alter the current network would be necessary if the station choice models were used to forecast demand for a new station or the effect of substantial service changes, as the predictor variables would need to be generated based on this new situation.

5.2.1.2 Visography TRACC

Visography TRACC is commercial software developed by Basemap Ltd that is popular among transport planners in consultancies and local government. The key advantage of Visography TRACC is that it can import standard UK public transport data formats: TransXChange, Association of Train Operating Companies (ATOC) common interface file, and NaPTAN. A key disadvantage is that the user is limited to the analysis tools provided in the software. The primary focus of TRACC is accessibility analysis, generating total travel time or distance between a set of OD pairs. The software can produce a 'Full OD-Path File' that contains more detailed information about each journey, such as walk time and interchange time, but the help pages warn that 'it would be best doing so with a small set of origin points [and] only the first 50 origin points will have a path report' (Basemap Ltd, 2014). Furthermore, given that the UK public transport data is freely available under various open data initiatives, a solution that is not reliant on commercial software was considered preferable. Open transport data is of little practical benefit to a researcher who does not have access to suitable tools with which to analyse it.

[1] For example, Google prominently displays the following warning in the Google Distance Matrix API developer information: 'use of the Distance Matrix API must relate to the display of information on a Google Map; for example, to determine OD pairs that fall within a specific driving time from one another, before requesting and displaying those destinations on a map. Use of the service in an application that doesn't display a Google map is prohibited.' (Google, 2015).

5.2.1.3 OpenTripPlanner

OTP is an open-source and cross-platform multi-modal route planner written in JAVA that uses imported OpenStreetMap (OSM) data for routing on the street and path network and supports multi-agency public transport routing through imported General Transit Feed Specification (GTFS) data. It can also apply a digital elevation model to the OSM street network, allowing, for example, cycle-friendly routes to be requested. OTP has a web front-end that can be used as a trip planner by end-users and a sophisticated RESTful API. This was considered the most promising platform, as scripts could be written in R to query the API and process the returned data. In addition, this work could be extended in the future to develop a comprehensive OTP API wrapper as an R package, which would benefit the wider research community.

5.2.2 Building the multi-modal network

OTP has a high random access memory (RAM) requirement when building the trip planner graph[2] from the large datasets involved in this study. The graph build stage was therefore carried out on a Microsoft Azure Linux Cloud Server with 56 GB of RAM[3], and the graph was then transferred to a local server with 16 GB RAM for normal operation of the trip planner. For testing purposes, an initial graph was built using current OSM data for Great Britain obtained from Geofabrik (Geofabrik, 2015) and GTFS data for GB National Rail services which had been converted from the ATOC common interface format (GB Rail, 2015). Although this initial work resulted in a fully-functioning trip planner suitable for walk, cycle and rail modes, a couple of deficiencies were identified:

- The release version of OTP assumes that OSM roads tagged with 'highway=trunk' can only be traversed by cars. While in some countries walking and cycling are not permitted on trunk roads, in the UK there is no real distinction between trunk and primary roads other than the body responsible for them. To correct this anomaly the source code was amended to give traversal permission to all modes on trunk roads.
- After testing recommended drive routes based on local knowledge it was apparent that OTP was suggesting unlikely routes via narrow unclassified roads. The UK OSM tagging guidelines indicate that roads tagged as 'tertiary' are considered to be busy unclassified roads wide enough to allow two cars to pass safely (OpenStreetMap, 2015). However, the release version of OTP has the average speed for tertiary roads set the same as unclassified and residential roads, at 25 mph. The source code was amended to increase the average speed of tertiary roads to 35 mph. Other adjustments included raising the average speed of secondary roads from 35 mph to 40 mph, and adjusting the speed

[2]The trip planner graph specifies every location in the region covered and how to travel between them. It is compiled from the OSM and GTFS data.

[3]It is only necessary to build a new graph when the underlying public transit data or OSM street network requires updating. Graph build is therefore likely to be an irregular occurrence.

of primary roads and motorways to 47 mph and 67 mph (from 45 mph and 65 mph) respectively, based on published free-flow road speeds (Department for Transport, 2015). These changes resulted in more realistic driving routes.

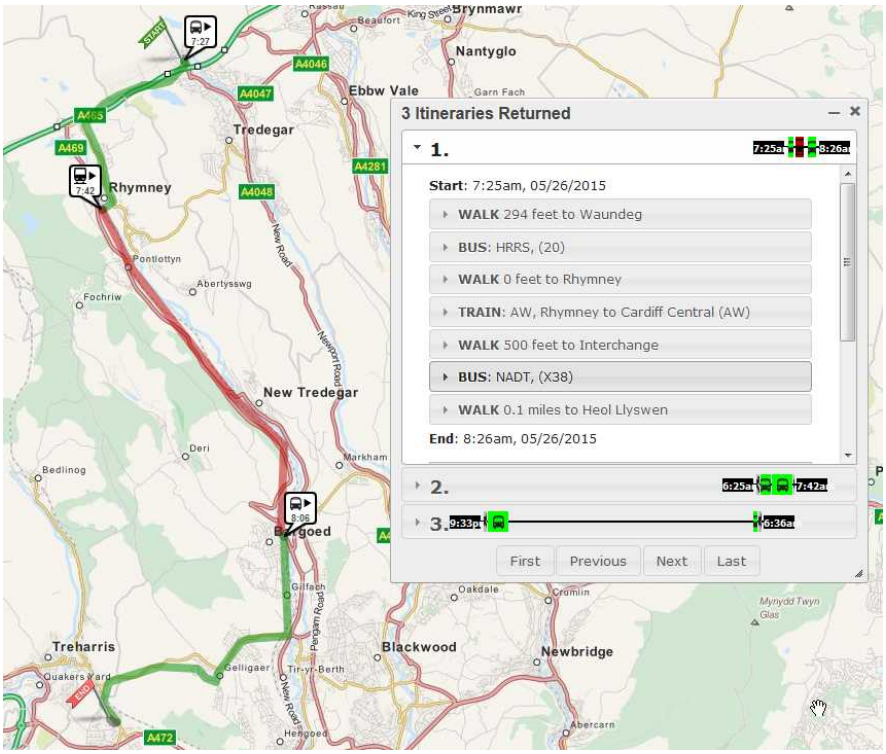


FIGURE 5.1: The OTP web interface, with example walk, bus and train trip itinerary.

The next stage in the test implementation of OTP was to incorporate public transit (bus) timetable data for Wales, obtained from the Traveline National Dataset (TNDS). This dataset is only available in the TransXChange format, a UK standard consisting of an XML schema for the exchange of bus routes and timetables. Attempts were made to locate a reliable tool to convert from TransXChange to GTFS format. The open source TransXChange2GTFS Converter (GoogleTransitDataFeed, 2016) was investigated but was found to abort when processing the vast majority of XML files in the TNDS, despite the files passing validation in the official TransXChange Publisher tool available from the Department for Transport. The converter has not been updated since 2012, probably as a result of GTFS becoming the de-facto standard for publishing public transport data around the world, with the UK now a notable exception, and it was rejected as a plausible solution. The only available alternative was Visography TRACC, which is able to import TransXChange files and export a public transport network to a GTFS feed. After completing this conversion, a number of error checking, correction and clean-up processes were performed on the GTFS feed before it was used for an OTP graph build, either to prevent fatal build errors or to improve performance. Figure 5.1 shows the OTP web interface once the transit data had been incorporated, with an example trip itinerary in the Rhymney Valley (Wales) using walk, bus, and train modes.

5.3 An automated framework to derive model variables

It was recognised early on that deriving the predictor variables for the station choice models would involve the collation of a large amount of data from a range of disparate open transport data sources, and that a set of automated processes would be needed to handle this in an efficient, reliable and accurate manner. In a discrete choice model variables must be derived for every alternative in the choice set, thus increasing the number of observations in the model by at least an order of magnitude. A data processing framework was therefore developed that could automatically populate database tables with attributes obtained from internal and external data sources. The framework consists of a PostgreSQL database, the R software environment, an internal OTP route planner, and various external data sources. A generic version of the framework is described in Young (2016), and the components and how they interact are illustrated in Figure 5.2. Further information about the framework's components, in the specific context of this research project, are given in the sections that follow.

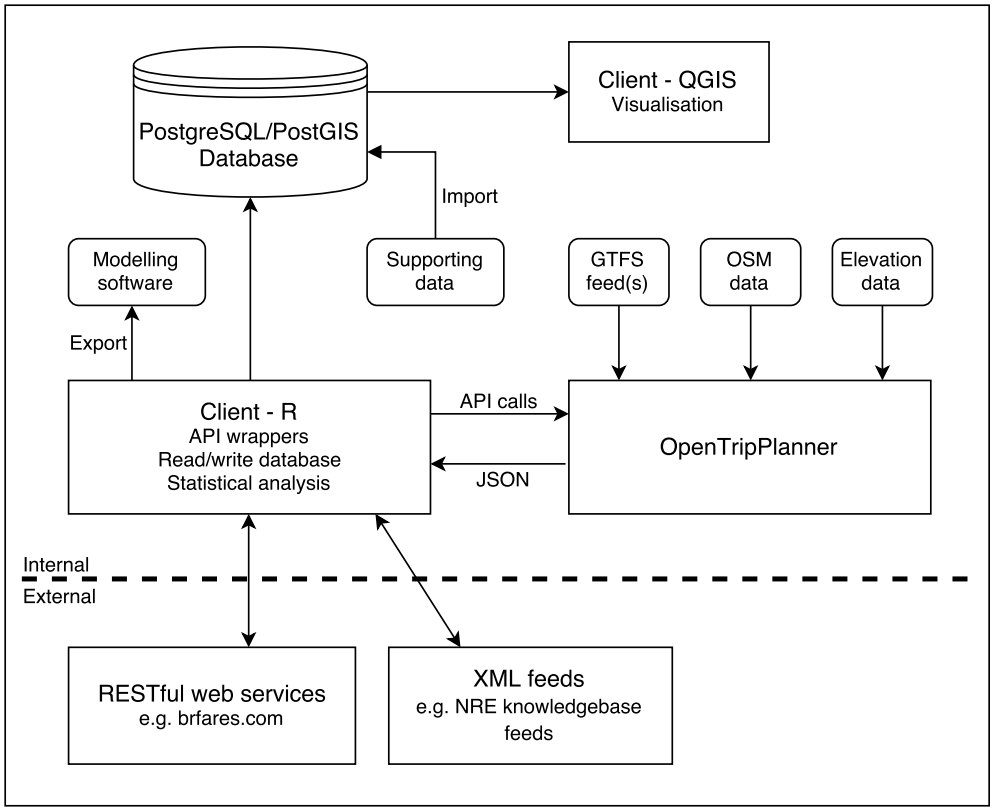


FIGURE 5.2: Framework to derive explanatory variables from disparate open transport data sources.

5.3.1 PostgreSQL database

The PostgreSQL database, spatially-enabled using PostGIS, is used to store data and perform spatial and non-spatial queries. Tables were grouped into three schemas, one for each dataset

and one for supporting data. The key tables for the 'latis' and 'data' schemas and their relationships, are shown in Figure 5.3 and described below:

- `data.stations` — this table contains information about every station in GB, with the station CRS code as the primary key. The table was initially populated from the NaPTAN database, and additional columns were added to store information related to station services and facilities.
- `latis.survey_val` — This table contains the validated revealed preference survey data (as described in Chapter 4). Each row in this table corresponds to an individual survey response, with columns corresponding to the survey questions. Additional columns specific to each survey response were added to this table.
- `latis.nearest_30_stations` — This table holds the 30 nearest stations to each unique origin (`originlatlong`) in `latis.survey_val`, calculated using the euclidean distance. For each origin the table has 30 rows, each a potential alternative station. This table was used to rank the stations, for example by road distance, and to generate a choice set for each survey response. The process used to populate this table and create the choice sets is described in Chapter 6.

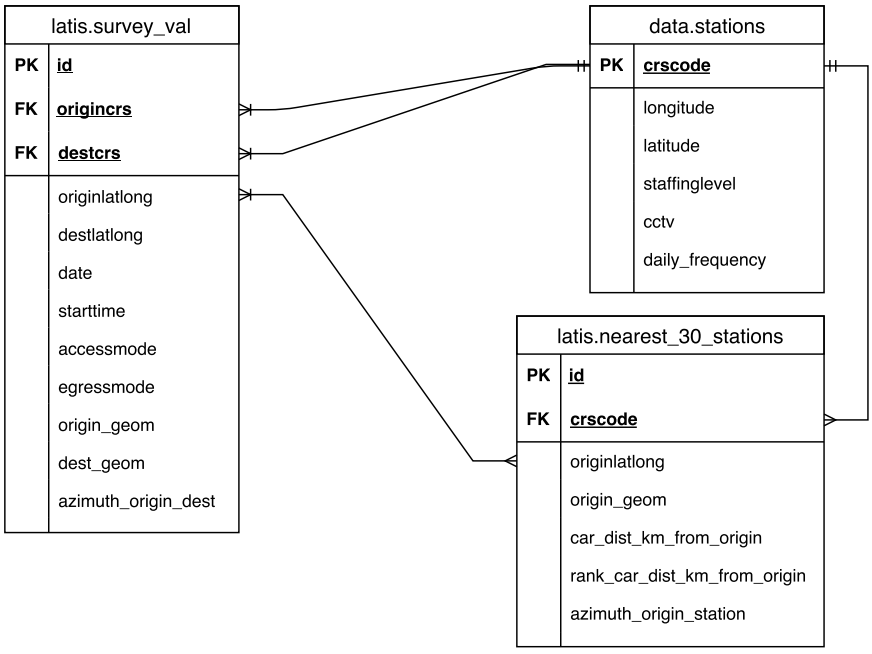


FIGURE 5.3: Key tables in the PostgreSQL database schema, showing primary and foreign keys and example columns.

5.3.2 R software environment

The R software environment is the hub of the framework. A set of functions were developed to query the OTP API and process the JSON response. These include a connect function

(otpConnect), a trip distance function (otpTripDistance), a trip time function (otpTripTime), and a function that returns an isochrone map in GeoJSON format (otpIsochrone). These functions form the beginnings of an API wrapper for OTP which could be released as an R package in the future. The R code for the set of functions can be found in Appendix A (R code segment A.1.) An example otpTripTime query and response is shown in Listing 5.1.

```
> otpTripTime(otpcon, from = '50.79877,-3.18689', to = '50.62158,-3.41228',
modes = 'rail,walk', date = '06-01-2015', time = '7:30pm', detail = TRUE)
$errorId
[1] "OK"
$itineraries
  start    end      duration  walkTime  transitTime waitingTime  transfers
1 20:12:12 22:07:01 114.82    8.78      49          57.03        1
```

LISTING 5.1: Example of an otpTripTime query to the OTP API and the parsed response

R is able to read from and write to the database by sending queries using the RPostgreSQL package (Conway et al., 2016). The steps used in a typical R script to populate a database table, for example the road distance between the trip origin and each alternative station (`car_dist_km_from_origin` in `latis.nearest_30_stations`), are illustrated in Figure 5.4. Data were also pulled from multiple database tables to populate the choice model datasets with the alternatives for each observation and associated predictor variables. The choice model datasets were stored as R data frames, and once complete were exported as CSV files in the format required by the choice modelling software.

5.3.3 External data sources

Data from external web services can be accommodated in the framework through appropriate API wrappers or feed parsers. Example feeds that were incorporated include the BR Fares website (BR Fares Ltd, 2016), and the National Rail Enquiries (NRE) Knowledgebase XML feeds (National Rail Enquiries, 2016). Further details are provided in Section 5.4.

5.3.4 Benefits of the framework

Developing and adopting the processing framework provided significant benefits to the research project. These extended beyond its initial use to generate the predictor variables, to enhance all aspects of the project. Key benefits included:

- It was efficient and reduced the opportunity for errors to arise, as the source data only had to be stored and maintained in a single location (a database table).
- All data processing and analysis was carried out using R scripts, providing an extremely detailed record of every step that was completed.

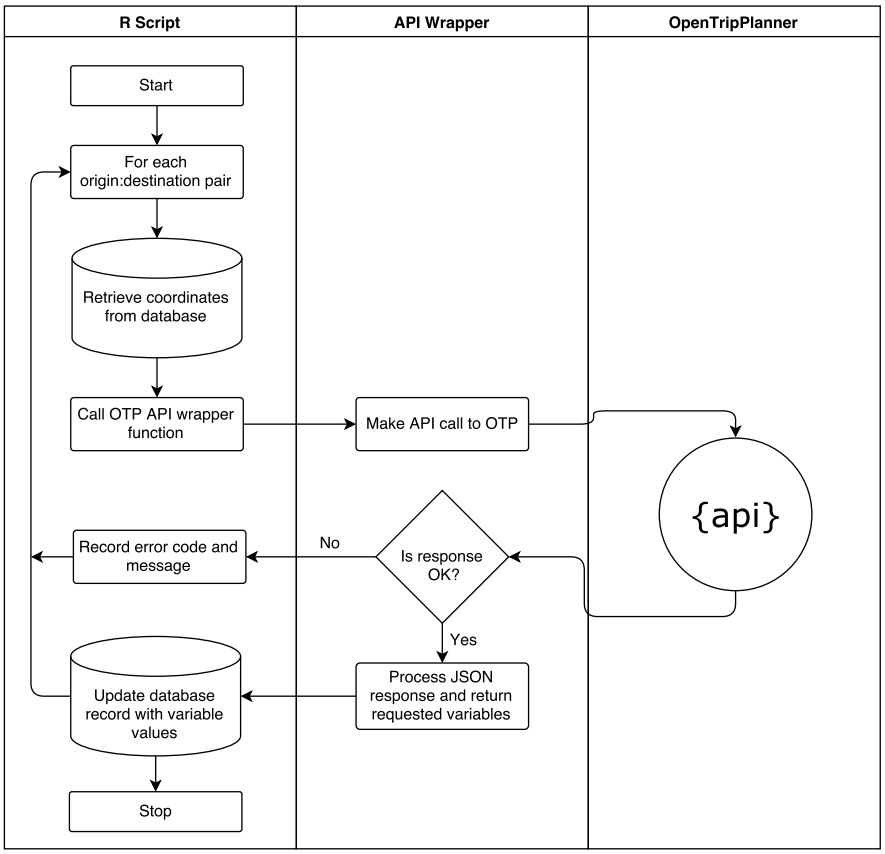


FIGURE 5.4: The steps in a typical R script to populate the choice model dataset using data from OTP.

- The approach gave a level of reproducibility that would not otherwise have been possible, which proved invaluable during the research process. For example, it enabled the choice model datasets to be readily regenerated based on different approaches or assumptions and with new or modified variables.
- It provided hugely enhanced analytical capabilities, both spatial and non-spatial. Some examples include: using SQL window functions to perform a calculation on grouped records, such as ranking alternatives by distance for each origin; calculating the difference in bearing variable described in Section 5.4.1 using the PostGIS `ST_Azimuth` function; and using SQL Procedural Language to calculate the accessibility term (see Section 7.5.3.1).

5.4 Deriving the predictor variables

Having established the processing framework, the next stage was to generate the predictor variables that were to be tested during calibration of the station choice models. Detailed information on how the variables were obtained is given in the sections that follow. Where

non-default values were used for OTP API parameters in the query request, these are identified and justified[4].

5.4.1 Access journey

Various measures of the access journey were obtained by querying the OTP API. These included the distance in km using drive mode, and the access time in minutes by the reported access mode. In a very small number of cases OTP reported that the trip was not possible by car. This was due to the nearest road to the origin postcode centroid not being available for car use, such as a pedestrianised street. In these cases the start point was manually adjusted in the OTP web interface until a valid route was returned, and the new coordinates for that origin were stored in a lookup table.

To generate journey data for access by bus (and also the Glasgow subway) the Scottish and Welsh components of the TNDS generated on 9 June 2015 were incorporated into OTP. As archived versions of TNDS are not publicly available, all bus and subway journeys were assumed to take place in the week beginning 8 June 2015. To take account of varying service levels throughout the week, the actual day of week of travel was calculated for each observation in the dataset, and this was matched to the same day in the week beginning 8 June 2015. The desired time to arrive at the origin station was set to the recorded train time[5], and the following three trip planner parameters were set to non-default values:

- The `maxWalkDistance` was set to 1,600 m (default: unlimited), notionally allowing 800 m (half a mile, or approximately a ten-minute walk) at both ends of the bus trip. This is a soft limit. If no solution is available that respects this limit, the route planner will increase it.
- The `walkReluctance` parameter, which is a multiplier that indicates the extent to which sitting on a bus is preferred over walking, was increased from the default value (2) to 5. This was based on experience requesting itineraries using the web interface, and ensured a more realistic balance between the walk and bus components of the trip. If set too low, the amount of walking may be excessive for someone who has chosen to travel by bus; and if set too high the planner will try to limit walking to the bare minimum, introducing unnecessary transfers and associated waiting time to avoid even a modest walk to/from the boarding or alighting bus stop.
- The `minTransferTime` was set to 600 seconds (10 minutes). This is the minimum time the planner will allow for a transfer between bus services.

[4]The full API documentation for OTP is available at: `http://dev.opentripplanner.org/apidoc/`

[5]For the WG dataset the scheduled station departure time is recorded, whilst for the LATIS dataset the start time of the particular service is recorded.

Two additional variables related to the access journey were generated. First, a 'nearest station' dummy variable which indicates whether or not a station in an individual's choice set is the closest station (this was determined based on both drive distance and mode-specific access time). Second, a 'bearing difference' variable, which gives the difference in bearing of origin:origin station and origin:destination in degrees (see Figure 5.5). This will identify whether passengers prefer to choose a departure station that is broadly in the same direction as their final destination. It was calculated using the PostGIS `ST_Azimuth` function (The PostGIS Development Group, 2018), which gives the angle measured in degrees referenced from the vertical (North) of point A to point B. This was calculated for origin:origin station (for all stations in an observation's choice set) and origin:destination (for the observation's reported trip). The absolute difference between the two azimuth angles was then calculated for each station in the choice set, using the equation: $180 - \text{abs}(\text{abs}(azimuth_{os} - azimuth_{od}) - 180)$.

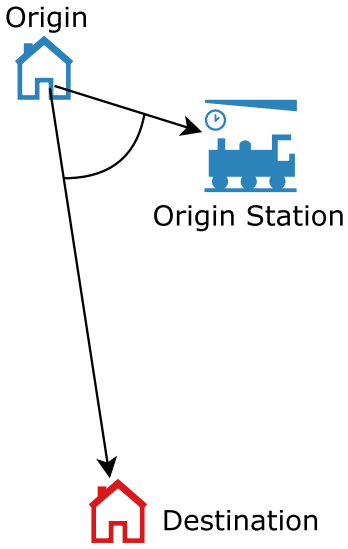


FIGURE 5.5: Difference in bearing (degrees) origin:origin station and origin:destination.

5.4.2 Station facilities and service frequency

Information on a range of potential facilities available at railway stations was obtained from the NRE Stations XML feed, which forms part of the NRE Knowledgebase. This was downloaded for every station in the UK and then parsed in R, primarily making use of the `xpathSApply()` function from the XML package (Lang & the CRAN Team, 2017). The variables recorded were: free car park (y/n), car park spaces (number), station CCTV (y/n), ticket machine (y/n), waiting room (y/n), station buffet (y/n), toilets (y/n), cycle spaces (number), cycle storage (y/n), cycle shelter (y/n), cycle CCTV (y/n), bus interchange (y/n), taxi rank (y/n), car hire (y/n), cycle hire (y/n), metro services (y/n), and staffing level (unstaffed, part-time, full-time). The script used to parse the feed is provided in Appendix A (R code segment A.2).

To generate service frequencies the GTFS feed for GB rail services dated 23 November 2013[6] was downloaded from the TransitFeeds archive (TransitFeeds, 2017) and converted into a PostgreSQL database, with each component CSV file becoming a separate database table. A SQL query was then used to count the number of daily services scheduled at each station on Monday 25 November 2013 (see PostgreSQL segment B.2 in Appendix B). All trains calling at the station were counted, even if passengers could alight only. This was considered appropriate, as terminal stations would otherwise appear to be served by fewer trains than intermediate stations on the same line.

5.4.3 Train journey

Two GTFS feeds for GB rail services dated 17 March 2014 and 4 April 2015 were downloaded from the TransitFeeds archive and incorporated into separate OTP graphs to cover the survey period for both the WG and LATIS datasets. In addition, to allow London transfers, a GTFS feed containing London Underground and Docklands Light Railway services was created from downloaded Transport for London Journey Planner timetables[7]. These timetables were provided in TransXChange format and were converted to GTFS, with bus and river services excluded, using Visography TRACC.

A single train journey itinerary from origin station to the observed destination station for the date of each trip was obtained by querying the OTP API. Walk mode was also permitted, primarily to enable an alternative destination station, for example on a different line, to be selected by the planner, with a walk to the observed destination station.[8] The `minTransferTime` parameter was set to 320 seconds (6 minutes), corresponding to the typical suggested connection time for a medium interchange station. The desired trip start time was set to the recorded train time.[9] The variables retrieved for testing in the choice models were the journey duration and its separate components: on-train time and waiting time.

Fare data was obtained using the independent BR Fares web service API (BR Fares Ltd, 2016). An API lookup was made for each unique origin:destination station pair in the choice model datasets. This returned all possibles fares between the two stations, in JSON format. Adult

[6]This was an oversight. In early modelling the GTFS feed dated 25 April 2015 had been used, which better corresponded to the date the surveys were carried out. However, only trains where passengers could board were counted, causing terminal stations to appear to have fewer trains than intermediate stations on the same line. As the frequency figure for *all* trains had already been calculated for the 23 November 2013 GTFS feed, this was inadvertently used. As it is unlikely that any major changes in station service frequencies occurred between these two dates, the impact is considered to be minimal.

[7]Available from: `https://api-portal.tfl.gov.uk/docs`

[8]Initially it was planned to request routes from each origin station to the ultimate destination. However, this is problematic as in some cases the egress mode is by car or coach with the final destination a considerable distance from the observed destination station, and the route planner will suggest a much longer rail journey to a station that is nearer the ultimate destination.

[9]See Footnote 5

walk-up fares were then selected, and from this subset the cheapest anytime return fare[10] and the cheapest off-peak return fare[11] were extracted. The fare variable was then populated dependent on the recorded train time, with the anytime return fare used for train times before 9 a.m., and the off-peak fare used for train times after 9 a.m. The code used to parse the API JSON response is included in Appendix A (R code segment A.3).

5.4.4 Land use and built environment

To investigate the effect of land use on station choice, a land-use mix measure was generated using the Ordnance Survey Points of Interest (POI) dataset, obtained from the EDINA Digimap service. As it was not possible to obtain the POI dataset for the entirety of the survey regions, due to a maximum area restriction, the POIs within an 800 m^2 buffer of each station were downloaded and merged into a database table. The number of points of interest for each of the nine top-level classifications[12] within a Euclidean distance of 400 m (about a five-minute walk) of each station were then summed using a spatial query. The Herfindahl-Hirschman Index (HHI) was then calculated for each station. The HHI indicates the extent to which one land use type dominates in an area, and is calculated by squaring the percentage market share of each classification, and then summing the squares:

$$HHI = \sum_{i=1}^{K} (P_i \times 100)^2 \tag{5.1}$$

where P_i is the proportion of land-use type i, and K is the number of land-use types (in this case the nine top-level classifications) (Song & Rodríguez, n.d.). With nine classifications the index can range from 1,111.$\overline{11}$, where each land use type is equally represented in the area, to 10,000 where only a single land use type is present.

5.4.5 Socio-economic variables

In discrete choice models each attribute must vary across the alternatives in a choice set. While this is usually the case for attributes of the alternatives, attributes that relate to the decision maker, such as socio-demographic variables, will be the same for each alternative. There are two methods that allow socio-demographic variables to be used in choice models. In the first, the variable is interacted (in some justifiable way) with an attribute that does vary across alternatives, for example a cost variable could be divided by income. The second, which can only be used if each alternative has a separate utility function, requires one of

[10] If available, the anytime day return fare was used (ticket type code: SDR), otherwise the lowest fare with code SOR, GOR or GTR was selected

[11] If available the cheap day return fare was used (ticket type code: CDR), otherwise the lowest fare with code SVR, BFR, G2R or SMG was selected.

[12] The nine top-level classifications are: accommodation, eating and drinking; commercial services; attractions; sport and entertainment; education and health; public infrastructure; manufacturing and production; retail; and transport.

the parameters to be normalised to zero by excluding it from one of the utility functions. For example, if the choice was between travelling by bus or car, income could be excluded from the car utility function. The estimated parameter would then be interpreted as the effect of income on utility of bus *compared to car* (see Train (2009, pp. 21–23) for a more detailed discussion). An alternative solution would be to calibrate entirely separate models for particular socio-demographic segments, such as different age groups or levels of car ownership.

The LATIS and WG surveys did include some supplementary socio-demographic questions, for example sex, age (WG only), and household car ownership (LATIS) or car availability (WG). However, as the station choice models would only define a single utility function (representing the utility of choosing a station), the variables would either have to be interacted with an attribute of the alternatives, or separate segmented models would need to be estimated. Another potential issue was that any variables included in the choice models would need to be available at the same spatial resolution when the aggregate models were calibrated or applied. However, the socio-demographic UK census data is generally not available at the unit postcode level, which was the zonal spatial resolution chosen for this research. Furthermore, even if a variable such as level of car ownership was available at postcode level, it would still represent an average for the postcode and introduce the problem of ecological fallacy. Given the absence of a justifiable interaction variable, a desire to maximise the information available to the station choice models by avoiding segmentation, and concerns about subsequent model application, it was decided to only include attributes of the alternatives in the station choice models.

5.5 Conclusions

This chapter has described how a range of potential station choice predictor variables have been derived in a reproducible manner from a variety of data sources, supported by a processing framework built around open source tools and accompanying code. The overriding approach has been to obtain variables that better represent the information that would have been available to each survey respondent. The OTP trip planner, and set of R functions written to query the API and parse the response, has enabled mode-specific station access journeys and several components of the train leg to be generated. These have been enhanced further by using the transit timetables that were in operation when the passenger surveys took place, and by matching to the appropriate day of week and trip time. Further code development has enabled station facility and fare information to be obtained directly from API services; and importing the rail timetable data into an SQL database has facilitated powerful relational queries, such as calculating daily station frequency. The OTP API functions have the potential to be developed further into an R package. This could be of enormous benefit to researchers across disciplines, enabling them to query their own bespoke trip planner.

The station choice predictor variables, along with the observed station choice data that was described in the previous chapter, can now be brought together to calibrate models that can predict station choice. The development of these models is the subject of the next chapter.

Chapter 6

Station choice models

6.1 Introduction

This chapter is concerned with the development of station choice models that have the potential to be incorporated into aggregate rail demand models. It begins by explaining which model forms were chosen and why (Section 6.2). The process of defining the choice sets is then outlined, and descriptive statistics for the two datasets are presented and discussed (Section 6.3). The calibration of MNL models (Section 6.4) and random parameter (mixed) logit (RPL) models (Section 6.5) is then described; followed by an appraisal of the models, considering their predictive performance and transferability (Section 6.6.1). The development of a station choice model specifically intended to be incorporated into a national-scale trip end model is then described (Section 6.7), before the chapter closes by summarising the outcomes of the model development process and drawing some conclusions (Section 6.8).

6.2 Choosing the model form

It was decided to initially develop a range of MNL models, as this model form has been widely used to model station choice in prior work, and it made sense to begin model development with this relatively simple closed-form model. The calibration of the MNL models is described in Section 6.4. The other commonly adopted approach in previous research has been to model combined access mode and station choice using NL, with access mode at the upper level and station choice at the lower level. However, there are several theoretical and practical issues with this approach, some of which were identified in Chapter 3. The main issues are summarised below:

- The NL model is intended to address the IIA problem and the proportional substitution behaviour that follows from it. It is far from clear how placing the same stations

into each access mode nest can be theoretically justified[1]. Crucially, this nesting structure fails to address unobserved spatial correlation (alternative nesting structures are considered in Section 6.2.1 below).

- As the same alternatives are not allowed to be in multiple nests, it is necessary to pair each alternative with an access mode (e.g. station1_car, station1_transit, station1_walk, station1_cycle), creating a much larger choice set.
- The NL model requires a universal choice set to be specified. In the case of the LATIS dataset with 328 unique stations, the universal choice set with four access modes would potentially contain 1,312 alternatives, exceeding the maximum of 500 allowed by the NLOGIT 5 software package that was chosen for this project (Econometric Software Inc, 2012). A universal choice set is also inappropriate for a study of station choice, where the choices available to individuals will depend upon their location.
- While the passenger surveys asked some questions that would be particularly important to include in the utility function for access mode choice, for example car ownership and/or availability, such data would not be available at the necessary spatial resolution when the station choice models were applied.

In view of these issues, it was decided not to pursue this model form. As MNL is unable to account for individual taste variation, it was decided to examine whether the MNL models could be improved upon by using the random parameter specification of the ML model, an open-form model that requires the probabilities to be calculated using simulation techniques. The calibration of these RPL models is described in Section 6.5.

6.2.1 Addressing spatial correlation

A weakness of almost all previous station choice research studies, is their failure to address the issue of spatial correlation between alternatives. This is a particular issue for models that will be used to predict demand for new stations, as it impacts their ability to represent realistic patterns of passenger abstraction from existing stations. In an MNL model, introducing a new station will reduce the probability of all existing stations in the choice set by the same percentage, when in reality it would be expected to exert a greater influence on stations closer to it. Several possible methods to address the issue were identified in Section 3.3.3, including: nested logit; generalized nested logit; specially formulated spatial choice models; and the introduction of an accessibility term. The potential of these four approaches was considered in the context of this project, and the findings are summarised in the sections that follow.

[1] This is a view shared by Professor William Greene, Professor of Economics at New York University Stern School of Business and developer of NLOGIT, who cast doubt on the validity of this approach (personal communication, 7 October, 2015).

6.2.1.1 Nested logit

The NL model has the potential to address the problem of proportional substitution, but only if appropriate groupings of stations can be defined. Although the IIA property is relaxed between nests, so that the ratio of probabilities of two alternatives in different nests can vary, IIA still holds for each nest and proportional substitution will occur. It is therefore necessary to define groups of stations where this would be appropriate, and a mechanism for objectively achieving this using a clustering algorithm was considered. The Partitioning Around Medoids (PAM) algorithm was chosen, which is available as part of the 'cluster' R Package (Maechler et al., 2017). This was preferred over K-means (which uses Euclidean distances), as the cluster centres (medoids) are data points and a dissimilarity matrix can be supplied (Mirkes, 2011). In this case the dissimilarity matrix was defined as the road distance between each station pair. This was created for the WG dataset, by obtaining the walk distance between each unique pair of stations in the dataset from OTP. As NLOGIT 5 allows a maximum of 25 nests to be specified, this was set as the number of required clusters. Once the clusters had been generated, coordinate data for the stations was attached and the clusters were plotted in QGIS. A selection of the clusters are shown in Figure 6.1, with the stations in each cluster identified by colour and cluster number. It is apparent that the clusters are large, in several cases larger than the anticipated individual choice set size of 10 stations (e.g. cluster 14), and the stations within them are widely spread geographically (e.g. cluster 21). Even when the number of clusters was increased to 60, as shown in Figure 6.2, large clusters of geographically spread stations remain. In addition, some stations are nearer to a station in another cluster than they are to stations within their own cluster, for example where cluster 59 meets cluster 27. Based on this analysis it was decided not to pursue NL as a method of addressing spatial correlation. As well as the clusters containing too many geographically dispersed stations to be useful for capturing spatial competition effects, there would be a more general problem with the transferability of such models.

6.2.1.2 Cross-nested logit

The potential to use CNL to address spatial correlation between stations was considered with particular reference to the approach adopted by Lythgoe et al. (2004), which was discussed in Section 3.3.3.1. This approach allowed for a natural grouping of stations within a nest structure, as shown in Figure 3.7, where the composite utility of travelling by rail from an origin station zone to a destination station via any of the (up to 15) competing stations is at the upper level. In the case of the models to be calibrated for this thesis there is no upper level above the individual station choice by which to group the alternatives. Therefore, to adopt a similar approach to Lythgoe et al. would imply each station in the universal choice set being paired with each alternative[2]. For the LATIS dataset with 328 unique stations it would be necessary to define 107,256 nests in the model. This is unlikely to be feasible,

[2]Note that order is important, as in the model nest $\left[\begin{smallmatrix} i \\ k \end{smallmatrix}\right]$ is distinct from nest $\left[\begin{smallmatrix} k \\ i \end{smallmatrix}\right]$.

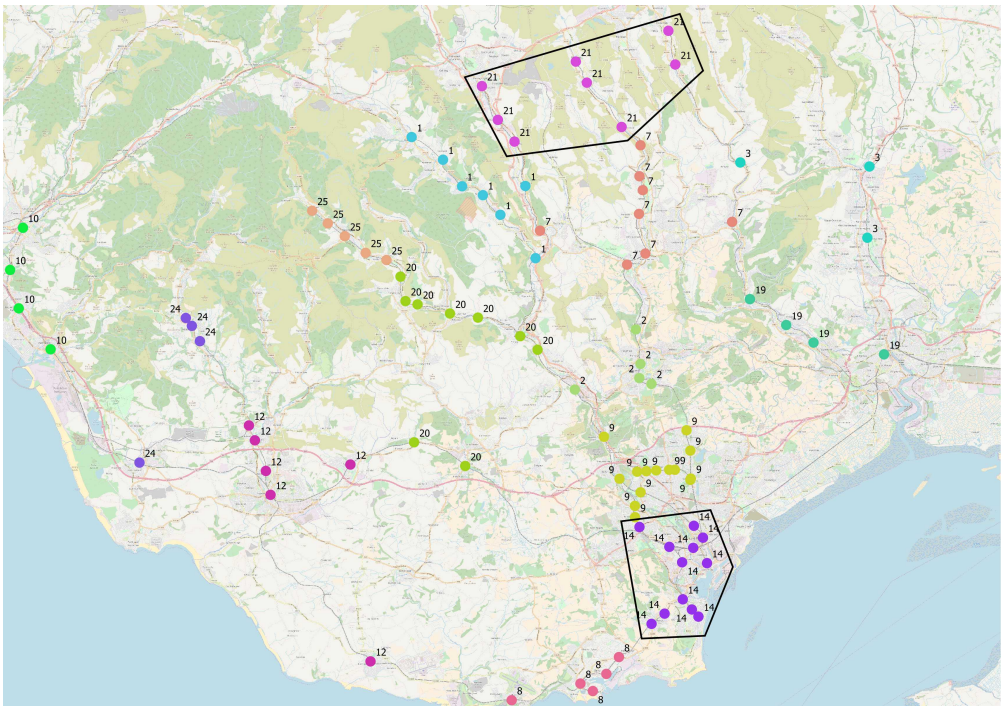


FIGURE 6.1: Stations in the WG dataset clustered using the PAM algorithm — 25 clusters (not all shown).

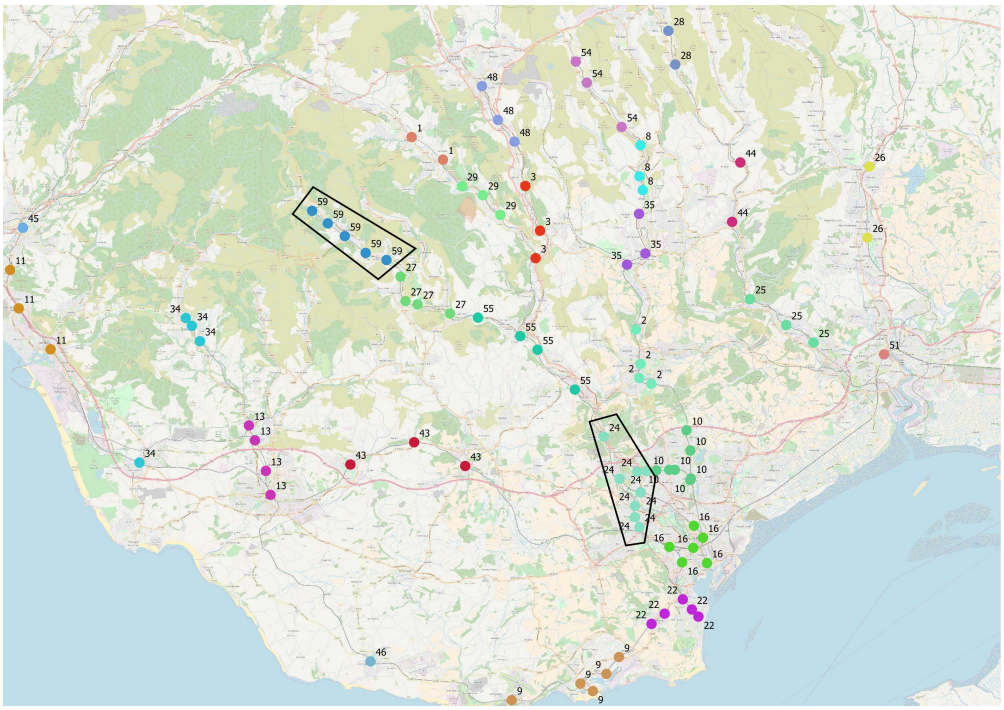


FIGURE 6.2: Stations in the WG dataset clustered using the PAM algorithm — 60 clusters (not all shown).

both in terms of defining the model (for example NLOGIT5 allows a maximum of 25 nests) and model calibration. In addition, it has already been noted that a universal choice set would not be appropriate for station choice, as each decision maker will clearly only consider a very limited subset of these stations. This issue could be overcome by considering each individual's choice set to be the upper level grouping, with the cross-nesting of station pairs only occurring within each choice set. However, for the LATIS model with 9,367 choice sets and assuming 10 stations in each (producing 90 nests), an infeasible total of 843,030 nests would need to be specified.

A potential solution to this enormous escalation in the number of nests would be to rank the stations by distance from the trip origin for each choice set, with the choice becoming a station of a particular rank, rather than a specific station. In this way the number of alternatives in the model could be reduced to just 10, with 90 cross-nested station pairs. However, a major issue with this approach is that a single set of allocation and dissimilarity parameters could not adequately represent the degree of variability in unobserved independence or correlation between pairs of stations of specific rank. For example, if a trip origin is in an area of high station density then the expected pattern of allocation would be very different from that in an area of very low station density. This is illustrated in Figure 6.3, where the hypothetical allocation of the nearest station (R1) to four other stations is shown. In choice set A the stations are close together and R1 is apportioned equally to the nests of the other stations. In choice set B the stations are geographically more disperse, and a much larger proportion of R1 is allocated to R2's nest than to the nests of the other stations. It might be possible to address this issue by part calculating the allocation parameter prior to model estimation. This was the approach adopted by Lythgoe et al., where a logit probability was used to calculate the allocations based on the road distance between station pairs (see Equation 3.19), with a parameter θ to be estimated (effectively a spread parameter). However, the ability to part calculate allocation and/or dissimilarity parameters was not available in the NLOGIT5 software selected for this research. It was also noted that Lythgoe et al. were unable to estimate θ but instead tested the model with 'the parameter set to different values'.

An additional consideration in the selection of a model for the station choice element of this research was the practicability of incorporating it into a trip end model that was to be calibrated for all of GB at a high zonal spatial resolution. This integration would require calculating choice probabilities for some 1.5 million postcodes, and a model form that imposed substantial additional overhead on that calculation, for example by evaluating it over at least 90 nests per postcode, is very unlikely to be practicable, either for calibration or the subsequent application of the model to generate demand forecasts for new stations.

In view of the range of issues associated with implementing CNL discussed above, it was decided not to pursue this model form and consideration turned to models specifically developed to address spatial correlation. These are discussed in the next two sections, but it is useful to note at this point that the work of Sener et al. (2011) and Weiss and Habib (2016) (referred to in Section 6.2.1.3) was carried out many years after development of the

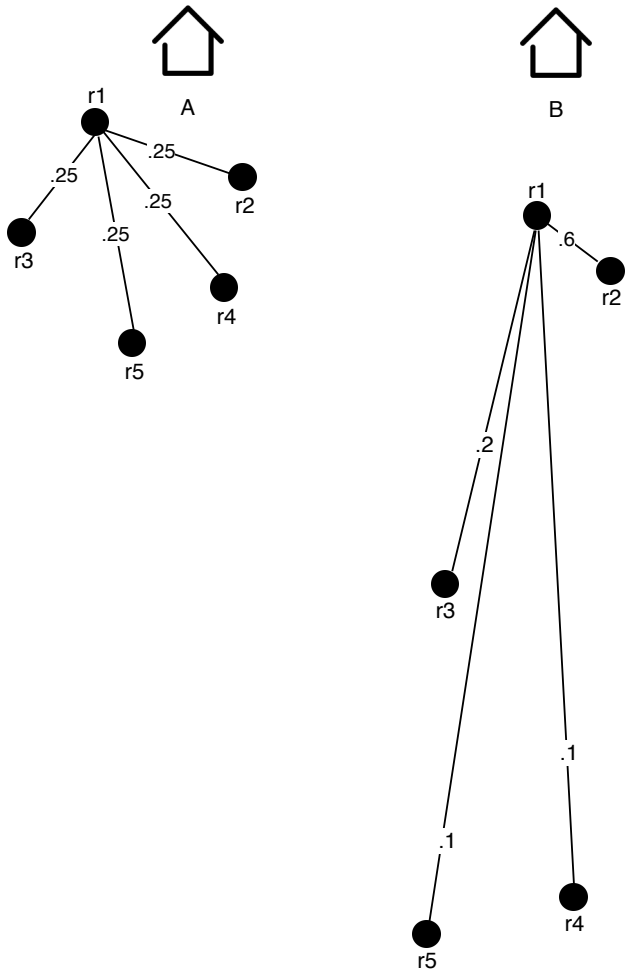


FIGURE 6.3: Example choice sets (A and B) where stations are ranked by distance from the trip origin (r1 – r5), showing hypothetical allocation of the first ranked station to each nest containing stations ranked 2 – 4.

CNL and GNL models. This suggests that the latter models may not be the most appropriate solution for addressing the issue of spatial choice in models of this type (i.e. flat MNL choice form with no obvious upper level nesting).

6.2.1.3 Spatial choice models

Two models that address spatial correlation and appear promising in the context of station choice were identified in Chapter 3 (Section 3.3.3). These were the GSCL model proposed by Sener et al. (2011); and the SWEC model proposed by Weiss and Habib (2016). Unfortunately, the functionality to run these models is not present in proprietary or open-source software packages. It would therefore be necessary to define the likelihood function programmatically, for example using the GAUSS matrix programming language, which would require advanced knowledge of econometrics. Attempts were made to contact Ipek Sener, with the view to

obtaining the code to run the GSCL model, but no response was forthcoming. In view of these obstacles, the decision was made not to pursue these models further.

6.2.1.4 Introducing an accessibility term

The fourth option identified in Chapter 3 to deal with spatial correlation, is introducing an accessibility term into the MNL model. To assess the potential of this approach, it was decided to test the following form of the accessibility term, as used in Fotheringham's CDM:

$$A_{ni} = \left(\frac{1}{M-1} \sum_{\substack{k \\ k \neq j}} \frac{W_k}{d_{jk}} \right)^{\theta}, \tag{6.1}$$

where M is the total number of stations in the choice set for individual n at origin i, W is a weight, d is the distance from station j to station k, and θ is a parameter to be estimated. As A increases a station is closer to more 'attractive' stations. The weight was defined as the total number of station entries and exits in 2014/15, and the expectation was for $\theta < 0$, indicating that a station has a lower utility, and is therefore less likely to be chosen, the nearer it is on average to more heavily used stations. Fotheringham states that the CDM can be derived, and under certain circumstances be consistent with random utility theory, simply by including the accessibility term in the utility function (Fotheringham, 1986), and that was the approach adopted, with the suggested logarithmic transformation of the term added to the models.

6.3 Choice set definition

Having decided on the model forms to be developed, and how the issue of spatial correlation was to be examined, the next step was to define the choice sets for each observation in the datasets. It is infeasible that someone choosing an origin departure station would consider the entire universal choice set of some 2,500 stations in GB. This issue of defining the individual choice set for spatial decisions where the universal choice set is often very large is well-recognised. For example, in a review of choice set formation in destination choice models, Thill (1992) argues:

> On the other hand, the set of possible alternatives is typically large for spatial decisions, so that it can hardly be argued that the individual is able to evaluate it all. More realistically, the individual considers only a portion of the universal set. (p. 364)

Erroneously defining the choice set as the universal set could result in significant model mis-specification, as the choice model will assign positive probabilities to *all* alternatives, irrespective of whether they are in the individual's true choice set, potentially resulting in biased parameters and/or prediction errors (Pagliara & Timmermans, 2009; Thill, 1992). A method was therefore required to reduce the universal choice set to a realistic and feasible set of alternatives for each individual. As discussed in Section 3.4.3, there are two main approaches to this choice set generation process, either deterministic or probabilistic (stochastic). In the deterministic approach the choice sets are defined exogenously by the researcher based on some constraint(s). This approach has been criticised for relying on assumptions made by the researcher on the basis of arbitrary criteria and therefore involving uncertainty (for example, see Cantillo and Ortúzar (2005) and Zolfaghari, Sivakumar, and Polak (2013)). The stochastic approach is usually based around the two stage model first suggested by Manski (as cited in Pagliara and Timmermans (2009)), which takes the following form:

$$P_d^i = \sum_{C \in G} P^i(d \mid C) \cdot P^i(C \mid G), \tag{6.2}$$

where P_d^i is the probability that individual i chooses alternative d; $P^i(d \mid C)$ is the probability that individual i chooses d given choice set C; and $P^i(C \mid G)$ is the probability (to be modelled) that the choice set of individual i is C; and G is a set of all non-empty subsets of the universal choice set M. A major problem with this general form is that the sum is across every possible combination of alternatives. The number of choice sets increases exponentially with the number of alternatives ($G = 2^M - 1$) and the model is only practicable when the number of alternatives is small (perhaps 6 or less). It would be virtually impossible to apply this type of model with 2,500 alternatives in the universal choice set. In order to create a tractable model many variations to the choice set generation stage have been proposed that impose constraints to restrict the choice sets and sets of choice sets (Horni, Charypar, & Axhausen, 2010). A comprehensive review of these models is provided by Pagliara and Timmermans (2009). They can be difficult and complex to estimate, and most cannot be estimated using standard software. Furthermore, like the deterministic approach, these rely on exogenous information for the choice set formation, as noted by Pagliara and Timmermans (2009):

> Even though the inclusion of latent stochastic thresholds and the simultaneous estimation of thresholds and utility functions represents an important step forward in discrete choice analysis, forecasting results still depend on the researchers' specification of the choice set. What seems lacking is a convincing process model that probably needs to be developed with a particular type of spatial choice behavior in mind. (p. 193)

An important objective of this research project, as outlined in Chapter 1, was to use the station choice model to define probabilistic station catchments at a high spatial resolution for incorporation into a trip end model. This would require defining a choice set for some

1.5 million GB postcodes. With this in mind a pragmatic approach was needed to define the individual choice sets. A highly complex stochastic method, untested in the field of station choice modelling and not based on a robust process model for station choice behaviour (which does not exist), was not thought to be practicable or appropriate. The methods used to define the choice sets in prior station choice research, which were reviewed in Section 3.4.3, were all based on a deterministic approach, with the nearest n stations to the trip origin the method most commonly adopted. It was therefore decided to adopt a similar approach in this study. This was considered preferable to using a distance- or time-based threshold which would produce widely varying choice set sizes depending on station density. For example, a 60-minute drive time threshold applied to a postcode origin in the Greater London area could produce an infeasibly large choice set, potentially containing hundreds of stations; while in a rural area the same threshold may contain only one or two stations, potentially excluding stations that were evaluated by travellers in reality. The nearest n method is intuitively more attractive as it implies that an individual's geographical area of consideration will be smaller when there is a high density of stations and larger when station density is low. This is consistent with conceptual models of spatial choice behaviour where consumers develop a 'spatial information field' or 'mental map' of the available facilities to satiate their demands (for example, see Hanson (1977); Potter (1979); Smith (1976)). In areas with low station density, the spatial information field may need to be wider to include more distant stations to meet the traveller's needs. The decision on the value of n was based on findings from an initial pilot study using a smaller survey dataset (Young & Blainey, 2016). This analysis found that the nearest 10 stations accounted for 99% of observed choice, and this was chosen as the criteria for generating the choice sets.

Both Thill (1992) and Pagliara and Timmermans (2009) make the observation that the consequences of a mis-specified choice set are only theoretical and the impact can be minimised by a well-specified model. If an alternative that was not evaluated is included in a choice set but is assigned a very low probability, close to zero, then the impact might be very small. The example given by Thill (1992) is a store located a long distance from the decision maker and with no characteristics that make it more attractive than other closer stores. As this store is unlikely to be chosen then its inclusion in the choice set 'is of no consequence either for predicted choice probabilities or for parameter estimates'. A similar point is made by Bierlaire, Hurtubia, and Flötteröd (2010), who observe that 'the more an alternative is dominated, the less important it is to know if it really belongs to the choice set'. In the context of station choice there are situations where you might expect one station to be dominant, for example the choice set for a postcode next to a major station with superior service levels and facilities; or a postcode in a market town located close to the only station and where the other stations in the choice set are in neighbouring towns on the same line and with similar services and facilities. In situations like this the dominant station was found to have an extremely high probability, as shown in the example in Table 6.1. This gave confidence that the model was well specified and that any bias would be minimised.

Station	Probability
Swansea	0.9993002895
Port Talbot Parkway	0.0002015849
Gowerton	0.0001988341
Neath	0.0001913671
Llansamlet	0.0000871365
Skewen	0.0000138374
Baglan	0.0000030092
Briton Ferry	0.0000029273
Pontarddulais	0.0000005616
Bynea	0.0000004525

TABLE 6.1: Predicted station choice probabilities for postcode SA1 5DZ, located close to Swansea railway station.

To generate the choice sets for the WG and LATIS datasets, a database table was first populated with the nearest 30 stations to each unique origin, based on Euclidean distance using the efficient PostGIS indexed nearest neighbour query (Ramsey, 2011). Any new stations that were not open during the relevant survey periods were excluded from the universal set of available stations. For each origin:station pair the drive distance was obtained using an API call to OTP and the 30 stations were then ranked by drive distance for each origin using a window function, enabling the nearest ten to be identified. These choice sets were found to account for 92% and 95% of observed choice in the LATIS and WG datasets respectively.

It was noted in Section 4.6.2 that a small but not insignificant proportion of survey respondents chose a station that was outside of their nearest ten, and even including the nearest 30 stations did not account for all observed choice. One likely explanation is that passengers sometimes choose to board at a major city-centre station, and reject many small- or medium-sized stations that are closer to their trip origin. It was therefore decided to try and improve the choice sets by ensuring the nearest major station to each origin was included. Although a strict criteria was not applied to select these 'major' stations, the starting point was those stations in Scotland or Wales with more than 50,000 annual interchanges. Suburban stations were excluded, and several stations in England that might realistically be chosen from origins in Wales and Scotland were added. The final list of stations identified as 'major' were: Aberdeen, Aberystwyth, Bridgend, Bangor (Gwynedd), Carlisle, Cardiff Central, Cardiff Queen Street, Carmarthen, Chester, Dundee, Edinburgh, Glasgow Central, Glasgow Queen Street, Hereford, Haymarket, Inverness, Llandudno Junction, Newcastle, Newport (S Wales), Perth, Shrewsbury, Stirling, Swansea, and Wrexham General. In the case of Glasgow, Edinburgh and Cardiff, the two main stations in these cities were included in the choice set if either of them was the nearest major station to the origin. By including the nearest major station in the choice sets, the proportion of observed choice accounted for increased to 97% in both datasets.

If an alternative station was also the destination station of an observed trip, then it was removed from the choice set, as this would clearly not be a valid option. In addition, if Glasgow Central or Glasgow Queen Street was the observed destination, then both of these stations were removed from the choice set if present. Using either of these stations to get to the other would be illogical. This is not the case for Cardiff or Edinburgh where travel between the two main stations by rail would be a logical trip. Any observation where the chosen station was not present in the choice set was, by necessity, removed prior to model calibration.

6.3.1 Threshold-based adjustments

A feature of logit models is that an alternative can never have a probability of zero, and if an alternative has no realistic prospect of being chosen it can be excluded from the choice set (Train, 2009). For example, if an individual has chosen to walk to a station, then a cut-off distance could be defined, after which a station is no longer considered a feasible alternative; and if travelling to a station by bus, then the choice set could be restricted to those stations that can realistically be accessed by bus from the individual's trip origin. However, refining the choice sets in this manner assumes that each individual only considered a single access mode, the one that they used to access their chosen station. As this is unlikely to be a valid assumption in many cases, applying adjustments of this nature may not be appropriate unless choice of access mode is simultaneously modelled. However, in the case of access by bus, it was considered reasonable to assume that a car was not available for the station access journey. Therefore, where access to the chosen station was by bus (or Glasgow subway) alternatives were only retained in the choice set if a route by that mode was available, or if the trip planner suggested walking to the station instead.

During data validation any trips where the respondent said they walked to the station were removed from the datasets if the access journey would have taken over 60 minutes (see Section 4.5.1.1). However, the choice sets for the retained observations where access mode was walk were *not* restricted to stations within 60 minutes of the origin. This is for the reasons outlined above; it is possible that someone who chose to walk to a reasonably close station also considered driving to a more distant one. Furthermore, restricting the choice set to stations within a 60-minute walk of the trip origin would have resulted in some choice sets containing only a single alternative, and the affected observations could not have been included in the model calibration.

As it was intended to estimate some models using mode-specific access time parameters, the small number of observations where access mode was recorded as 'other' were removed prior to model calibration[3]. This ensured that identical choice sets could be used for all model calibrations, allowing models to be compared using measures of model fit (log likelihood,

[3]These were largely unspecified or modes for which the trip planner could not be used to generate the access time variable, for example 'boat' and 'ferry'.

adjusted rho-squared and AIC). A summary of the choice set composition for the two datasets is provided in Table 6.2.

Dataset	No. of choice situations	No. of cases	Average choice set size
WG	5680	59833	10.53
LATIS	9367	97838	10.44

TABLE 6.2: Summary of choice sets prepared for model calibration.

6.3.2 Descriptive statistics

Summary statistics for most of the model variables are provided for the two datasets in Tables 6.3 and 6.4[4]. The statistics are summarised for all the alternatives (cases) present within the dataset, and also for the chosen alternative in each choice set only. The mean of the boolean variables indicates the proportion of survey observations or cases where that variable was true. As parameters for the two car parking variables were only estimated against those observations that accessed the station by car (see Section 6.4.1.2), the summary statistics for these variables only relate to those observations. As there were only a few observations where bicycle or taxi was used to access the chosen station, no variables that specifically related to these modes (for example, cycle parking) were included in the models.

Correlation matrices for the two datasets, prepared using the R package 'corrplot' (Wei & Simko, 2017), are shown in Figures 6.4 and 6.5. The upper triangular matrix represents the Pearson correlation coefficient for each pair of variables using a shaded circle, where the area of the circle and the depth of shading is proportional to the size of the correlation coefficient. Purple shading indicates a positive correlation, and brown shading a negative correlation. The lower triangular matrix shows the actual correlation coefficient in percentage format (for reasons of clarity). Where the correlation coefficient is not significant at the 95% level, neither a circle nor coefficient is shown. The variables are ordered using the first principal component method.

The highest positive correlations occur between the station service, facility and staffing-level variables, with this effect more pronounced in the WG dataset. For example, there is a strong correlation between full-time staffing level and service frequency (WG: 0.88; LATIS: 0.83), and a moderate correlation between service frequency and the number of car parking spaces (WG: 0.68; LATIS: 0.53). In the WG dataset a station with a toilet is very likely to also have a waiting room (0.90), although interestingly there is a small negative correlation (−0.09) between these two variables in the LATIS dataset. The presence of correlations between these variables is to be expected, as they are all influenced by the 'size' of the station. Larger stations that serve more passengers will have a greater service frequency, and they are more

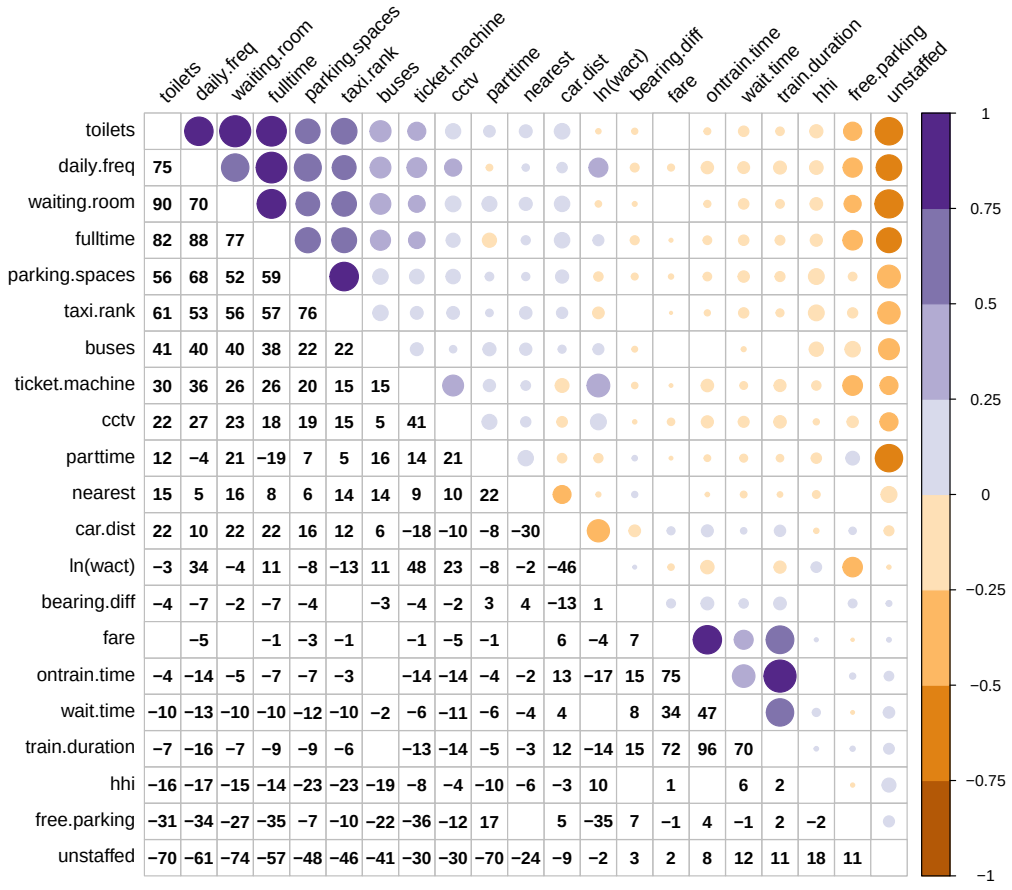
[4]For reasons of brevity the various measures of the access journey that were tested in the models, other than access distance by road, are not included in the summary statistics tables or correlation matrices.

Variable	All cases					Observed choice only				
	Mean	Std.Dev.	Minimum	Maximum	Number of cases	Mean	Std.Dev.	Minimum	Maximum	Number of cases
Nearest (by distance)	0.095		0	1	59833	0.705		0	1	5680
Car distance (km)	8.286	7.02	0.02	79.64	59833	2.736	3.86	0.02	79.64	5680
Full-time staff	0.133		0	1	59833	0.312		0	1	5680
Part-time staff	0.184		0	1	59833	0.484		0	1	5680
Unstaffed	0.683		0	1	59833	0.205		0	1	5680
Daily service frequency	143.459	174.34	7	686	59833	222.595	203.94	8	686	5680
CCTV	0.822		0	1	59833	0.973		0	1	5680
Ticket machine	0.700		0	1	59833	0.868		0	1	5680
Toilets	0.186		0	1	59833	0.501		0	1	5680
Waiting room	0.204		0	1	59833	0.532		0	1	5680
Bus interchange	0.443		0	1	59833	0.740		0	1	5680
Taxi-rank	0.115		0	1	59833	0.384		0	1	5680
HHI	2007.000	423.13	1422.22	7222.22	59833	1863.720	228.25	1422.22	3968.25	5680
Ln(wact)	11.705	1.37	5.18	14.11	59833	11.459	1.57	6.17	14.11	5680
Train leg duration (mins)	65.863	72.74	1	859	59833	54.259	63.19	2	512	5680
Waiting time (mins)	11.349	22.85	0	662	59833	5.467	12.34	0	311	5680
On-train time (mins)	54.456	58.94	1	632	59833	48.745	55.96	2	399	5680
Bearing difference (deg)	76.566	56.16	0	180	59833	80.392	52.48	0	180	5680
Fare (£)	18.779	36.99	2	325	59833	17.386	33.68	2.3	315	5680
Parking spaces (car mode)	47.005	90.24	0	1140	16476	123.405	122.52	0	402	1537
Free car park (car mode)	0.540		0	1	16476	0.524		0	1	1537

TABLE 6.3: Summary statistics for choice model variables — WG dataset.

Variable	All cases					Observed choice only				
	Mean	Std.Dev.	Minimum	Maximum	Number of cases	Mean	Std.Dev.	Minimum	Maximum	Number of cases
Nearest	0.096		0	1	97838	0.638		0	1	9367
Car distance (km)	11.238	14.00	0.01	212.83	97838	3.634	7.36	0.02	200.55	9367
Full-time staff	0.100		0	1	97838	0.385		0	1	9367
Part-time staff	0.405		0	1	97838	0.486		0	1	9367
Unstaffed	0.495		0	1	97838	0.129		0	1	9367
Daily service frequency	195.521	272.288	2	1263	97838	422.105	391.72	6	1263	9367
CCTV	0.961		0	1	97838	0.998		0	1	9367
Ticket machine	0.590		0	1	97838	0.846		0	1	9367
Toilets	0.361		0	1	97838	0.793		0	1	9367
Waiting room	0.974		0	1	97838	0.951		0	1	9367
Bus interchange	0.993		0	1	97838	0.997		0	1	9367
Taxi-rank	0.997		0	1	97838	0.996		0	1	9367
HHI	2087.750	559.06	1278.35	10000.00	97838	2012.480	396.42	1362.85	5200.00	9367
Ln(wact)	12.428	1.79	3.41	15.95	97838	12.249	1.88	6.93	15.38	9367
Train leg duration (mins)	63.142	69.21	1	1304	97838	45.526	49.23	3	1187	9367
Waiting time (mins)	10.836	27.03	0	604	97838	0.918	7.97	0	463	9367
On-train time (mins)	51.462	51.31	1	1003	97838	44.396	45.88	1	737	9367
Bearing difference (deg)	80.529	57.08	0	180	97838	77.238	52.02	0	180	9367
Fare (£)	14.763	21.60	0.90	437.00	97838	14.040	20.96	1.50	432.00	9367
Parking spaces (car mode)	108.304	162.77	0	940	26480	228.338	225.62	0	940	2515
Free car park (car mode)	0.010		0	1	26480	0.004		0	1	2515

TABLE 6.4: Summary statistics for choice model variables — LATIS dataset.



	toilets	daily.freq	waiting.room	fulltime	parking.spaces	taxi.rank	buses	ticket.machine	cctv	parttime	nearest	car.dist	ln(wact)	bearing.diff	fare	ontrain.time	wait.time	train.duration	hhi	free.parking	unstaffed
toilets																					
daily.freq	75																				
waiting.room	90	70																			
fulltime	82	88	77																		
parking.spaces	56	68	52	59																	
taxi.rank	61	53	56	57	76																
buses	41	40	40	38	22	22															
ticket.machine	30	36	26	26	20	15	15														
cctv	22	27	23	18	19	15	5	41													
parttime	12	−4	21	−19	7	5	16	14	21												
nearest	15	5	16	8	6	14	14	9	10	22											
car.dist	22	10	22	22	16	12	6	−18	−10	−8	−30										
ln(wact)	−3	34	−4	11	−8	−13	11	48	23	−8	−2	−46									
bearing.diff	−4	−7	−2	−7	−4		−3	−4	−2	3	4	−13	1								
fare		−5		−1	−3	−1		−1	−5	−1		6	−4	7							
ontrain.time	−4	−14	−5	−7	−7	−3		−14	−14	−4	−2	13	−17	15	75						
wait.time	−10	−13	−10	−10	−12	−10	−2	−6	−11	−6	−4	4		8	34	47					
train.duration	−7	−16	−7	−9	−9	−6		−13	−14	−5	−3	12	−14	15	72	96	70				
hhi	−16	−17	−15	−14	−23	−23	−19	−8	−4	−10	−6	−3	10		1		6	2			
free.parking	−31	−34	−27	−35	−7	−10	−22	−36	−12	17		5	−35	7	−1	4	−1	2	−2		
unstaffed	−70	−61	−74	−57	−48	−46	−41	−30	−30	−70	−24	−9	−2	3	2	8	12	11	18	11	

FIGURE 6.4: Correlation matrix for model variables — WG dataset.

likely to be staffed on a full-time basis, to provide better facilities for passengers, and to have larger car parks. Conversely, the strongest negative correlations are seen between a station being unstaffed and these service and facility measures, for example toilets (WG: —0.70; LATIS: —0.73) and service frequency (WG: —0.61; LATIS: —0.46).

The other notable positive correlation is between fare and on-train time (WG: 0.75; LATIS: 0.90). This is not surprising, given that rail ticket pricing in the UK is generally dependent upon the distance travelled for walk-up fares. It should be noted that on-train time and wait-time are both components of the train duration variable, so a strong positive correlation between these variables would be expected.

6.4 Model calibration — multinomial logit models

A series of MNL models were calibrated separately for the WG and LATIS datasets using NLOGIT 5. During the calibration of the models, the predictor variables were entered using a manual forward selection procedure, and variables were retained or rejected based on several factors, including the statistical significance of the estimated parameter, whether the sign of the parameter matched that intuitively expected, and the contribution of the variable

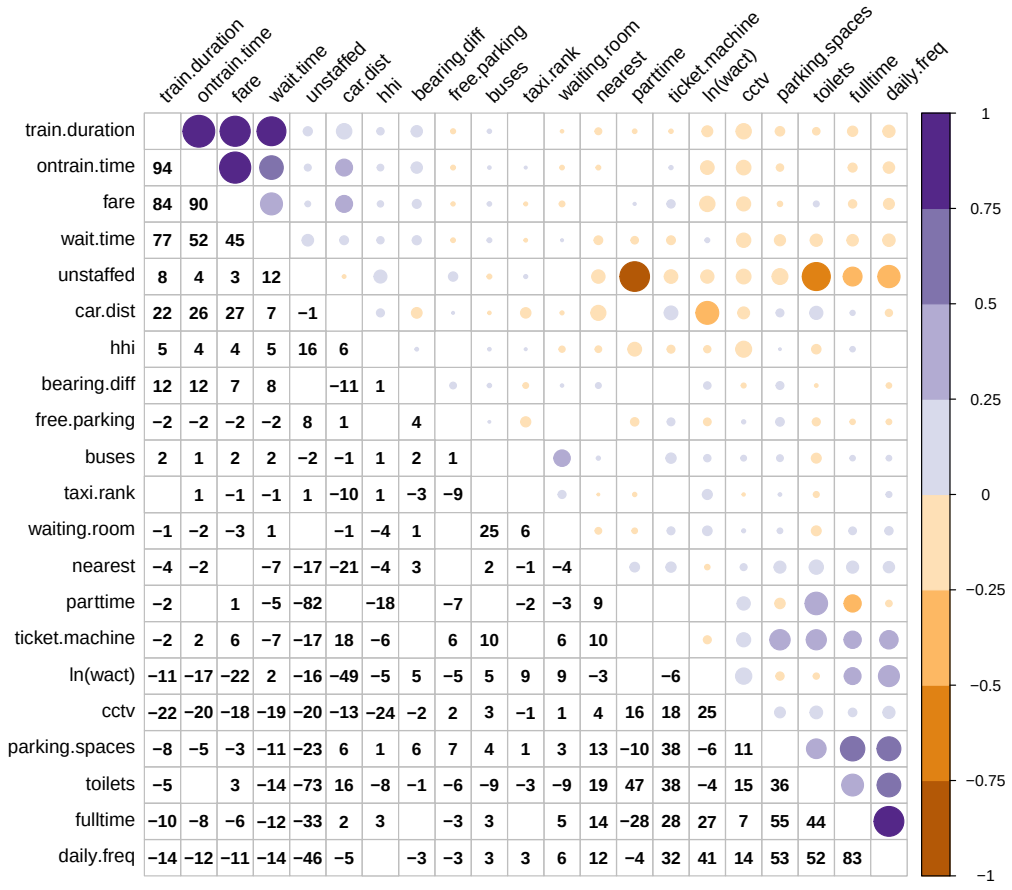


FIGURE 6.5: Correlation matrix for model variables — LATIS dataset.

to model performance. The performance of the models was assessed using log-likelihood, McFadden's adjusted pseudo R^2 (rho-squared), and the AIC, which is considered a more appropriate in-sample measure to compare the predictive accuracy of models (see Section 7.6 for a fuller discussion). The initial log-likelihood (i.e. the NULL model used to calculate the adjusted rho-squared) assumes that there is an equal probability of each alternative in a choice set being chosen. Choice models suitable for use in trip end rail demand models were distinguished from those suitable for flow models, with the latter additionally incorporating variables relating to the train leg and destination.

Although the alternatives comprising the choice sets are named and identifiable, as far as the model construct is concerned the approach adopted is equivalent to an unlabelled choice experiment. As the calibrated models are intended to be used for predictive purposes, when entirely different alternatives will be under consideration, the parameter estimates are considered to be generic and not specific to a particular alternative, and therefore ASCs are immaterial and have not been estimated. This approach differs, for example, from that of Blainey and Evens (2011) where the alternatives were identified by their distance rank within the choice set, and ASCs were estimated for each rank. An additive linear utility function was specified for all the models (see Equation 3.2). In some models non-linear transformations of predictor variable were entered, and some variables were interacted with

dummy variables so that parameters were only estimated on a subset of the choice situations. These cases are described when the relevant model is discussed in the sections that follow. In addition to measures of model fit, the calibration result tables include a measure of model predictive performance, called the 'predictive performance difference'. This is the absolute difference between actual and predicted choice for each station summed across the model and expressed as a percentage of the total number of choice situations, with a *lower* value therefore indicating a better performing model. The measure is discussed at greater length in Section 6.6.1.

6.4.1 Trip end variant models

6.4.1.1 Station access variables

The initial set of models that were calibrated (models TE1 through to TE12) incorporated variables related to accessing the station. The results for these models are shown in Tables 6.5 and 6.6 for the WG and LATIS datasets respectively.

In the first model (TE1), the nearest station (by drive distance) dummy variable was added. As would be expected, given that in 60–70% of the choice situations the nearest station was chosen, this model was a considerable improvement over the null model for both datasets. The WG model performed rather better than the LATIS model, presumably reflecting the larger proportion of choice situations where the nearest station was chosen (70.5% vs. 63.8%). In model TE2 an alternative measure of the nearest station was tested, based on drive time rather than distance. For both datasets, this was an inferior model.

The next stage of calibration concentrated on identifying which measures of the access journey produced the best performing model, with both distance and time-based variables tested. In addition to estimating a single parameter for each variable, which represents only an average effect on utility, mode-specific parameters were estimated by interacting dummy variables for each access mode, or for motorised and non-motorised modes, with the time or distance measure. The models that used time-based measures were found to consistently out-perform those based on distance measures.

The best model for both the WG and LATIS datasets, with an adjusted rho-squared of 0.58 and 0.51 respectively, incorporated mode-specific parameters for access time (model TE12). This was the only model where access journey times were retrieved from OTP for the *actual*

	WG-TE1			WG-TE2			WG-TE3			WG-TE4			WG-TE5			WG-TE9			WG-TE11			WG-TE12		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	3.127	107.2	***				1.814	40.6	***	1.139	23.6	***	1.114	23.1	***	1.761	39.5	***	1.256	28.0	***	1.063	23.2	***
Nearest by time				2.861	102.9	***																		
Distance							-0.277	-28.2	***															
Distance (motorised)										-0.199	-23.4	***												
Distance (non-motorised)										-0.994	-30.5	***												
Distance (walk)													-1.051	-30.3	***									
Distance (cycle)													-0.431	-7.1	***									
Distance (bus)													-0.151	-9.8	***									
Distance (car)													-0.214	-22.8	***									
Time																-0.169	-30.2	***						
Time (motorised)																			-0.084	-21.5	***			
Time (non-motorised)																			-0.099	-30.7	***			
Time (walk)																						-0.106	-31.7	***
Time (cycle)																						-0.140	-7.3	***
Time (bus)																						-0.047	-13.0	***
Time (car)																						-0.136	-23.3	***
Sample size (# trips)	5680			5680			5680			5680			5680			5680			5680			5680		
Initial log-likelihood[a]	-13355			-13355			-13355			-13355			-13355			-13355			-13355			-13355		
Final log-likelihood	-7215			-8194			-6550			-5868			-5836			-6548			-5722			-5627		
McFadden's adjusted R^2	0.46			0.39			0.51			0.56			0.56			0.51			0.57			0.58		
AIC	14433			16391			13104			11742			11681			13100			11450			11264		
Predictive perf. diff. (%)	64.0			76.0			59.9			57.3			57.4			60.7			56.8			56.5		

[a]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.

***, **, * indicate significance at 1%, 5%, 10% level.

TABLE 6.5: Results of station choice MNL models — WG trip end variants (1 of 3).

	LATIS-TE1			LATIS-TE2			LATIS-TE3			LATIS-TE4			LATIS-TE5			LATIS-TE9			LATIS-TE11			LATIS-TE12		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	2.808	130.5	***				1.802	56.6	***	1.202	34.8	***	1.116	31.8	***	1.468	44.3	***	1.009	30.4	***	0.861	25.3	***
Nearest by time				2.691	126.9	***																		
Distance							-0.162	-32.1	***															
Distance (motorised)										-0.109	-25.4	***												
Distance (non-motorised)										-0.729	-35.3	***												
Distance (walk)													-0.880	-35.1	***									
Distance (cycle)													-0.217	-9.6	***									
Distance (bus/subway)													-0.074	-10.9	***									
Distance (car)													-0.128	-25.5	***									
Time																-0.134	-41.1	***						
Time (motorised)																			-0.075	-31.0	***			
Time (non-motorised)																			-0.098	-37.7	***			
Time (walk)																						-0.105	-38.0	***
Time (cycle)																						-0.085	-10.5	***
Time (bus/subway)																						-0.046	-19.1	***
Time (car)																						-0.119	-31.1	***
Sample size (# trips)	9367			9367			9367			9367			9367			9367			9367			9367		
Initial log-likelihood[a]	-21945			-21945			-21945			-21945			-21945			-21945			-21945			-21945		
Final log-likelihood	-13751			-14453			-12752			-11582			-11418			-12439			-10945			-10785		
McFadden's adjusted R2	0.37			0.34			0.42			0.47			0.48			0.43			0.50			0.51		
AIC	27505			28908			25509			23170			22846			24883			21896			21580		
Predictive perf. diff. (%)	72.0			78.2			70.2			65.0			64.7			68.0			61.8			62.0		

[a]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen

***, **, * indicate significance at 1%, 5%, 10% level.

TABLE 6.6: Results of station choice MNL models — LATIS trip end variants (1 of 3).

mode used by the survey respondents[5]. The utility function for model TE12 is as follows:

$$V_{nik} = \beta N + \sum_{m=1}^{4} \gamma_m (Dmode_m \times T_{ikm}), \tag{6.3}$$

where N is a dummy variable with value 1 if alternative k is the nearest station, and zero otherwise; β is the parameter for nearest station; $Dmode_m$ is a dummy variable with value 1 if individual n uses access mode m, and zero otherwise; T_{ikm} is access time from origin i to alternative k using mode m; and γ_m is the access time parameter for mode m.

The estimated parameters suggest that access time is a slightly greater cost to car drivers than to pedestrians, but a substantially lower cost to bus passengers. For example, using the WG model, a 30-minute access journey would reduce the utility of a station by 4.1 units for a car driver, but by only 1.4 units for a bus passenger. There are likely to be more critical considerations than access time for someone reliant on getting a bus to a station, such as which station(s) is(are) served and the bus schedule, and to an extent the travel time has to be accepted. In contrast the car driver has greater control and flexibility, including the option not to travel by train at all.

6.4.1.2 Service and facility variables

The next set of models (TE16 through to TE28) used model TE12 as the starting point, and introduced variables related to station service levels and facilities. The results for these models are shown in Tables 6.7 and 6.9, for the WG dataset, and Tables 6.8 and 6.10 for the LATIS dataset.

The station staffing level dummy variables (part-time and full-time) were added first (model TE16), and these need to be interpreted with reference to the excluded unstaffed level. The utility of a station was found to be higher for staffed stations than unstaffed stations, and the models were substantially improved, particularly on the predictive performance measure. It is not clear how important actual staffing level is in the decision-making process, as it could be an indicator of a range of other station facilities, and full-time staffing level is highly correlated with daily service frequency (WG: 0.88; LATIS: 0.83). In model TE17 staffing level was replaced with daily service frequency, but it was a far inferior model, indicating that staffing level is capturing additional information.

There are a few stations in both datasets that have very high service frequencies relative to the other stations, and this produces a right-skewed distribution with a long tail (see Figure 6.6). By applying a log-normal transformation a distribution that is closer to the normal distribution was obtained (see Figure 6.7). This transformed variable was tested in

[5]Car was specified as the available routing mode in the OTP API query when taxi or motorcycle was given as the access mode; and where access was by bus, bus and walk were specified as the available modes. For the LATIS dataset, subway was made available as an additional routing mode for any observation where either bus or subway was the chosen access mode.

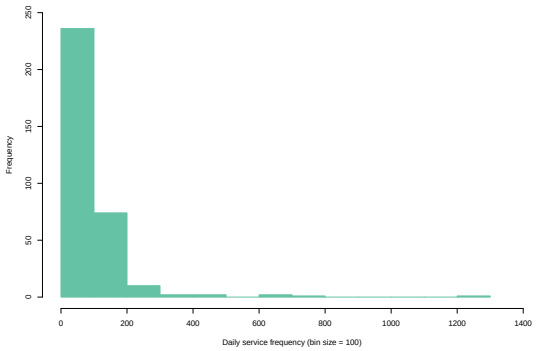


FIGURE 6.6: Histogram of service frequency — unique stations in LATIS dataset.

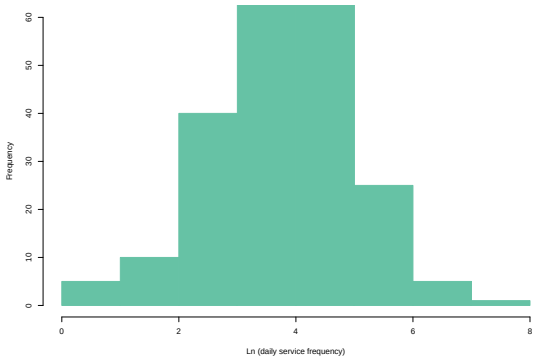


FIGURE 6.7: Histogram of log transformed service frequency — unique stations in LATIS dataset.

model TE18, and found to perform substantially better than the untransformed version, with adjusted rho-squared improving from 0.66 to 0.71 in the WG dataset, and from 0.61 to 0.67 in the LATIS dataset.

In model TE19, both staffing level and log-transformed daily frequency were included. For the LATIS dataset this model performed better than either of the models where these two variables were present alone, although the effect of the correlation between daily frequency and full-time staffing can be seen in the lower parameter estimates for these variables. In the case of the WG dataset, the full-time staffing variable was no longer significant, and in the subsequent model, WG-TE20, the full-time and part-time variables were replaced with the unstaffed dummy. The estimated parameter for this variable was significant and negative, as would be intuitively expected, and the model was also an improvement over models TE16 and TE18.

In the subsequent models (WG-TE21 to WG-TE28; and LATIS-TE20 to LATIS-TE28) the station facilities variables were introduced. Overall, these produced a relatively small improvement in adjusted rho-squared, although there was a distinct improvement in the model predictive performance measure, particularly for the WG dataset.

With respect to the WG dataset, the CCTV, car parking spaces, free car park, ticket machine, toilets, bus interchange and taxi rank parameters were all positive and significant at the 99% level, and each resulted in a significant incremental increase in the log-likelihood ($p < 0.001$)[6]. By introducing these variables the predictive performance measure was reduced from 28.9% (model WG-TE20) to 20.9% (model WG-TE28), indicating a substantial improvement. The variable for waiting room was only significant at the 5% level and caused a slight reduction in predictive performance and was not retained in subsequent models.

For the LATIS dataset, the CCTV, car parking spaces, ticket machine and toilets parameters were positive and significant at the 99% level, and each resulted in a significant incremental increase in the log-likelihood ($p < 0.001$). The improvement in the predictive performance

[6]Calculated using the log likelihood ratio test, with one degree of freedom.

	WG-TE16			WG-TE17			WG-TE18			WG-TE19			WG-TE20			WG-TE21			WG-TE22			WG-TE23		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	0.950	18.2	***	1.044	21.1	***	1.007	19.1	***	0.964	17.9	***	0.957	17.8	***	0.946	17.5	***	0.902	16.7	***	0.920	16.9	***
Time (walk)	-0.137	-30.3	***	-0.140	-30.6	***	-0.140	-30.9	***	-0.135	-29.9	***	-0.141	-30.6	***	-0.140	-30.5	***	-0.134	-29.8	***	-0.134	-29.8	***
Time (cycle)	-0.140	-6.3	***	-0.152	-7.3	***	-0.161	-7.0	***	-0.160	-6.8	***	-0.156	-6.6	***	-0.163	-6.9	***	-0.156	-6.8	***	-0.156	-6.8	***
Time (bus)	-0.042	-11.0	***	-0.054	-14.4	***	-0.061	-15.6	***	-0.057	-14.7	***	-0.054	-13.9	***	-0.055	-13.9	***	-0.051	-13.1	***	-0.052	-13.3	***
Time (car)	-0.146	-24.7	***	-0.162	-28.6	***	-0.187	-30.5	***	-0.180	-27.9	***	-0.175	-28.0	***	-0.177	-27.8	***	-0.199	-28.7	***	-0.190	-27.0	***
Fulltime staffing[a]	3.221	44.4	***							-0.005	0.0	ns												
Part-time staffing[a]	2.079	37.0	***							0.967	14.3	***												
Unstaffed													-1.128	-16.7	***	-1.038	-15.3	***	-1.110	-16.0	***	-1.070	-15.3	***
Weekday service frequency				0.006	42.3	***																		
Ln(service frequency)							1.982	46.4	***	2.042	24.4	***	1.455	28.4	***	1.425	27.7	***	1.199	22.5	***	1.251	22.7	***
CCTV (yes)																0.976	7.1	***	0.968	7.0	***	0.954	7.0	***
Car parking spaces (no.)																			0.005	13.5	***	0.005	13.6	***
Free car park (yes)																						0.465	4.5	***
Sample size (# trips)	5680			5680			5680			5680			5680			5680			5680			5680		
Initial log-likelihood[b]	-13355			-13355			-13355			-13355			-13355			-13355			-13355			-13355		
Final log-likelihood	-4068			-4585			-3899			-3713			-3757			-3727			-3628			-3618		
McFadden's adjusted R^2	0.69			0.66			0.71			0.72			0.72			0.72			0.73			0.73		
AIC	8150			9182			7811			7441			7528			7470			7274			7256		
Predictive perf. diff. (%)	34.7			41.6			32.5			26.9			28.9			28.1			24.8			24.8		

[a]Unstaffed removed from model as reference.

[b]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.

***, **, * indicate significance at 1%, 5%, 10% level.

TABLE 6.7: Results of station choice MNL models — WG trip end variants (2 of 3).

	LATIS-TE16			LATIS-TE17			LATIS-TE18			LATIS-TE19			LATIS-TE20			LATIS-TE21			LATIS-TE22			LATIS-TE23		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	0.873	21.5	***	1.006	27.3	***	0.825	21.6	***	0.808	20.1	***	0.812	20.1	***	0.807	20.0	***	0.806	20.0	***	0.813	19.9	***
Time (walk)	-0.115	-34.9	***	-0.098	-33.0	***	-0.106	-35.0	***	-0.112	-34.9	***	-0.112	-34.9	***	-0.110	-34.6	***	-0.111	-34.6	***	-0.111	-34.8	***
Time (cycle)	-0.085	-10.0	***	-0.075	-10.0	***	-0.096	-11.2	***	-0.098	-11.0	***	-0.098	-10.9	***	-0.096	-10.8	***	-0.096	-10.8	***	-0.096	-10.7	***
Time (bus/subway)	-0.042	-16.0	***	-0.045	-17.8	***	-0.054	-19.3	***	-0.052	-17.7	***	-0.052	-18.1	***	-0.051	-17.9	***	-0.051	-17.9	***	-0.051	-17.7	***
Time (car)	-0.143	-35.4	***	-0.128	-34.2	***	-0.181	-42.6	***	-0.175	-40.6	***	-0.175	-40.4	***	-0.179	-40.3	***	-0.179	-40.3	***	-0.177	-39.7	***
Fulltime staffing[a]	4.401	65.2	***							2.444	27.8	***	2.446	27.9	***	2.398	27.2	***	2.416	27.0	***	2.359	27.0	***
Part-time staffing[a]	1.930	36.6	***							1.112	19.1	***	1.082	18.6	***	1.054	18.1	***	1.064	18.1	***	1.089	18.7	***
Weekday service frequency				0.003	66.6	***																		
Ln(service frequency)							1.850	68.0	***	1.136	31.4	***	1.112	30.3	***	1.059	28.3	***	1.056	28.2	***	0.916	23.7	***
CCTV (yes)													2.047	4.7	***	2.100	4.9	***	2.097	4.9	***	1.893	4.4	***
Car parking spaces (no.)																0.001	7.3	***	0.001	7.2	***	0.001	6.6	***
Free car park (yes)																			0.507	1.4	ns			
Ticket machine (yes)																						0.861	13.8	***
Sample size (# trips)	9367			9367			9367			9367			9367			9367			9367			9367		
Initial log-likelihood[b]	-21945			-21945			-21945			-21945			-21945			-21945			-21945			-21945		
Final log-likelihood	-7348			-8480			-7238			-6811			-6794			-6767			-6766			-6665		
McFadden's adjusted R2	0.66			0.61			0.67			0.69			0.69			0.69			0.69			0.70		
AIC	14710			16971			14488			13639			13606			13553			13554			13353		
Predictive perf. diff. (%)	30.0			45.2			33.8			25.0			24.8			25.3			25.3			23.1		

[a]Unstaffed removed from model as reference

[b]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen

***, **, * indicate significance at 1%, 5%, 10% level.

TABLE 6.8: Results of station choice MNL models — LATIS trip end variants (2 of 3).

measure was less pronounced than in the WG dataset, reducing from 25% (model LATIS-TE19) to 22.4% (model LATIS-TE25). The free car park, waiting room, and bus interchange variables were not significant ($p > 0.05$); while the taxi rank variable was only significant at the 10% level, produced only a small increase in the log-likelihood and no improvement to the predictive performance. These variables were not retained in subsequent models.

It is noticeable that although the staffing level parameters became less important once the range of station facilities and service variables were added to the models, they did remain statistically significant and fairly large. Moving from LATIS model TE16 to TE28, the full-time parameter reduced from 4.4 to 1.9, and the part-time parameter from 1.9 to 0.7; while in the WG models, the negative weighting applied to an unstaffed station reduced from 1.1 (TE20) to 0.6 (TE28). These results suggest that while the service and facilities variables help to explain choice behaviour that was previously being captured collectively by the staffing level variables acting as a proxy, staffing level is an important factor in and of itself. However, there is a potential endogeneity (simultaneity) problem at play. While it is likely that some passengers do prefer stations which have higher staffing levels, stations which are more frequently chosen due to factors not adequately captured by the model will have a better business case to provide more staff.

The parameters for the car parking spaces and free car park variables were only estimated against those choice situations where access mode was car, and this was achieved by interacting these two variables with a dummy variable that took the value of 1 if access mode was car, and 0 otherwise. The parameter appears very small for both datasets, but this only represents the effect of a single extra parking space. For example, model WG-TE28 (where the coefficient is 0.004), predicts that an extra 500 parking spaces would increase the utility of a station by 2 units.

The presence of CCTV was found to have a strong and significant positive effect on station utility, and when introduced to the models had relatively little impact on the other parameters. This result is surprising as this variable has not been included in previous studies of station choice. However, the main source of advice on passenger demand forecasting for the rail industry in Britain, the PDFH (Association of Train Operating Companies, 2013), does recommend a demand uplift when upgrading a station from no CCTV to CCTV of 8% for business and leisure trips and 5% for commuter trips.

6.4.1.3 Land-use (HHI)

When the HHI variable was initially introduced, the models failed to converge. This was resolved by dividing each value of HHI (which, as calculated, could range from $1{,}111.\overline{11}$ to 10,000), by 10,000. The measure entered into the reported models (model TE29 in Tables 6.9 and 6.10) therefore ranged from $0.\overline{11}$ to 1. A range from close to zero to one is a common alternative variant of the HHI index. In the WG model, HHI was significant

	WG-TE24			WG-TE25			WG-TE26			WG-TE27			WG-TE28			WG-TE29			WG-TE31		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	0.916	16.6	***	0.913	16.6	***	0.912	16.6	***	0.878	15.7	***	0.925	16.1	***	0.932	16.2	***	0.936	16.1	***
Time (walk)	-0.132	-29.5	***	-0.131	-29.6	***	-0.132	-29.5	***	-0.133	-29.6	***	-0.135	-29.4	***	-0.136	-29.4	***	-0.137	-29.3	***
Time (cycle)	-0.154	-6.7	***	-0.155	-6.8	***	-0.155	-6.8	***	-0.159	-6.8	***	-0.158	-6.6	***	-0.158	-6.6	***	-0.156	-6.5	***
Time (bus)	-0.055	-14.3	***	-0.056	-14.5	***	-0.056	-14.4	***	-0.057	-14.5	***	-0.055	-13.8	***	-0.053	-13.5	***	-0.053	-13.1	***
Time (car)	-0.195	-27.6	***	-0.199	-27.7	***	-0.197	-27.6	***	-0.199	-27.9	***	-0.192	-27.0	***	-0.190	-26.7	***	-0.189	-26.6	***
Unstaffed	-0.998	-14.1	***	-1.020	-14.5	***	-1.045	-14.2	***	-0.605	-7.9	***	-0.631	-8.1	***	-0.631	-8.1	***	-0.578	-7.4	***
Ln(service frequency)	1.222	22.2	***	1.442	19.1	***	1.316	19.3	***	1.225	22.2	***	1.038	17.6	***	1.013	17.1	***	0.991	16.6	***
CCTV (yes)	0.770	5.6	***	0.68	4.8	***	0.729	5.2	***	0.860	6.2	***	0.947	6.8	***	0.986	7.0	***	1.008	7.2	***
Car parking spaces (no.)	0.005	13.4	***	0.00	13.2	***	0.005	13.3	***	0.005	13.4	***	0.004	10.3	***	0.004	10.0	***	0.004	10.0	***
Free car park (yes)	0.526	5.0	***	0.51	4.8	***	0.504	4.8	***	0.644	6.1	***	0.550	5.2	***	0.513	4.9	***	0.579	5.5	***
Ticket machine (yes)	0.917	9.8	***	0.95	10.1	***	0.900	9.6	***	1.038	10.2	***	1.002	9.8	***	0.930	9.1	***	0.902	8.8	***
Toilets (yes)				-0.38	-4.4	***															
Waiting room (yes)							-0.192	-2.4	**												
Bus interchange (yes)										0.923	13.9	***	0.818	12.1	***	0.825	12.1	***	0.864	12.6	***
Taxi rank (yes)													0.551	9.0	***	0.434	6.6	***	0.272	3.6	***
POI (HHI/10000)																-5.162	-5.0	***	-4.476	-4.3	***
Ln(accessibility term)																			-0.282	-4.7	***
Sample size (# trips)	5680			5680			5680			5680			5680			5680			5680		
Initial log-likelihood[a]	-13355			-13355			-13355			-13355			-13355			-13355			-13355		
Final log-likelihood	-3567			-3558			-3564			-3468			-3427			-3414			-3403		
McFadden's adjusted R^2	0.73			0.73			0.73			0.74			0.74			0.74			0.74		
AIC	7156			7139			7152			6959			6879			6855			6835		
Predictive perf. diff. (%)	24.5			23.9			24.2			23.1			20.9			21.0			20.5		

[a]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.

***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.9: Results of station choice MNL models — WG trip end variants (3 of 3).

	LATIS-TE25			LATIS-TE26			LATIS-TE27			LATIS-TE28			LATIS-TE29			LATIS-TE31		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	0.807	19.80	***	0.808	19.81	***	0.805	19.74	***	0.807	19.79	***	0.800	19.64	***	0.769	18.60	***
Time (walk)	-0.110	-34.90	***	-0.110	-34.88	***	-0.110	-34.90	***	-0.111	-34.90	***	-0.111	-34.95	***	-0.111	-35.13	***
Time (cycle)	-0.095	-10.67	***	-0.095	-10.67	***	-0.095	-10.67	***	-0.095	-10.65	***	-0.096	-10.75	***	-0.097	-10.88	***
Time (bus/subway)	-0.050	-17.46	***	-0.050	-17.39	***	-0.050	-17.45	***	-0.050	-17.44	***	-0.050	-17.48	***	-0.053	-17.70	***
Time (car)	-0.177	-39.53	***	-0.177	-39.51	***	-0.177	-39.53	***	-0.177	-39.55	***	-0.178	-39.59	***	-0.183	-38.82	***
Fulltime staffing[a]	1.931	18.68	***	1.935	18.74	***	1.911	18.31	***	1.944	18.75	***	1.976	18.94	***	2.117	19.18	***
Part-time staffing[a]	0.720	9.44	***	0.723	9.49	***	0.704	9.11	***	0.729	9.53	***	0.766	9.88	***	0.775	10.03	***
Ln(service frequency)	0.876	22.53	***	0.877	22.55	***	0.877	22.55	***	0.872	22.36	***	0.868	22.30	***	0.823	20.35	***
CCTV (yes)	1.922	4.45	***	1.910	4.42	***	1.882	4.35	***	1.918	4.44	***	1.996	4.60	***	2.063	4.79	***
Car parking spaces (no.)	0.001	6.84	***	0.001	6.86	***	0.001	6.81	***	0.001	6.81	***	0.001	6.68	***	0.001	7.47	***
Ticket machine (yes)	0.687	10.40	***	0.693	10.47	***	0.674	10.11	***	0.689	10.43	***	0.687	10.37	***	0.677	10.19	***
Toilets (yes)	0.532	7.58	***	0.523	7.39	***	0.556	7.70	***	0.536	7.63	***	0.533	7.58	***	0.531	7.56	***
Waiting room (yes)				-0.167	-1.11	ns												
Bus interchange (yes)							0.314	1.40	ns									
Taxi rank (yes)										0.774	1.81	*						
POI (HHI/10000)													1.456	3.41	***			
Ln(accessibility term)																0.160	4.94	***
Sample size (# trips)	9367			9367			9367			9367			9367			9367		
Initial log-likelihood[b]	-21945			-21945			-21945			-21945			-21945			-21945		
Final log-likelihood	-6636			-6635			-6635			-6634			-6630			-6624		
McFadden's adjusted R2	0.70			0.70			0.70			0.70			0.70			0.70		
AIC	13296			13297			13296			13295			13287			13273		
Predictive perf. diff. (%)	22.4			22.4			22.3			22.4			22.2			21.6		

[a]Unstaffed removed from model as reference.
[b]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.
***, **, * indicate significance at 1%, 5%, 10% level.

TABLE 6.10: Results of station choice MNL models — LATIS trip end variants (3 of 3).

at the 1% level, and had a negative sign. This corresponds to the a priori expectation for this variable, that passengers will gain greater utility from stations with a diverse mix of the top-level POI classifications in their immediate vicinity, than from stations located within a more homogeneous land-use environment. For example, passengers may prefer a station where they can carry out a range of other activities prior to catching their train, such as shopping, getting refreshments, or going to the bank. The coefficient appears quite large (−5.162), and as the HHI of stations in the WG dataset ranges from 0.14 to 0.72, representing an effect on utility of between −0.7 and −3.71 units, has the potential to substantially impact relative station utility. However, the variable had a minimal impact on the model, which had a slightly higher log-likelihood than the prior model, but was marginally worse in terms of predictive performance. In the case of the LATIS model, although the parameter was also significant at the 1% level, it had a positive sign, suggesting that stations would have a higher utility as land-use mix becomes less diverse. As a behavioural explanation for this result is difficult to justify, the HHI variable was removed from subsequent LATIS models.

6.4.1.4 Accessibility term

The process of deriving the accessibility term proved unexpectedly computationally intensive[7], and involved the following main steps:

- Every possible combination of two stations present within the choice sets was identified.
- The walk distance between each unique station pair was obtained from OTP.
- The accessibility term was then calculated for each station in each choice set. This required a series of processes for every row in the dataset:
    1. identify the current station (alternative) for the current row.
    2. identify which choice set the current station belongs to.
    3. identify the other stations in this choice set.
    4. retrieve the distance to each of the other stations from the current station.
    5. retrieve the entries/exits for each of the other stations
    6. calculate the accessibility term.

The calculated accessibility term was introduced in the final trip end variant model (TE31). In the WG model the parameter was significant at the 1% level and was negative. As explained in Section 6.2.1.3, when the accessibility term increases, a station is on average nearer to more attractive alternatives *within a specific choice set*, and a negative parameter therefore

[7]To the extent that scripting this procedure in R proved impractical when the station choice models were run for every postcode in GB during calibration of national trip end models, and SQL procedural code was written instead (see Section 7.5.3.1 for details).

suggests the presence of a competition effect. As the accessibility term ranges between 5.18 and 14.11 in the WG dataset, the estimated parameter (−.282) has the potential to reduce station utility by between −1.46 and −3.98 units. However, the maximum difference in the accessibility term of stations within any given choice set is lower, at 3.71, indicating a maximum utility difference of −1.05 units. The choice set for the observation with the maximum difference is mapped in Figure 6.8. It can be seen that Swansea (SWA) has by far the lowest weighted accessibility term (−1.88), reflecting the fact that the other stations within the choice set have substantially fewer annual trips (the attraction variable used to weight the accessibility term). Llansamlet station (LAS) has the highest weighted accessibility term (−2.93), reflecting its proximity to Swansea; and the weighted term then gradually reduces to −2.53 at Llandeilo (LLO) as the influence of Swansea diminishes. In this example, the chosen station was Swansea.

In the LATIS model, the accessibility term parameter is positive (0.160), which suggests an agglomeration effect, where stations are more likely to be chosen if they are nearer to other (more attractive) stations. The purpose of this variable was to attempt to address the proportional substitution behaviour of MNL models, so that when the model is used in a planning capacity it can allow a new station to have a greater influence (i.e. to abstract proportionally more passengers) from closer stations than more distant ones. A positive parameter would have the opposite effect, so would not be a useful mechanism to address this issue. In the case of both the WG and LATIS datasets the accessibility term improved the model, despite the difference in parameter sign, with a small (but statistically significant) reduction in log-likelihood, and a small improvement in the predictive performance measure.

The inconsistency in the sign of the accessibility term parameter between the two datasets is clearly problematic and does not give confidence that this is an appropriate mechanism for capturing spatial competition effects and modifying the proportional substitution behaviour of the MNL model. However, the decision to append the nearest major station to each choice set (in order to increase the proportion of observed choice accounted for) will have created artificial spatial relationships between stations that may have undermined the CDM. This issue is discussed further in Section 6.7 when the definition of choice sets for calibration of a combined dataset model is considered.

6.4.1.5 Summary of best performing models

The most suitable models for incorporating into trip end rail demand models, calibrated using the two datasets, are WG-TE31 and LATIS-TE25 which have an adjusted rho-squared of 0.74 and 0.70 respectively, and a predictive performance measure of 20.5% and 22.4% respectively. The utility function (V) for model WG-TE31, for individual n at origin i choosing

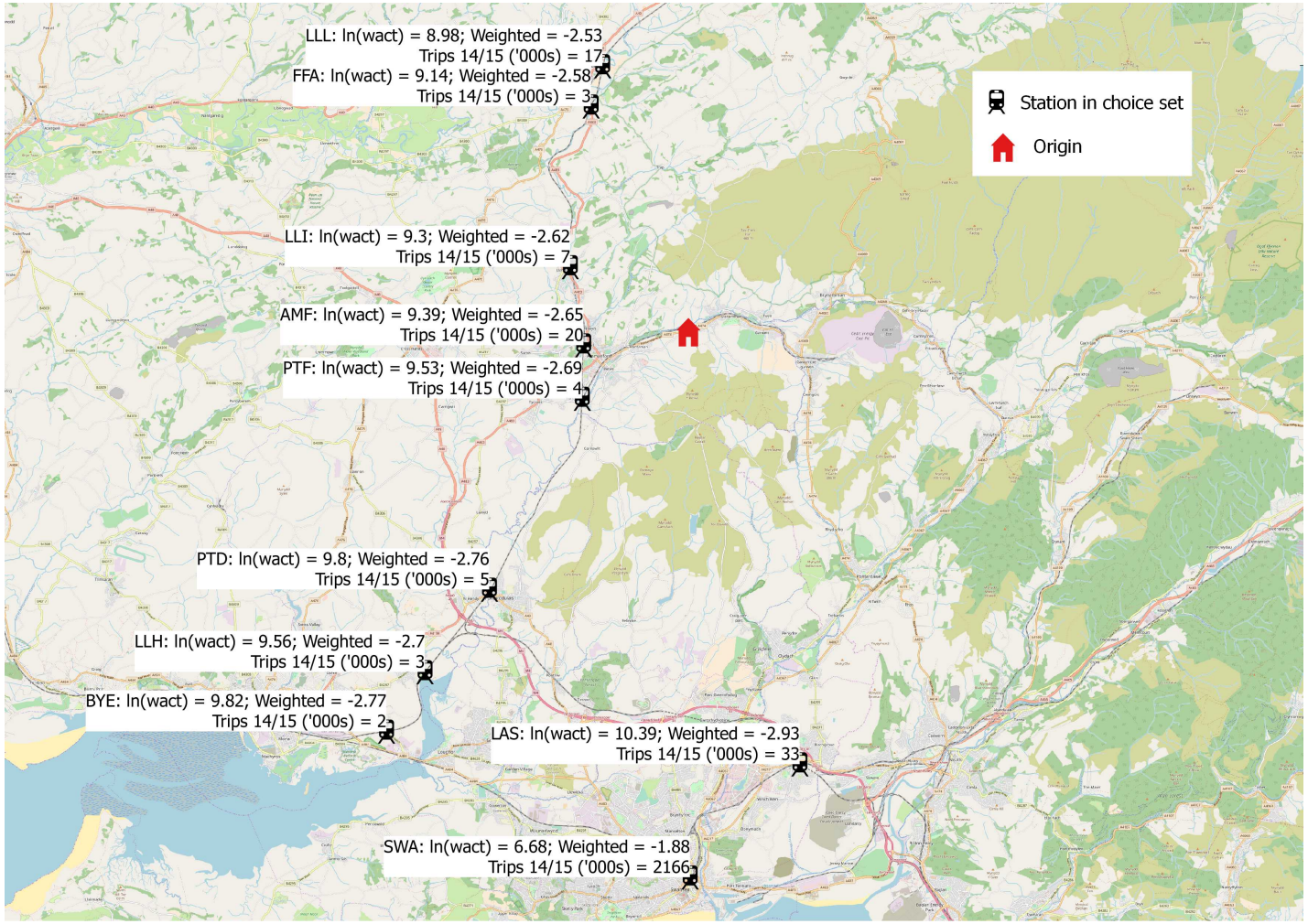


FIGURE 6.8: Choice set for a single observation in WG dataset, showing the accessibility term for each station, and the weighted term using the estimated parameter from model TE31 (−0.282).

station k, is as follows:

$$V_{nik} = \beta N + \sum_{m=1}^{4} \gamma_m (Dmode_m \times T_{ikm}) + \delta U + \epsilon \ln F_k + \zeta C_k + \eta (Dcar \times Ps_k) + \theta (Dcar \times Pf) + \iota Tm_k + \kappa B_k + \lambda Tr_k + \mu H_k + \nu \ln A_k, \tag{6.4}$$

where $Dmode_m$ is a dummy variable with value 1 if individual n uses access mode m, and zero otherwise; T_{ikm} is access time from origin i to alternative k using mode m; F is the daily service frequency; $Dcar$ is a dummy variable with value 1 if individual n accessed the station by car; Ps is the number of car parking spaces; H is the HHI, A is the accessibility term; N, U, C, Pf, Tm, B, and Tr are dummy variables that take the value of 1 if station k is the nearest station (by distance), unstaffed, has CCTV, has a free car park, has a ticket machine, has a bus interchange, or has a taxi-rank respectively, and zero otherwise; and β, γ, δ, ϵ, ζ, η, θ, ι, κ, λ, μ and ν are parameters to be estimated. The utility function (V) for model LATIS-TE25, for individual n at origin i choosing station k, takes the following form:

$$V_{nik} = \beta N + \sum_{m=1}^{4} \gamma_m (Dmode_m \times T_{ikm}) + \delta Ft + \epsilon Pt + \zeta \ln F_k + \eta C_k + \theta (Dcar \times Ps_k) + \iota Tm_k + \kappa To_k, \tag{6.5}$$

where Ft, Pt, and To are dummy variables that take the value of 1 if station k is full-time staffed, part-time staffed, or has toilets, and zero otherwise.

6.4.2 Flow variant models

The starting point for calibrating the flow variant models are models WG-TE29 (in preference to WG-TE31, as the accessibility term is introduced again at the end of flow variant calibration) and LATIS-TE25. The results are shown in Tables 6.11 and 6.12.

6.4.2.1 Train leg variables

The duration of the train leg (in minutes) was introduced in model FM1, and produced an improvement over the previous models, especially for the LATIS dataset where adjusted rho-squared increased from 0.70 to 0.78 and there was a substantial uplift in predictive performance (with the predictive performance difference measure reducing from 22.4% to 14.5%). An effect of introducing the train leg variable was to increase the size of the mode-specific access time parameters, which had been very consistent up to this point (since introduction of the service frequency variable). The most notable change was for car mode, where the parameter reduced from −0.190 to −0.229 in the WG model, and from −0.177 to −0.281 in the LATIS model. It may be that the prior models were unable to adequately explain longer access journeys to a chosen station. If decisions to travel further by car to

	LATIS-FM1			LATIS-FM2			LATIS-FM3			LATIS-FM4			LATIS-FM5			LATIS-FM6			LATIS-FM7			LATIS-FM8		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	0.691	16.0	***	0.692	15.7	***	0.770	17.1	***	0.745	16.5	***	0.703	15.6	***	0.751	16.3	***	0.750	16.7	***	0.731	15.9	***
Time (walk)	-0.133	-54.2	***	-0.122	-34.7	***	-0.126	-40.5	***	-0.122	-35.0	***	-0.123	-34.6	***	-0.125	-35.1	***	-0.130	-52.3	***	-0.122	-34.4	***
Time (cycle)	-0.136	-12.1	***	-0.121	-11.3	***	-0.130	-11.9	***	-0.124	-11.5	***	-0.114	-10.7	***	-0.125	-11.3	***	-0.128	-11.5	***	-0.115	-10.7	***
Time (bus/subway)	-0.100	-32.9	***	-0.077	-20.6	***	-0.088	-23.9	***	-0.077	-20.5	***	-0.071	-19.2	***	-0.083	-22.4	***	-0.087	-37.1	***	-0.070	-19.4	***
Time (car)	-0.281	-50.1	***	-0.247	-43.1	***	-0.272	-46.7	***	-0.249	-43.1	***	-0.231	-38.6	***	-0.258	-41.5	***	-0.263	-45.5	***	-0.230	-38.5	***
Fulltime staffing[a]	1.929	16.3	***	1.727	15.3	***	1.992	16.7	***	1.920	16.6	***	1.900	16.5	***	1.968	16.6	***	1.781	13.9	***	1.716	13.8	***
Part-time staffing[a]	0.687	7.6	***	0.673	7.8	***	0.696	7.8	***	0.716	8.3	***	0.702	8.1	***	0.675	7.6	***	0.636	7.0	***	0.660	7.6	***
Ln(service frequency)	0.418	8.5	***	0.453	9.8	***	0.405	8.4	***	0.442	9.6	***	0.469	10.1	***	0.440	9.0	***	0.466	9.4	***	0.491	10.5	***
CCTV (yes)	2.362	5.7	***	2.008	4.8	***	2.292	5.5	***	2.152	4.9	***	1.912	4.5	***	2.080	5.1	***	2.082	5.0	***	1.793	4.2	***
Car parking spaces (no.)	0.002	10.3	***	0.002	10.2	***	0.002	10.3	***	0.002	9.9	***	0.001	9.1	***	0.002	9.5	***	0.002	8.6	***	0.001	8.4	***
Ticket machine (yes)	0.518	6.5	***	0.536	7.0	***	0.528	6.7	***	0.553	7.2	***	0.525	6.8	***	0.500	6.4	***	0.507	6.4	***	0.532	7.0	***
Toilets (yes)	0.428	5.2	***	0.417	5.3	***	0.448	5.5	***	0.437	5.5	***	0.458	5.8	***	0.465	5.7	***	0.486	5.9	***	0.474	5.9	***
Train duration	-0.131	-56.5	***				-0.121	-44.2	***							-0.118	-40.7	***	-0.124	-51.7	***			
On-train time				-0.092	-32.6	***				-0.081	-28.1	***	-0.076	-25.9	***							-0.079	-25.9	***
Wait-time				-0.141	-29.3	***				-0.142	-29.4	***	-0.142	-29.4	***							-0.144	-29.5	***
Bearing difference							-0.002	-5.7	***	-0.005	-11.0	***												
Bearing difference (0-5km access)													-0.004	-8.7	***	-0.002	-4.1	***	-0.002	-4.2	***	-0.004	-8.9	***
Bearing difference (5-10km access)													-0.007	-8.9	***	-0.004	-5.1	***	-0.004	-4.9	***	-0.007	-8.8	***
Bearing difference (10-15km access)													-0.011	-8.2	***	-0.007	-5.2	***	-0.006	-4.7	***	-0.010	-7.9	***
Bearing difference (15-20km access)													-0.014	-6.9	***	-0.010	-4.5	***	-0.009	-4.0	***	-0.013	-6.5	***
Bearing difference (20+ km access)													-0.019	-7.5	***	-0.014	-5.6	***	-0.013	-4.9	***	-0.017	-6.8	***
Ln(accessibility term)																			-0.179	-4.4	***	-0.157	-3.9	***
Sample size (# trips)	9367			9367			9367			9367			9367			9367			9367			9367		
Initial log-likelihood[b]	-21945			-21945			-21945			-21945			-21945			-21945			-21945			-21945		
Final log-likelihood	-4785			-5221			-4774			-5162			-5134			-4758			-4745			-5123		
McFadden's adjusted R2	0.78			0.76			0.78			0.76			0.77			0.78			0.78			0.77		
AIC	9596			10470			9575			10354			10305			9552			9528			10286		
Predictive perf. diff. (%)	14.5			14.7			14.7			15.0			14.5			14.4			14.2			14.4		

[a]Unstaffed removed from model as reference.

[b]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.

***, **, * indicate significance at 1%, 5%, 10% level.

TABLE 6.11: Results of station choice MNL models — LATIS flow variants.

	WG-FM1			WG-FM2			WG-FM3			WG-FM4			WG-FM5			WG-FM6		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	0.895	15.1	***	0.896	15.1	***	0.888	15.0	***	0.891	14.9	***	0.902	15.1	***	0.896	15.1	***
Time (walk)	-0.141	-29.6	***	-0.141	-29.6	***	-0.141	-29.4	***	-0.140	-29.4	***	-0.142	-29.5	***	-0.142	-29.5	***
Time (cycle)	-0.170	-6.7	***	-0.170	-6.7	***	-0.170	-6.7	***	-0.169	-6.6	***	-0.167	-6.6	***	-0.167	-6.6	***
Time (bus)	-0.069	-14.7	***	-0.069	-14.3	***	-0.069	-14.7	***	-0.066	-13.9	***	-0.068	-14.7	***	-0.069	-14.9	***
Time (car)	-0.229	-28.7	***	-0.229	-27.6	***	-0.228	-28.7	***	-0.218	-25.3	***	-0.227	-28.5	***	-0.228	-28.7	***
Unstaffed	-0.584	-7.2	***	-0.585	-7.2	***	-0.607	-7.4	***	-0.608	-7.4	***	-0.538	-6.6	***	-0.525	-6.5	***
Ln(service frequency)	0.521	8.1	***	0.524	8.1	***	0.503	7.8	***	0.511	7.8	***	0.491	7.5	***	0.514	8.1	***
CCTV (yes)	1.188	8.3	***	1.187	8.3	***	1.184	8.3	***	1.175	8.2	***	1.212	8.5	***	1.210	8.5	***
Car parking spaces (no.)	0.004	9.1	***	0.004	9.0	***	0.004	9.1	***	0.004	9.1	***	0.004	9.1	***	0.004	9.6	***
Free car park (yes)	0.718	6.5	***	0.715	6.5	***	0.714	6.5	***	0.713	6.4	***	0.780	7.0	***	0.801	7.2	***
Ticket machine (yes)	0.870	8.2	***	0.875	8.1	***	0.894	8.4	***	0.908	8.4	***	0.837	7.8	***	0.835	7.8	***
Bus interchange (yes)	0.870	12.4	***	0.872	12.4	***	0.862	12.3	***	0.863	12.3	***	0.902	12.8	***	0.925	13.5	***
Taxi rank (yes)	0.279	4.0	***	0.276	3.9	***	0.260	3.7	***	0.255	3.6	***	0.122	1.6	ns			
POI (HHI/10000)	-5.877	-5.5	***	-5.839	-5.4	***	-5.751	-5.4	***	-5.680	-5.3	***	-5.257	-4.9	***	-5.676	-5.4	***
Train duration	-0.063	-22.5	***				-0.065	-22.6	***	-0.065	-22.1	***	-0.063	-22.5	***	-0.063	-22.6	***
On-train time				-0.061	-14.0	***												
Wait-time				-0.064	-16.2	***												
Bearing difference							0.002	3.7	***									
Bearing difference (0-5km access)										0.002	4.3	***						
Bearing difference (5-10km access)										0.002	2.1	**						
Bearing difference (10-15km access)										0.000	0.1	ns						
Bearing difference (15-20km access)										-0.008	-2.8	***						
Bearing difference (20+ km access)										-0.004	-1.2	ns						
Ln(accessibility term)													-0.29	-4.4	***	-0.332	-5.7	***
Sample size (# trips)	5680			5680			5680			5680			5680			5680		
Initial log-likelihood[a]	-13355			-13355			-13355			-13355			-13355			-13355		
Final log-likelihood	-3074			-3074			-3067			-3059			-3064			-3065		
McFadden's adjusted R^2	0.77			0.77			0.77			0.77			0.77			0.77		
AIC	6178			6181			6167			6157			6161			6161		
Predictive perf. diff. (%)	19.4			19.4			19.4			19.4			19.0			19.0		

[a]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.
***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.12: Results of station choice MNL models — WG flow variants.

board at a station with faster direct train services can now be accounted for by a smaller train leg disutility, then the disutility associated with the access journey per se can increase.

In model FM2, the train leg was split into on-train time and wait-time (due to transfers). In the LATIS model the wait-time parameter was 1.53 times larger than the on-train parameter, which is reasonably consistent with the convention that wait time is valued at twice the rate of in-vehicle time (Association of Train Operating Companies, 2013) (in subsequent models, once the 'bearing difference' variable had been introduced the differential was greater, for example wait-time was valued at 1.82 times on-train time in model LATIS-FM8). However, this was not replicated in the WG model where wait-time was valued only marginally higher than on-train time, and both parameters were very similar to the train duration parameter. There is a potential problem with the datasets that may have impacted the estimation of train leg parameters. The questionnaire used in both the WG and LATIS surveys asked respondents for the boarding and alighting station of the train they were *currently travelling on*, rather than their ultimate boarding and alighting station. To ensure that the ultimate origin and destination stations were accurately identified it was therefore necessary to exclude any observations where the respondent indicated that their access or egress mode was another train. In theory this should mean that none of the retained observations involved a transfer between trains. In reality, this is not the case, presumably because some respondents had the entirety of their trip in mind rather than the current train. However, this does mean that there are likely to be artificially fewer observations in the dataset where the train leg from the chosen station involved a transfer between trains than would be the case in reality (and the extent of this might differ between the two datasets).

The LATIS FM2 model, with the train leg split, performed somewhat worse than the FM1 model on all the measures, whilst there was no significant difference between the two WG models. For subsequent WG models only the train leg duration was retained, while both measures of the train leg were tested with additional variables in subsequent LATIS models.

The train fare variable was not included in the models due to a very high correlation with other train leg variables, for example a 0.9 correlation with on-train time in the LATIS dataset.

6.4.2.2 Difference in bearing variable

The 'difference in bearing' variable, described in Section 5.4.1, was added next. In the LATIS models (LATIS-FM3 and LATIS-FM4) this had the expected negative sign, indicating that a station is less likely to be chosen as the difference in bearing from origin:origin station and origin:destination increases, suggesting a preference for a station that is in the same direction of travel as the ultimate destination. However, the variable did not have the expected sign in the WG model (LATIS-FM3). It was hypothesised that this may become a more important factor as the access journey distance increases, and might be of little consequence for short access journeys. This was investigated in subsequent models by estimating five separate

parameters for the variable based on banded access journey time. In the LATIS models (LATIS-FM5 and LATIS-FM6) the parameters showed the expected effect, with a gradual increase in the size of the negative parameter as access distance increases, and produced a small improvement in model fit and predictive performance over the models without this variable. The effect of a 45-degree difference in bearing ranged from −0.1 for access journeys < 5 km, to −0.6 for access journeys > 20 km (using model LATIS-FM6). In the WG model only the parameters for the two longer access bands had the expected negative sign, but only the parameter for the 15–20 km band was significant. It is possible that the geography of the South Wales valleys has affected this variable in the WG dataset. Each of the valley rail lines, which mostly radiate out from central Cardiff, are confined to their respective valley along with the associated road network used for station access. As a consequence, stations in any given choice set might be largely confined to the same valley, thus limiting the variability of the bearing difference amongst alternatives.

6.4.2.3 Accessibility term

When the accessibility term was re-introduced to the WG model (WG-FM5), the dummy variable for taxi-rank was no longer significant, and this was removed and the model re-run. In model WG-FM6 the accessibility term has a negative sign and a similar parameter value to that estimated in the trip end variant (−0.282 in WG-TE31, compared to −0.332 in WG-FM6). Unlike the trip end variant models, the accessibility term also had a negative sign in the LATIS models (LATIS-FM7 and LATIS-FM8), and was significant at the 1% level. In both datasets the models with the accessibility term performed slightly better, both in terms of measures of fit and predictive performance, than those without.

6.4.2.4 Summary of best performing models

The most suitable models for incorporating into flow rail demand models, calibrated using the two datasets, are WG-FM6 and LATIS-FM7 which have an adjusted rho-squared of 0.77 and 0.78 respectively, and a predictive performance measure of 19.0% and 14.2% respectively. The utility function (V) for model WG-FM6, for individual n at origin i choosing station k and travelling to destination station j, is as follows:

$$\begin{aligned} V_{nikj} = \beta N + \sum_{m=1}^{4} \gamma_m (Dmode_m \times T_{ikm}) + \delta U + \epsilon \ln F_k + \zeta C_k + \eta (Dcar \times Ps_k) \\ + \theta (Dcar \times Pf_k) + \iota Tm_k + \kappa B_k + \lambda H_k + \mu Tl_{kj} + \nu \ln A_k, \end{aligned} \tag{6.6}$$

where Tl_{kj} is the duration of the train leg from origin station k to destination station j. The utility function (V) for model LATIS-FM7, for individual n at origin i choosing station k and

travelling to destination station j takes the following form:

$$V_{nik} = \beta N + \sum_{m=1}^{4} \gamma_m (Dmode_m \times T_{ikm}) + \delta Ft + \epsilon Pt + \zeta \ln F_k + \eta C_k + \theta (Dcar \times Ps_k) + \iota Tm_k + \kappa To_k + \lambda Tl_{kj} + \sum_{b=1}^{5} \mu_b (Dbearing_b \times Bdif_{ikij}) + \nu \ln A_k, \tag{6.7}$$

where $Dbearing_b$ is a dummy variable with value 1 if the access journey to alternative k falls within distance band b, and zero otherwise; and $Bdif_{ikij}$ is the bearing difference between origin i to alternative k and origin i and destination station j.

6.5 Model calibration — random parameter (mixed) logit models

A potential weakness of the MNL model is that it does not allow for individual taste variation in the estimated parameters. The random parameter specification of the mixed logit model allows some or all of the parameters to vary by individual, from a distribution specified by the researcher. However, the model is more complex than MNL and the calculation of probabilities does not take a closed form. Instead the probabilities have to be simulated, and model estimation takes significantly longer to complete. Utility is specified in the same way as with the MNL model, except the vector of coefficients is now able to vary by individual, and the probability of individual n choosing alternative i from a choice set of J alternatives is an integral given by the following equation:

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}} \right) f(\beta) d\beta, \tag{6.8}$$

where β' is a vector of coefficients for variables x for individual n, and the coefficients vary over the population with density $f(\beta)$ (Train, 2009).

6.5.1 Trip end variant models

Initial RPL models were run, using the best performing trip end variant MNL models (WG-TE31 and LATIS-TE25) as the starting point, with all parameters specified as random (apart from the accessibility term and HHI[8]) to test whether the standard deviation (SD) of each

[8]The accessibility term is included to capture spatial correlation effects, and it was therefore considered inappropriate to specify its parameter as random. Due to difficulties with the MNL models converging when the HHI variable was included, it was decided to specify its parameter as non-random in the more complex simulation models.

parameter was significantly different from zero. If the SD is not significant, it indicates that there is no individual taste variation for that parameter. As the parameter for all the model variables was expected to have the same sign for all individuals, $f(\beta)$ was specified as log-normal, with those variables expected to have a negative sign entered as negative values. Halton draws were used for the simulation, with 75 and 100 draws for the WG and LATIS datasets respectively. The results of these initial models are shown in Table 6.13.

The SD of the nearest station and mode-specific access time parameters were significant at the 1% level for both the WG and LATIS models, with the exception of cycle mode in the WG model, where the SD was only significant at the 10% level. In addition, the SD of the part-time staffing and car park spaces parameters was significant at the 1% level in the LATIS model. With the exception of the taxi-rank parameter in the WG model, where the SD was significant at the 5% level, the SD of the remaining parameters had low z-values in both datasets, and were not close to critical values. Based on these findings an RPL model (model RPL1) was run for both datasets, with the parameters that had significant SDs at the 1% level specified as random. For both datasets, the SD of the nearest station parameter was not significant in this first model, and neither was the SD of the part-time staffing parameter in the LATIS model. An additional model was therefore run with these variables no longer specified as random (model RPL2). In the LATIS RPL2 model, the SD of the car park spaces parameter was no longer significant, and so a third model was run with this variable no longer specified as random (RPL3). The results of the various models are shown in Tables 6.14 and 6.15. These also show the median, mean, and standard deviation of the random parameters, calculated from the log-normal parameters using the formulae below, following Train (2009, p. 150):

$$\tilde{B} = \exp(m), \tag{6.9a}$$

$$\bar{B} = \exp(m + (s^2/2)), \tag{6.9b}$$

$$\text{std}(B) = \bar{B} \times \sqrt{(\exp(s^2) - 1)}, \tag{6.9c}$$

where m is the mean of $\ln(B)$ and s is the standard deviation of $\ln(B)$.

Both the WG and LATIS models (RPL2 and RPL3 respectively) had higher log-likelihood and adjusted rho-squared values than the equivalent MNL model, and although predictive performance was slightly better for the WG model (20.2% vs. 20.5%), it was marginally worse for the LATIS model (22.85% vs. 22.4%). The SD of the random parameters was significant, indicating that the parameter estimates are individual-specific and for any individual the parameter may be different from the mean parameter estimate (Hensher et al., 2016). Interestingly, the variability in the parameter for walk access time was much greater in the WG model (SD 0.18) than it was in the LATIS model (SD 0.06), while there was greater variability in the parameter for car access time in the LATIS model (SD 0.31) compared with the WG model (SD 0.15). The RPL model also had an effect on the non-random parameters, when compared to the MNL model, most noticeably a substantially smaller parameter for

	WG initial model (starting TE31)									LATIS initial model (starting TE25)								
	Random parameters[a]						Non-random parameters			Random parameters[a]						Non-random params.		
variable	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*z*	*Sig*	*B*	*z*	*Sig*	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest station (yes)	-1.079	-3.5	***	1.260	4.7	***				-1.277	-4.7	***	1.024	4.0	***			
Time - walk (mins)	-1.544	-26.7	***	0.659	8.6	***				-1.912	-50.5	***	0.413	7.1	***			
Time - cycle (mins)	-1.476	-8.3	***	0.573	1.9	*				-1.786	-12.5	***	0.777	4.4	***			
Time bus/pt (mins)	-2.640	-23.0	***	0.567	8.3	***				-2.547	-46.5	***	0.775	29.3	***			
Time (car) mins	-1.367	-25.3	***	0.445	7.6	***				-1.270	-34.2	***	0.731	19.2	***			
Ln(frequency)	0.148	2.0	**	0.035	0.1	ns				0.030	0.7	ns	0.006	0.0	ns			
Full-time (yes)[b]										0.763	11.6	***	0.015	0.1	ns			
Part-time (yes)[b]										-0.357	-2.1	**	0.565	4.1	***			
Unstaffed (yes)	-0.270	-1.7	*	0.124	0.2	ns												
CCTV (yes)	0.109	0.5	ns	0.748	1.5	ns				1.120	2.7	***	0.017	0.0	ns			
Car park spaces (#)	-5.485	-39.8	***	0.350	1.1	ns				-8.129	-135.5	***	2.394	178.4	***			
Free car park (yes)	-0.382	-1.1	ns	0.266	0.2	ns												
Ticket machine (yes)	-0.046	-0.3	ns	0.016	0.0	ns				-0.076	-0.6	ns	0.131	0.2	ns			
Buses	-0.155	-1.1	ns	0.150	0.2	ns												
Taxi-rank	-1.926	-2.4	**	1.433	2.4	**												
Toilets										-0.330	-2.2	**	0.090	0.1	ns			
HHI							-5.283	-3.7	***									
ln(wact)							-0.334	-4.0	***									
Sample size (# trips)	5680									9366								

[a]Log normal distributions specified and inverse of variables expected to have negative coefficients entered into model
[b]Unstaffed removed from model as reference. ***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.13: Initial RPL models to identify variables with significant standard deviation.

the nearest station variable (0.63 vs. 0.94 in the WG model, and 0.37 vs. 0.81 in the LATIS model). This presumably reflects the ability of the RPL models to better explain choice decisions through the access journey variables as a result of individual-specific parameters.

6.5.2 Flow variant models

For the flow variant RPL models, the best performing flow variant MNL models were initially selected as the starting point (WG-FM6 and LATIS-FM8). The parameters specified as random were the mode-specific access time parameters, as identified in the trip end variant RPL models, and the parameters for the relevant train leg variables. The WG RPL model based on WG-FM6 failed to converge after 100 iterations, and a model using WG-FM1 as the starting point was estimated instead. The results of the flow variant RPL models (RPL4) are shown in Tables 6.14 and 6.16.

In the WG model, the SD of the three mode-specific parameters remained significant at the 1% level, and the SD of the parameter for train leg duration was also significant at the 1% level. When compared to the MNL model (WG-FM1), there was a very small improvement in the goodness of fit measures, and in the predictive performance measure (19.17% vs. 19.4%). In the LATIS model, the SD of the mode-specific access time parameters was no longer significant, and neither was the SD of the on-train time and waiting-time parameters. This model was virtually identical to the MNL equivalent, both in terms of the estimated parameters and the goodness of fit and performance measures.

6.6 Model appraisal

6.6.1 Predictive performance

Rather than use the fundamentally flawed 'percent correctly predicted' measure (this applies in all choice contexts, see Train (2009, p. 69) for a discussion), which assesses a model by assuming each individual would choose the station with the highest predicted probability and compares that to the station actually chosen, predictive performance was measured by comparing the sum of predicted probabilities for each station with the number of times that station was actually chosen (as preferred by Hensher et al. (2016, p. 502)). To assess the overall performance of the models reported in this thesis, the absolute difference between the two figures was summed for all stations and expressed as a percentage of the total number of choice situations in the model. A 'predictive performance difference' of zero percent would therefore indicate no deviation between observed and predicted choice. There is no theoretical upper limit to the measure. The predictive performance of the best models, as discussed in the previous sections, is summarised in Table 6.17. Given that the aim of this research is to improve on the simplistic models that assume the nearest station is chosen,

	WG trip end variant (RPL1) (TE31 starting model)												WG trip end variant (RPL2) (TE31 starting model)												WG flow variant (RPL4) (FM1 starting model)											
	Random parameters[a]									Non-random parameters			Random parameters[a]									Non-random parameters			Random parameters[a]									Non-random parameters		
variable	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*z*	*Sig*	*Med. B*	*Mean B*	*Std. Dev B*	*B*	*z*	*Sig*	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*z*	*Sig*	*Med. B*	*Mean B*	*Std. Dev B*	*B*	*z*	*Sig*	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*z*	*Sig*	*Med. B*	*Mean B*	*Std. Dev B*	*B*	*z*	*Sig*
Nearest station (yes)	-0.521	-1.9	*	0.17	0.1	ns	0.59	0.60	0.10													0.629	9.4	***										0.654	9.6	***
Time - walk (mins)	-1.666	-31.9	***	0.58	8.2	***	0.19	0.22	0.14				-1.631	-31.1	***	0.65	9.0	***	0.20	0.24	0.18				-1.650	-33.4	***	0.54	8.4	***	0.19	0.22	0.13			
Time - cycle (mins)										-0.187	-8.4	***										-0.184	-8.3	***										-0.217	-6.7	***
Time bus (mins)	-2.682	-18.3	***	1.21	10.9	***	0.07	0.14	0.26				-2.681	-24.7	***	0.55	8.8	***	0.07	0.08	0.05				-2.435	-22.9	***	0.58	10.5	***	0.09	0.10	0.07			
Time (car) mins	-1.363	-27.1	***	0.51	10.6	***	0.26	0.29	0.16				-1.394	-27.3	***	0.51	10.5	***	0.25	0.28	0.15				-1.214	-27.2	***	0.30	5.9	***	0.30	0.31	0.10			
Ln(frequency)										1.099	14.2	***										1.113	14.5	***										0.591	7.3	***
Unstaffed (yes)										-0.683	-7.2	***										-0.685	-7.2	***										-0.648	-6.6	***
CCTV (yes)										1.141	7.0	***										1.126	7.1	***										1.312	8.2	***
Car park spaces (#)										0.004	10.1	***										0.004	9.7	***										0.004	8.5	***
Free car park (yes)										0.675	5.3	***										0.658	5.3	***										0.845	6.7	***
Ticket machine (yes)										0.873	6.8	***										0.892	7.0	***										0.919	7.1	***
Buses										0.882	10.5	***										0.851	10.2	***										0.901	10.6	***
Taxi-rank										0.317	3.7	***										0.286	3.4	***										0.303	3.9	***
Train duration (mins)																									-2.646	-49.2	***	0.47	5.5	***	0.07	0.08	0.04			
HHI										-4.555	-3.4	***										-5.221	-4.0	***										-6.388	-5.0	***
ln(wact)										-0.309	-4.1	***										-0.328	-4.4	***												
Sample size (# trips)	5680												5680												5680											
Initial log-likelihood[b]	-13355												-13355												-13355											
Final log-likelihood	-3351												-3339												-3007											
McFadden's adjusted R2	0.75												0.75												0.77											
AIC	6740.50												6713.90												6052.80											
Predictive perf. diff. (%)	20.36												20.20												19.17											

[a]Log normal distributions specified and inverse of variables expected to have negative coefficients entered into model

[b]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen. ***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.14: RPL model results — WG.

	LATIS trip end variant (RPL1) (TE25 starting model)												LATIS trip end variant (RPL2) (TE25 starting model)												LATIS trip end variant (RPL3) (TE25 starting model)											
	Random parameters[a]									Non-random parameters			Random parameters[a]								Non-random parameters				Random parameters[a]								Non-random parameters		
variable	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*z*	*Sig*	*Med. B*	*Mean B*	*Std. Dev B*	*B*	*z*	*Sig*	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*z*	*Sig*	*Med. B*	*Mean B*	*Std. Dev B*	*B*	*z*	*Sig*	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*Sig*	*Med. B*	*Mean B*	*Std. Dev B*	*B*	*z*	*Sig*
Nearest station (yes)	-0.853	-2.4	**	0.414	0.5	ns	0.43	0.46	0.20													0.413	8.8	***									0.37	7.9	***
Time - walk (mins)	-1.940	-50.7	***	0.413	7.2	***	0.14	0.16	0.07				-1.952	-56.0	***	0.382	6.7	***	0.14	0.15	0.06				-1.934	-55.7	***	0.385	***	0.14	0.16	0.06			
Time - cycle (mins)	-1.785	-11.5	***	0.901	4.5	***	0.17	0.25	0.28				-2.110	-19.0	***	0.496	4.8	***	0.12	0.14	0.07				-2.111	-19.2	***	0.490	***	0.12	0.14	0.07			
Time PT (mins)	-2.773	-28.8	***	1.152	25.8	***	0.06	0.12	0.20				-2.570	-40.7	***	0.882	27.2	***	0.08	0.11	0.12				-2.544	-46.6	***	0.827	***	0.08	0.11	0.11			
Time (car) mins	-1.308	-33.6	***	0.781	19.7	***	0.27	0.37	0.34				-1.360	-38.6	***	0.771	21.8	***	0.26	0.00	0.31				-1.326	-38.6	***	0.760	***	0.27	0.35	0.31			
Ln(frequency)										0.914	19.7	***										0.984	22.4	***									0.96	22.0	***
Full-time (yes)b										2.414	17.4	***										2.061	16.2	***									2.07	16.3	***
Part-time (yes)b	0.055	0.5	ns	0.009	0.0	ns	1.06	1.06	0.01													0.817	8.1	***									0.80	7.9	***
CCTV (yes)										2.218	2.9	***										1.940	2.7	***									1.94	2.7	***
Car park spaces (#)	-7.040	-26.3	***	1.368	5.4	***	0.00	0.00	0.01				-6.783	-21.9	***	0.119	0.1	ns	0.00		0.00												0.00	8.5	***
Ticket machine (yes)										0.865	9.1	***										0.841	9.2	***									0.85	9.3	***
Toilets										0.509	5.3	***										0.639	6.8	***									0.65	7.0	***
On train time (mins)																																			
HHI																																			
ln(wact)																																			
Sample size (# trips)	9366												9366												9366										
Initial log-likelihood[c]	-21945												-21945												-21945										
Final log-likelihood	-6636												-6410												-6410										
McFadden's adjusted R2	0.71												0.71												0.71										
AIC	12847.10												12853.60												12851.90										
Predictive perf. diff. (%)	22.54												22.70												22.85										

[a]Log normal distributions specified and inverse of variables expected to have negative coefficients entered into model. [b]Unstaffed removed from model as reference.
[c]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen. ***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.15: RPL model results — LATIS (trip end variant).

	LATIS flow variant (RPL4) (FM8 starting model)											
	Random parameters[a]									Non-random parameters		
variable	*Mean ln(B)*	*z*	*Sig*	*Std. dev ln(B)*	*z*	*Sig*	*Med. B*	*Mean B*	*Std. Dev B*	*B*	*z*	*Sig*
Nearest station (yes)										0.732	16.4	***
Time - walk (mins)	-2.099	-73.4	***	0.001	0.0	ns	0.12	0.12	0.00			
Time - cycle (mins)	-2.135	-36.7	***	0.000	0.0	ns	0.12	0.12	0.00			
Time PT (mins)	-2.649	-152.1	***	0.002	0.0	ns	0.07	0.07	0.00			
Time (car) mins	-1.471	-67.0	***	0.002	0.0	ns	0.23	0.23	0.00			
Ln(frequency)										0.498	10.5	***
Full-time (yes)b										1.703	12.9	***
Part-time (yes)b										0.653	6.4	***
CCTV (yes)										1.838	3.5	***
Car park spaces (#)										0.001	8.3	***
Ticket machine (yes)										0.534	6.4	***
Toilets										0.473	5.0	***
On train time (mins)	-2.529	-81.2	***	0.003	0.0	ns	0.08	0.08	0.00			
Waiting-time (mins)	-1.938	-93.8	***	0.002	0.0	ns	0.14	0.14	0.00			
Bearing diff. (0-5km)										-0.004	-8.8	***
Bearing diff. (5-10km)										-0.007	-8.5	***
Bearing diff. (10-15km)										-0.010	-8.6	***
Bearing diff. (15-20km)										-0.013	-6.7	***
Bearing diff. (20+ km)										-0.017	-7.9	***
Ln(wact)										-0.160	-4.3	***
Sample size (# trips)	9366											
Initial log-likelihood[c]	-21945											
Final log-likelihood	-5122											
McFadden's adjusted R2	0.77											
AIC	10296.40											
Predictive perf. diff. (%)	14.43											

[a]Log normal distribution; inverse of variables entered where negative coefficients expected

[b]Unstaffed removed from model as reference.

[c]Initial LL assumes there is an equal probability of each alternative in a choice set being chosen.

***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.16: RPL model results — LATIS (flow variant).

the models are compared with a base model where the probability of choosing the nearest station is equal to one. The graphs in Figures 6.9 and 6.10 show the number of times each station was actually chosen and by how much the model under or over-predicted this choice, for the WG base model and WG-FM6, and similar graphs for the LATIS models are provided in Figures 6.11 and 6.12. These graphs clearly illustrate the substantially better predictive performance of the flow variant models compared to the base models.

An alternative method of viewing model predictive performance, on a local scale, is to overlay the under- and over-prediction for each station on a map. Figures 6.13 and 6.14 show the central Cardiff area with the under- and over-prediction represented as scaled bars positioned alongside each station, for the base model and WG-FM6 model respectively. Similar maps are shown for the Central Glasgow area in Figures 6.15 and 6.16. In both cities it is apparent

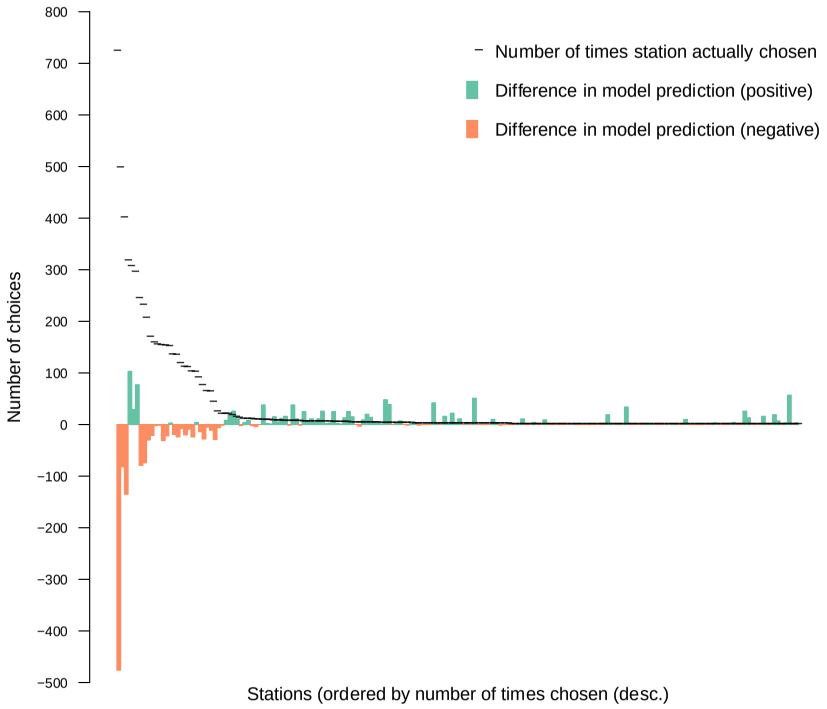


FIGURE 6.9: Model predictive performance — WG base model (nearest station probability = 1).

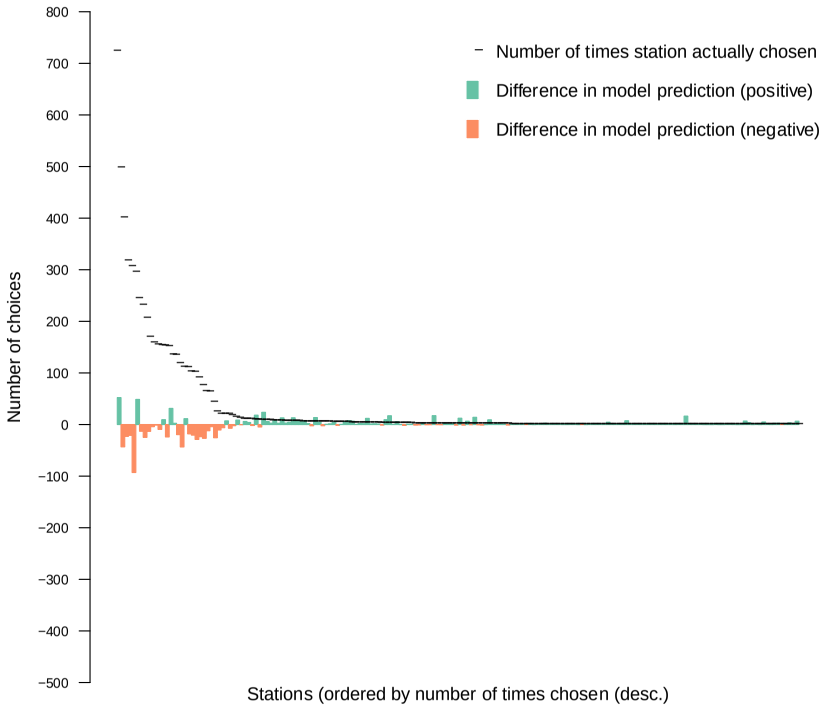


FIGURE 6.10: Model predictive performance — WG model FM6.

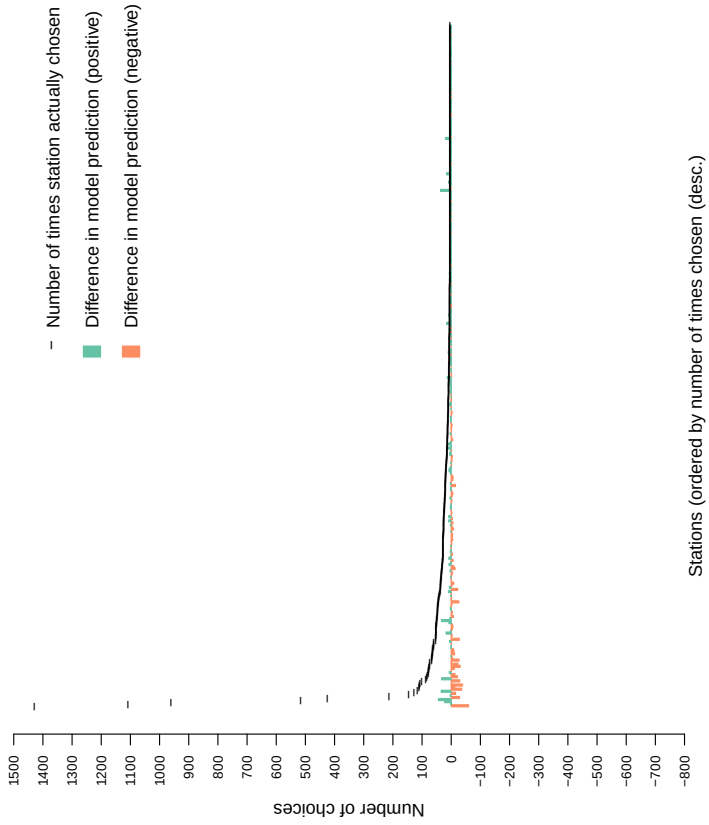


FIGURE 6.12: Model predictive performance — LATIS model FM6.

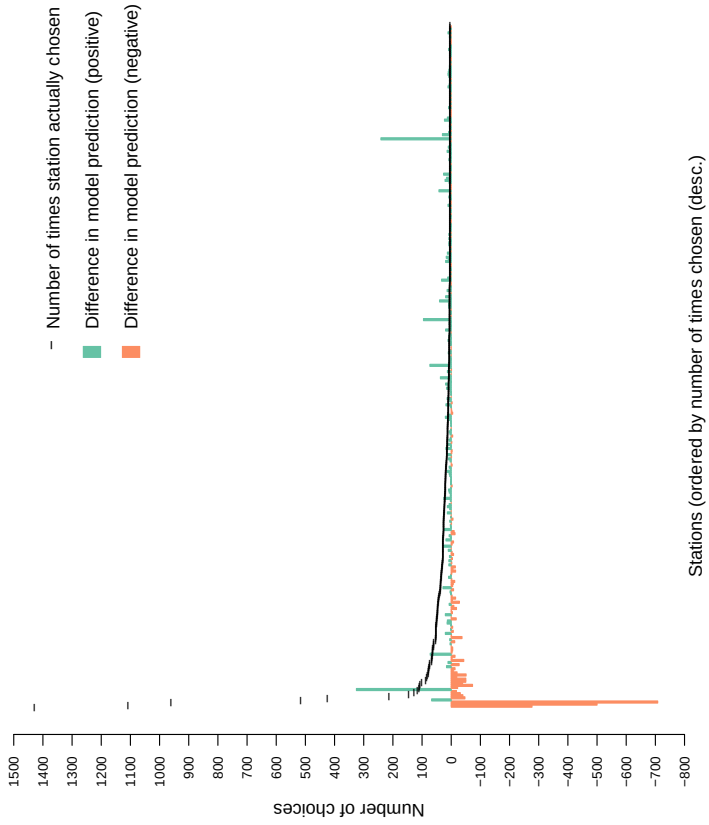


FIGURE 6.11: Model predictive performance — LATIS base model (nearest station probability = 1).

Model Type	WG		LATIS	
	Model	Predictive performance difference (%)	Model	Predictive performance difference (%)
Nearest station probability = 1	Base model	41.0	Base model	50.9
MNL trip-end variant	TE31	20.5	TE25	22.4
MNL flow variant (best performing)	FM6	19.0	FM7	14.2
MNL flow variant (comparator)	FM1	19.4	FM8	14.4
RPL trip-end variant	RPL2 (start TE31)	20.2	RPL3 (start TE25)	22.9
RPL flow variant	RPL4 (start FM1)	19.2	RPL4 (start FM8)	14.4
Transferability test (trip-end variant)	LATIS-TE25	34.5	WG-TE28	28.0
Transferability test (flow variant)	LATIS-FM5	33.1	WG-FM2	25.6

TABLE 6.17: Summary of station choice model predictive performance. Note: a lower value of the 'predictive performance difference' measure is better.

that the base model considerably under-predicts choice of the major stations (Cardiff Central (CDF), Glasgow Central (GLC) and Glasgow Queen Street (GLQ)), while over-predicting choice at nearby smaller stations. This problem is largely corrected by the station choice models, which is particularly encouraging given the very complex interaction of observed station catchments in these city centre locations (See Figures 4.28 and 4.29 in Chapter 4).

6.6.2 Transferability

One of the ultimate objectives of this research is to develop a generalised station choice model that is readily transferable and has wide applicability, rather than one that is restricted to application in the local context in which it was developed. A weakness of the predictive performance assessment reported above is that the models are validated against the sample that was used to calibrate them, which can result in an overly optimistic assessment of model performance. As an initial step to assess model transferability, the graph in Figure 6.17 plots the parameter estimates, along with the 95% and 99% confidence intervals, for the FM2 models[9]. The plot indicates reasonable correspondence of many of the parameters for shared variables, but also identifies potentially problematic variables, such as the provision of CCTV. This parameter has very wide confidence intervals in the LATIS model, and the large standard error may be due to the very high proportion of chosen stations (99.8%) that have CCTV installed. This could indicate that chosen stations have CCTV because nearly all stations have CCTV (96.1% of the alternatives in the LATIS dataset), and it may only be a factor that actually influences choice for a few observations.

[9]Model FM2 was selected for this exercise, as these are the most suitable for comparison — subsequent LATIS models include the 'bearing difference' variable which was inconsistent in the WG models. The HHI variable, which only appears in the WG model, is excluded from the plot for reasons of clarity, as its parameter has a large value relative to the other variables

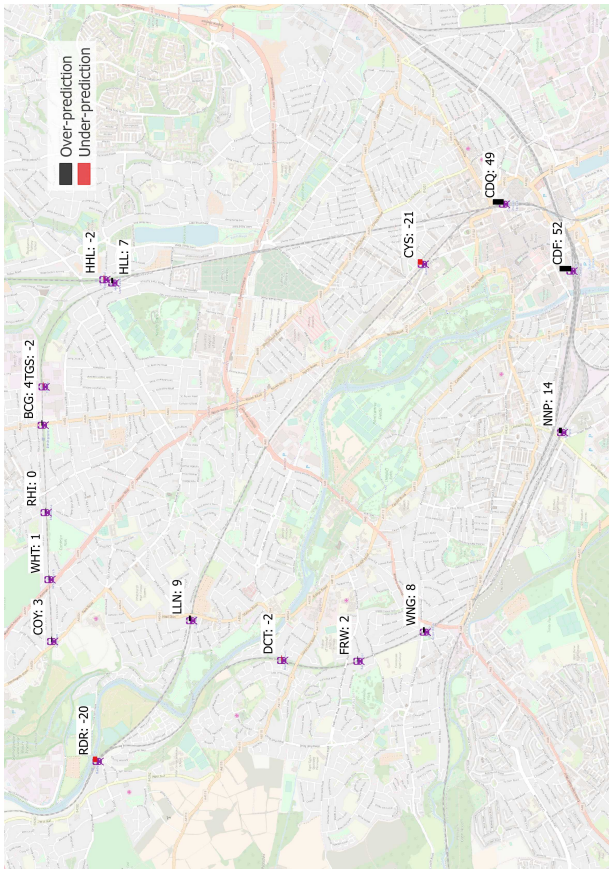


FIGURE 6.14: Central Cardiff stations — predictive performance WG model FM6.

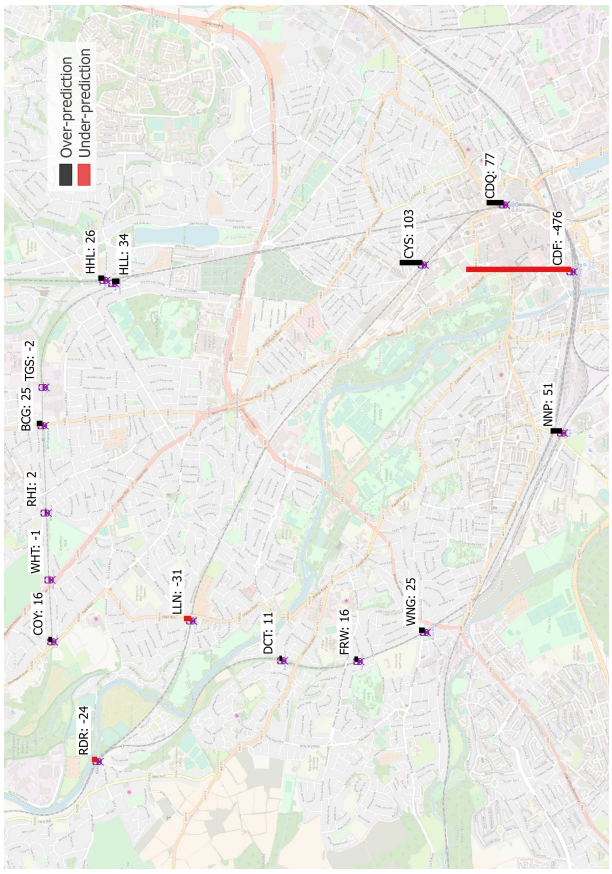


FIGURE 6.13: Central Cardiff stations — predictive performance WG base model.

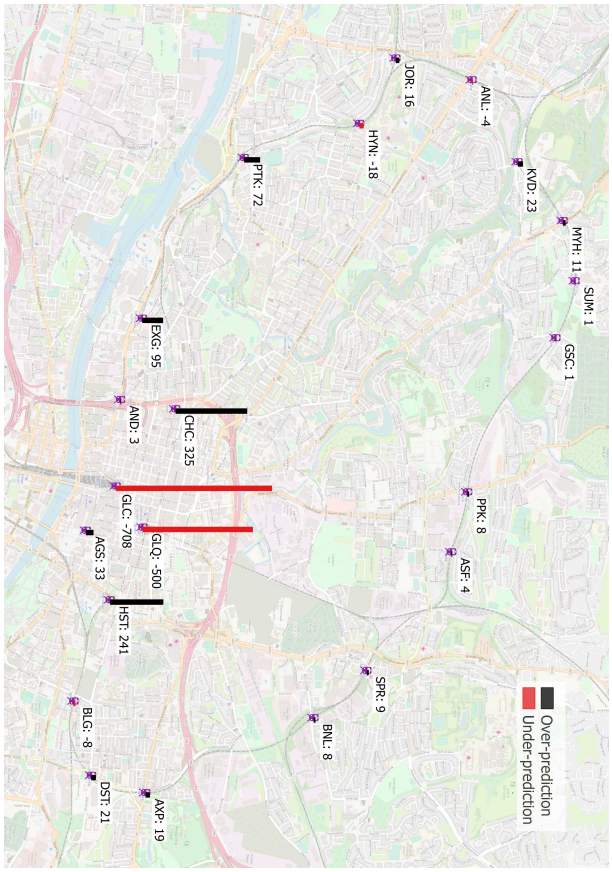


FIGURE 6.15: Central Glasgow stations — predictive performance LATIS base model.

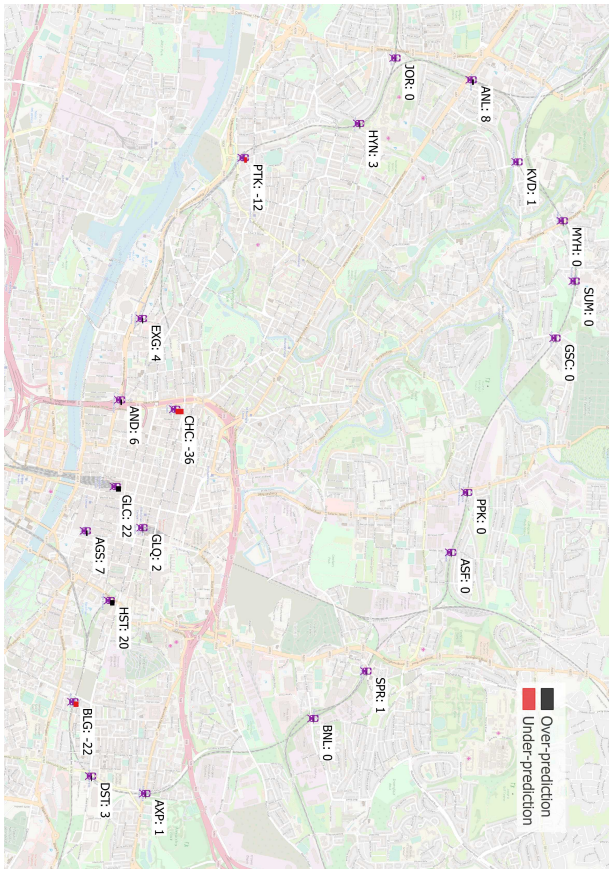


FIGURE 6.16: Central Glasgow stations — predictive performance LATIS model FM6.

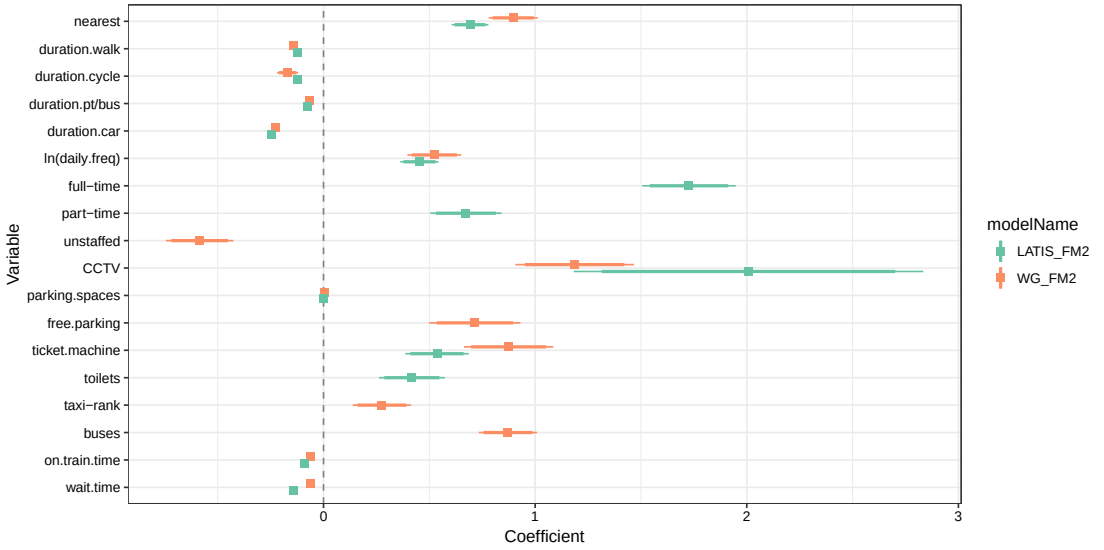


FIGURE 6.17: Parameter estimates for WG and LATIS model FM2 showing 95% and 99% confidence intervals.

To assess model transferability, the parameters from the WG-TE28 and WG-FM2 models were used to predict choice in the LATIS dataset; and parameters from the LATIS-TE25 and LATIS-FM5 models were used to predict choice in the WG dataset. The predictive performance of these models when applied to the alternative dataset are reported in Table 6.17. The WG-TE28 model performed quite well against the LATIS dataset, with a predictive performance of 28.0%, which compares favourably to the best in-sample trip end variant (LATIS-TE25: 22.4%). The WG-FM2 model performed slightly better, but its predictive performance was still below that of LATIS-TE25. Neither of the LATIS models performed particularly well against the WG dataset, with the predictive performance of LATIS-FM5 (33.1%) some way short of the predictive performance of the best in-sample trip end model (WG-TE31: 20.5%), although both of the models were an improvement over the base model.

6.7 Combined dataset models

This section is concerned with the calibration of station choice models that were specifically designed to be incorporated into a national-scale trip end model able to forecast demand for new local railway stations in GB, which is the subject of Chapter 7. These station choice models were calibrated using a combined dataset, formed by merging the WG and LATIS datasets, and as the trip end model methodology does not incorporate an access mode choice component, mode-specific access journey variables were not included.

The choice sets for these models were composed only of the 10 nearest stations. Unlike the earlier models, the nearest major station was not appended to the choice set (but may have been present as one of the nearest 10). This decision was based on concerns related to the accessibility term. By adding the nearest major station, the choice sets are no longer

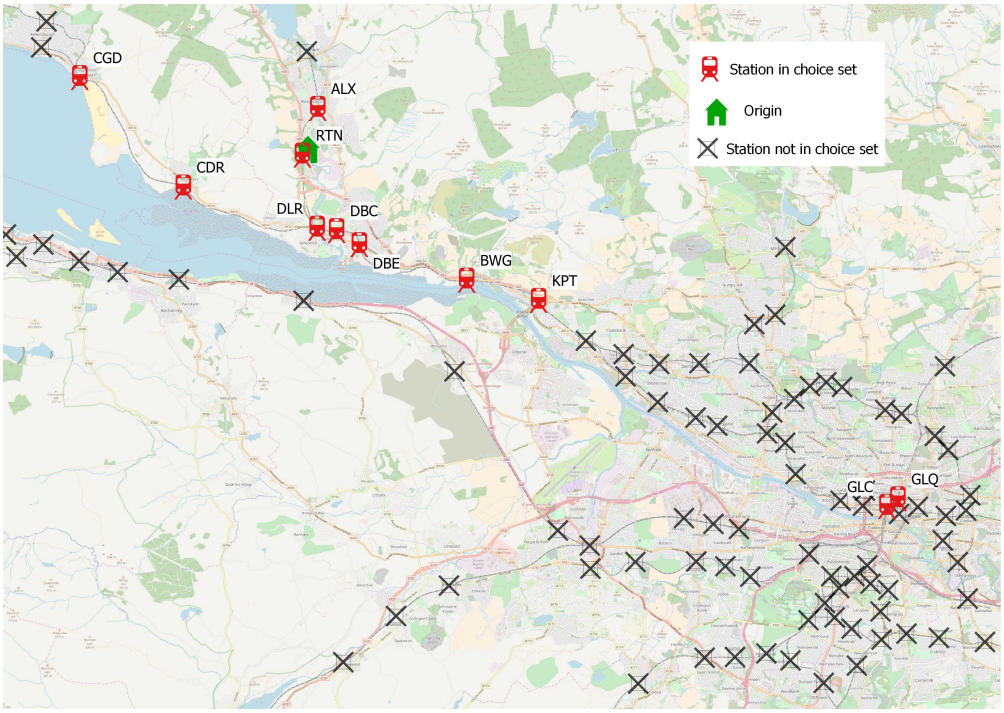


FIGURE 6.18: Example choice set where the nearest major stations (GLC and GLQ) have been appended.

representative of the true spatial relationships between stations, as all the stations that exist within the geographic area that encompasses the nearest 10 stations and the appended major station are not necessarily present in the choice set. This is illustrated in Figure 6.18, which shows a choice set where Glasgow Queen Street and Glasgow Central have been appended as the nearest major station(s). These two stations are surrounded by other stations that are in close proximity, but as far as the choice set is concerned they appear to be spatially isolated from the other stations, and this artificial spatial construct will impact the calculation of the accessibility term. As a major station that is appended to a choice set is likely to *appear* relatively isolated from other stations, *and* will only rarely be the chosen alternative (by definition as appending major stations results in a relatively small increase in the proportion of observed choice accounted for), this could impose an agglomeration effect on the model (a positive influence on θ) which moderates an otherwise underlying competition effect.

6.7.1 Model calibration

The first set of models calibrated using the combined dataset were aimed at improving the representation of the access journey, given that mode-specific access time variables were no longer included. The results of these models are summarised in Table 6.18. The best initial model was CMB-TE3, which included the nearest station (by distance) dummy variable and access distance (adjusted rho-squared: 0.51). This model was improved by transforming the access distance variable, with a square root transformation performing

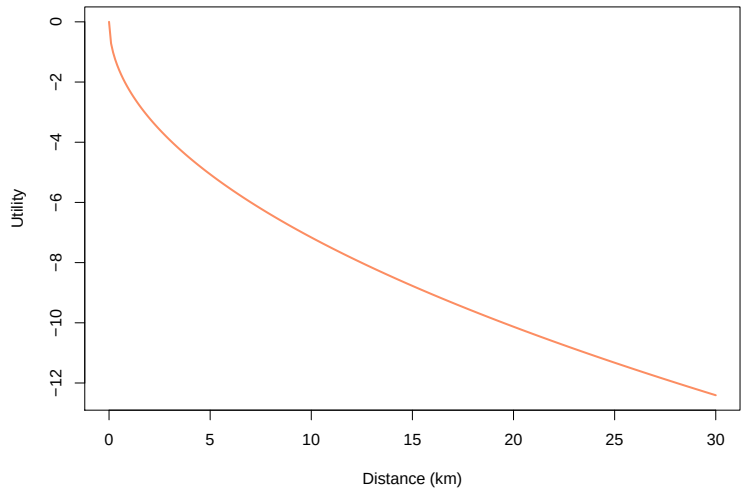


FIGURE 6.19: Utility associated with square root of access distance (0–30 km) using estimated coefficient −2.26517 (from model CMB-TE24).

slightly better (CMB-TE4: adjusted rho-squared: 0.55) than a log-normal transformation (CMB-TE5: adjusted rho-squared: 0.54). Both of these models were better in terms of model fit and the predictive performance measure than similar models estimated using access time (CMB-TE6 to CMB-TE8).

The predictive performance of model CMB-TE4 (56.5%) was not dissimilar to that achieved by the separate LATIS and WG models with mode-specific access time variables and the nearest station dummy (WG-TE12: 56.5%; LATIS-TE12: 62.0%). In Figure 6.19, the implied (dis)utility of access distance, when using a square root transformation and applying the estimated coefficient of −2.26517, is plotted over a distance of 30 km. This shows that disutility increases more rapidly over shorter access distances. The disutility of walking for 30 minutes (at 3 mph) is −3.516, and of driving for 30 minutes (at 30 mph) is −11.13; implying an average disutility per km travelled of −1.46 and −0.46 respectively. These figures are higher than the distance-based parameter estimates for walk and car modes obtained from models WG-TE5 (−1.05 and −0.21) and LATIS-TE5 (−0.88 and −0.13), but not hugely dissimilar. Given that it is much more likely that shorter access distances will be walked, and longer access distances will be by a motorised mode, this model does appear able to capture, to a certain extent, a mode-specific element.

In the subsequent models (CMB-TE10 to CMB-TE20), shown in Tables 6.19 and 6.20, the same service and facilities variables were tested as in the separate dataset models, with the exception of the staffing-level variables. When the data on designated staffing level for stations in England[10] (obtained from the NRE knowledgebase) was reviewed, it was found to

[10]The calibrated choice models would need to be applied throughout GB, so while the staffing level data appeared reliable for stations in Scotland and Wales (that formed the calibration dataset) a variable that was accurate across the country was preferable.

	CMB-TE1			CMB-TE2			CMB-TE3			CMB-TE4			CMB-TE5			CMB-TE6			CMB-TE7			CMB-TE8		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	2.993	165.9	***				1.366	48.2	***	0.578	18.3	***	0.423	12.4	***	1.275	45.8	***	0.992	33.9	***	1.111	36.6	***
Nearest by time				2.809	160.7	***																		
Distance							-0.367	-50.9	***															
Sqrt(distance)										-2.290	-65.5	***												
Ln(distance)													-2.016	-68.8	***									
Time																-0.217	-57.8	***						
Sqrt(time)																			-1.498	-65.5	***			
Ln(time)																						-1.894	-62.0	***
Sample size (# trips)	14422			14422			14422			14422			14422			14422			14422			14422		
Initial log-likelihood[a]	-33025			-33025			-33025			-33025			-33025			-33025			-33025			-33025		
Final log-likelihood	-18609			-20301			-16048			-14919			-15228			-15914			-15600			-16107		
McFadden's adjusted R^2	0.44			0.39			0.51			0.55			0.54			0.52			0.53			0.51		
AIC	37220			40604			32099			29841			30460			31832			31205			32219		
Predictive perf. diff. (%)	65.5			74.8			59.7			56.5			56.9			59.8			59.7			60.8		

[a]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.

***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.18: Results of station choice MNL models — combined dataset (1 of 3).

be unreliable. For example, stations known to be unstaffed were recorded as having full-time staff. It was therefore decided to use the 'Category F' Network Rail station category, which only includes unstaffed stations, as a proxy for staffing level. This variable was compared to the unstaffed variable (models CMB-TE10 and CMB-TE11) and the estimated parameter and model performance were found to be very similar.

The best performing model, prior to including the accessibility term (discussed below) was CMB-TE19, with an adjusted rho-squared of 0.71 and predictive performance measure of 24.9%. By comparison, the predictive performance of the base model, where the nearest station has a probability of one, was 42.2%. The utility function for model CMB-TE19, for individual n at origin i choosing station k is given by the following formula:

$$V_{nik} = \exp(\beta N_k + \gamma \sqrt{D_{ik}} + \delta U_k + \epsilon \ln F_k + \zeta C_k + \eta Ps_k + \theta T_k + \iota B_k), \tag{6.10}$$

where D is the access distance by road from origin i to station k; F is the daily service frequency at station k; Ps is the number of car parking spaces at station k; N, U, C, T and B are dummy variables that take the value of 1 if station i is the nearest station, unstaffed, has CCTV, has a ticket machine, or has a bus interchange respectively, and zero otherwise; and β, γ, δ, ϵ, ζ, η, θ, and ι are the estimated parameters.

The HHI variable was not tested in the combined dataset models. This is because the download of POI data from the EDINA Digimap service is restricted to a maximum area of 10,000 km^2. It is therefore not possible to download all the POIs for mainland GB. While the download limit would be sufficient to obtain the POIs within a 400 m^2 buffer of every station, the buffers would have to be defined separately for each station and added to the download basket one at a time. Given the mixed results obtained using this variable in the WG and LATIS models, it was felt that the available time should be allocated to higher priority tasks.

6.7.2 Accessibility term

The accessibility term incorporates a weighting, which is the annual number of station entries and exits. Clearly, this figure will not be known for proposed new stations, as it forms the dependent variable in the trip end demand model. Models with three variants of the accessibility term were therefore tested. In the first (CMB-TE21), the weight was defined as the total number of entries and exits at the station in 2014/15. In the second (model CMB-TE22), the median number of trip entries/exits for each station category (excluding stations in Inner London) was used; and in the third (model CMB-TE24) a fixed weight for each station category was chosen, based on the thresholds specified in the category definitions (see Green and Hall (2009, Annex C)), as shown in Table 6.21. The logarithmic transformation of the accessibility term was added to each of the models, as suggested by Fotheringham,

	CMB-TE10			CMB-TE11			CMB-TE12			CMB-TE13			CMB-TE14			CMB-TE15			CMB-TE16			CMB-TE17		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest distance	0.617	18.4	***	0.612	18.4	***	0.690	19.0	***	0.677	18.7	***	0.717	19.4	***	0.718	19.4	***	0.709	19.0	***	0.711	19.1	***
Sqrt(distance)	-2.139	-59.1	***	-2.134	-59.4	***	-2.258	-57.3	***	-2.278	-57.7	***	-2.253	-56.3	***	-2.252	-56.3	***	-2.271	-56.6	***	-2.269	-56.4	***
Category F	-2.124	-63.1	***				-0.733	-18.1	***	-0.681	-16.8	***	-0.760	-18.4	***	-0.759	-18.4	***	-0.759	-18.3	***	-0.741	-16.3	***
Unstaffed				-2.231	-63.4	***																		
Ln(service frequency)							1.714	59.1	***	1.698	58.3	***	1.416	42.4	***	1.417	42.4	***	1.252	36.2	***	1.235	32.2	***
CCTV										1.070	9.1	***	1.151	9.8	***	1.152	9.8	***	0.986	8.2	***	0.998	8.3	***
Car parking spaces													0.001	15.9	***	0.001	15.9	***	0.001	16.2	***	0.001	16.2	***
Free car park																0.068	0.7	ns						
Ticket machine																			0.960	18.9	***	0.951	18.5	***
Toilets																						0.047	1.0	ns
Sample size (# trips)	14422			14422			14422			14422			14422			14422			14422			14422		
Initial log-likelihood[a]	-33025			-33025			-33025			-33025			-33025			-33025			-33025			-33025		
Final log-likelihood	-12376			-12244			-10116			-10068			-9937			-9937			-9747			-9747		
McFadden's adjusted R^2	0.63			0.63			0.69			0.69			0.70			0.70			0.70			0.70		
AIC	24757			24494			20240			20146			19887			19889			19508			19509		
Predictive perf. diff. (%)	44.7			44.0			29.0			28.6			27.5			27.5			25.7			25.7		

[a]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.

***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.19: Results of station choice MNL models — combined dataset (2 of 3).

	Nearest main NOT added to choice set																		Nearest main added to choice set					
	CMB-TE18			CMB-TE19			CMB-TE20			CMB-TE21			CMB-TE22			CMB-TE24			CMB-MN-TE12			CMB-MN-TE14		
Variable	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*	*B*	*z*	*Sig*
Nearest by distance	0.698	18.8	***	0.691	18.4	***	0.685	18.3	***	0.686	18.3	***	0.691	18.4	***	0.691	18.4	***	0.972	29.5	***	0.951	28.9	***
Sqrt(distance)	-2.302	-56.9	***	-2.262	-56.3	***	-2.266	-56.5	***	-2.268	-56.4	***	-2.265	-56.4	***	-2.265	-56.4	***	-1.836	-65.0	***	-1.864	-64.6	***
Category F	-0.828	-19.6	***	-0.677	-16.0	***	-0.701	-16.4	***	-0.652	-15.3	***	-0.636	-14.4	***	-0.638	-14.5	***	-0.582	-14.8	***	-0.624	-15.5	***
Ln(service frequency)	1.364	36.5	***	1.199	34.6	***	1.207	34.8	***	1.210	34.7	***	1.216	34.6	***	1.214	34.6	***	1.332	43.8	***	1.288	41.1	***
CCTV	0.956	7.9	***	1.071	8.6	***	1.079	8.7	***	1.054	8.4	***	1.070	8.6	***	1.076	8.6	***	0.856	7.5	***	0.897	7.8	***
Car parking spaces	0.001	15.5	***	0.001	16.5	***	0.001	16.8	***	0.001	9.6	***	0.001	11.7	***	0.001	13.7	***	0.001	13.4	***	0.001	13.8	***
Ticket machine	0.919	18.0	***	0.984	19.1	***	0.969	18.7	***	0.976	18.9	***	0.975	18.9	***	0.963	18.6	***	0.971	20.6	***	0.973	20.6	***
Waiting room	-0.457	-8.8	***																					
Bus interchange				0.758	13.6	***	0.814	14.1	***	0.735	13.1	***	0.733	13.0	***	0.731	13.0	***	0.616	11.9	***	0.644	12.3	***
Taxi rank							-0.186	-3.6	***															
ln(accessibility term)[1]										-0.131	-3.7	***										0.159	5.4	***
ln(accessibility term)[2]													-0.069	-3.1	***									
ln(accessibility term)[3]																-0.141	-3.2	***						
Sample size (# trips)	14422			14422			14422			14422			14422			14422			15047			15047		
Initial log-likelihood[a]	-33025			-33025			-33025			-33025			-33025			-33025			-35300			-35300		
Final log-likelihood	-9709			-9651			-9645			-9644			-9646			-9646			-11707			-11693		
McFadden's adjusted R^2	0.71			0.71			0.71			0.71			0.71			0.71			0.67			0.67		
AIC	19433			19318			19307			19306			19310			19309			23430			23403		
Predictive perf. diff. (%)	25.2			24.9			24.7			24.4			24.5			24.4			28.4			29.0		

[a]Initial log-likelihood assumes there is an equal probability of each alternative in a choice set being chosen.

[1]weighted by entries/exits; [2]weighted by median entries/exits for station category; [3]fixed weight for station category applied

***, **, * indicate significance at 1%, 5%, 10% level

TABLE 6.20: Results of station choice MNL models — combined dataset (3 of 3).

with the utility function for individual n at origin i choosing station k becoming:

$$V_{nik} = \exp(\beta N_k + \gamma\sqrt{D_{ik}} + \delta U_k + \epsilon \ln F_k + \zeta C_k + \eta Ps_k + \theta T_k + \iota B_k + \kappa \ln A_k), \qquad (6.11)$$

where A is the accessibility term, and κ the associated parameter to be estimated.

The models incorporating the accessibility term performed slightly better than the best model without (CMB-TE19), with higher log-likelihood and lower AIC values. The parameter for the accessibility term was negative and significant at the 1% level in all three models, indicating that a competition effect is at play. The estimated parameter in CMB-TE24, which uses the fixed weights, was very similar to the parameter in model CMB-TE21, which uses actual entries/exits (−0.141 and −0.131 respectively). This suggests that the fixed category-specific weight is a suitable proxy for the actual number of entries and exits. As the trip end models are only intended to predict demand at new local stations, which are defined as Category E or F, the appropriate weight will always be known for any proposed new station (given that category F stations are unstaffed).

Station category	Median entries/exits (2014/15)	Chosen fixed weight
A	14,870,920	2,000,000
B	4,498,966	2,000,000
C	1,886,992	1,000,000
D	828,660	500,000
E	330,295	250,000
F	52,486	125,000

TABLE 6.21: Alternative derived weights for each main station category, used in the accessibility term.

6.7.3 Models with nearest major station appended to choice set

For purposes of comparison, the same calibration process was repeated using choice sets with the nearest major station appended. The final models, with and without the accessibility term, are shown in Table 6.20 (CMB-MN-TE12 and CMB-MN-TE14). The accessibility term, while significant at the 1% level, had a positive sign, indicating an agglomeration rather than competition effect. This would appear to justify the decision not to use choice sets with the nearest major station appended, due to them not representing the true spatial relationships between stations. The models with the nearest major station appended were also inferior in terms of the predictive performance measures, with 28.4% for model CMB-MN-TE12, compared to 24.9% for model CMB-TE19. This is not surprising given the additional difficulty of explaining the choice of a more distant station, especially when no account is taken of components of the train leg in these models, such as fewer transfers or a faster overall journey time.

6.7.4 Assessing model predictive accuracy

The predictive performance of the best combined dataset models reported in Table 6.20 represent an in-sample assessment against the data that was used to fit (train) the models. However, due to inevitable idiosyncrasies ('noise') of the training dataset and the risk of over-fitting, a predictive model will nearly always perform less well against a new dataset; a phenomenon known as 'validity shrinkage' or 'training optimism' (Fortmann-Roe, 2018; Ivanescu et al., 2015). In order to quantify the extent of this problem two techniques were adopted: a *k*-fold cross-validation; and application of the model(s) to data from the LATIS 2013 survey (which was not used in model calibration[11]).

By combining these two methods the shortcomings of each can be overcome and a more comprehensive understanding of the likely predictive performance of the model(s) on new data is possible. The LATIS 2013 dataset was relatively small (1,190 choice situations) and as predictive accuracy assessed against a single independent sample is subject to high variability, assessment of the model(s) against another survey might give quite different results. While a single *k*-fold cross-validation is also subject to high variability, it can be repeated multiple times enabling the stability of the model to be assessed and an average estimate of the model accuracy to be calculated (Vanwinckelen & Blockeel, 2012). An advantage of validating against an independent survey is the ability to consider the problem of an individual's observed choice not necessarily appearing in the researcher-defined choice set, thus allowing this additional cause of validity shrinkage associated with choice models to be investigated. As it was considered important that the maximum amount of information was available to the models during calibration, the *k*-fold cross-validation technique was chosen in preference to using a holdout sample.

6.7.4.1 *k*-fold cross-validation

In *k*-fold cross-validation the dataset is randomly divided into *k* (typically 5 or 10) equally sized subsets, known as folds. Each fold is, in turn, excluded from the dataset and the model is estimated on the remaining folds. The estimated model is then applied to the excluded fold and the desired measure of predictive performance is calculated. Each fold therefore acts as the validation dataset once. The average of the *k* predictive performance measures is considered to be an estimate of the predictive accuracy of the model. There is potential for high variance in this estimate, as a second *k*-fold cross-validation, with a different random division into folds, could produce a very different result. This can be investigated by performing repeated cross-validations, with an average of the estimates from each repeat taken as the predictive accuracy of the model (Vanwinckelen & Blockeel, 2012).

[11] Although the data from the 2013 survey was processed along with that of 2014 and 2015, when the choice sets were compiled observations from the 2013 survey were excluded. This was to enable a common universal set of stations to be defined from which the alternatives for each choice set were selected (several new stations were opened subsequent to the 2013 survey).

A 10-fold cross-validation repeated 10 times was completed for models CMB-TE24 and CMB-TE19 (the 'best' models with and without the accessibility term). The `sample()` function in R was used to allocate each choice situation in the dataset to a fold. This was repeated ten times. A procedure was written in NLOGIT to automate the process of estimating the model on $k-i$ folds and calculating the choice probabilities for fold i. For comparison, the same fold and repeat structure was used to calculate predictive accuracy of the base model (i.e. the predictive performance of each fold was estimated on the basis that the probability of the nearest station being chosen was one). The results for each fold and each repeat are summarised in Tables 6.22, 6.23 and 6.24. In addition to showing the accuracy estimate for each repeat (CV pred. perf. %) and the average of this estimate for all repeats, these tables include several summary measures (mean, maximum, and standard deviation) of the absolute difference (between sum of actual choice and sum of probabilities) for each station in the dataset.

	Predictive performance difference (%) of each fold											Summary measures of station absolute difference		
Rpt.	1	2	3	4	5	6	7	8	9	10	CV pred. perf. (%)	Mean	Max	Sd
1	26.90	25.02	27.98	28.49	26.44	27.75	29.60	32.17	28.96	28.35	28.17	8.01	236.67	19.08
2	29.12	27.70	29.58	28.65	28.45	28.89	26.55	29.43	27.75	28.22	28.43	8.09	236.90	19.15
3	28.72	26.70	31.48	23.76	28.97	27.45	26.60	28.30	30.95	29.72	28.26	8.04	236.78	19.12
4	29.54	27.06	28.04	28.97	29.41	28.22	29.78	26.99	26.67	29.70	28.44	8.09	236.73	18.98
5	29.84	28.83	28.78	27.60	29.94	28.08	28.77	28.12	28.03	27.09	28.51	8.11	236.75	19.05
6	28.45	30.00	27.81	28.97	26.82	27.33	28.07	31.77	27.68	26.80	28.37	8.07	236.97	18.97
7	27.59	28.38	27.45	28.37	31.68	26.28	27.42	28.72	27.71	27.50	28.11	8.00	236.90	18.99
8	29.95	28.67	26.20	26.67	30.22	30.66	26.80	28.44	25.69	28.00	28.13	8.00	236.82	19.07
9	26.37	29.34	27.65	26.89	28.64	30.37	26.40	28.13	28.52	30.82	28.31	8.05	236.93	18.95
10	29.40	26.28	27.91	28.86	30.80	25.66	28.34	29.15	27.54	28.22	28.22	8.03	236.68	18.99
										Average of all repeats	28.30	8.05	236.81	19.04

TABLE 6.22: Summary of the predictive performance difference (%) for 10-fold cross validation of model CMB-TE24 repeated 10 times.

The results show that the average predictive performance measure of all repeats is 28.3% for model CMB-TE24, which is marginally better than the 28.6% for model CMB-TE19. This represents a small reduction in model predictive performance, of 3.9 and 3.7 percentage points respectively, compared to the in-sample assessment. It should be noted that there are potential sources of both pessimistic and optimistic bias to this estimate of predictive performance. As the model is only ever calibrated on a maximum of 90% of the choice situations in the dataset it is likely to be slightly less accurate than a model calibrated on the full dataset, and therefore pessimistically biased (Vanwinckelen & Blockeel, 2012). However, as the *full* dataset was used to select the predictor variables and identify the 'best' model(s), there is also potential for optimistic bias as information from the excluded folds informed this procedure (Ivanescu et al., 2015).

	Predictive performance difference (%) of each fold											Summary measures of station absolute difference		
Rpt.	1	2	3	4	5	6	7	8	9	10	CV pred. perf. (%)	Mean	Max	Sd
1	26.88	25.07	28.40	28.99	26.75	28.13	30.02	32.57	28.96	28.66	28.44	8.09	257.28	19.58
2	29.55	28.25	29.57	29.08	28.79	29.31	26.97	29.90	27.69	28.42	28.75	8.18	257.42	19.71
3	29.16	26.81	31.89	24.20	29.48	27.78	26.99	28.42	31.33	30.19	28.62	8.14	257.28	19.68
4	30.01	27.02	28.46	29.40	29.80	28.65	30.03	27.48	26.56	30.13	28.76	8.18	257.22	19.52
5	30.25	29.21	28.72	28.01	30.39	28.57	29.18	28.58	28.31	27.24	28.85	8.21	257.37	19.62
6	28.92	30.31	28.05	29.42	27.19	27.77	28.51	31.95	28.19	27.09	28.74	8.18	257.41	19.53
7	27.88	28.73	27.44	28.38	31.90	26.67	27.84	28.84	28.21	27.97	28.39	8.07	257.32	19.50
8	30.28	28.67	26.66	27.12	30.58	30.91	27.24	28.87	25.82	27.99	28.41	8.08	257.34	19.56
9	26.51	29.58	28.08	27.32	29.07	30.85	26.82	28.63	28.88	30.79	28.65	8.15	257.42	19.50
10	29.53	26.35	28.38	29.34	31.16	26.01	28.79	29.56	28.01	28.61	28.57	8.13	257.25	19.53
								Average of all repeats			28.62	8.14	257.33	19.57

TABLE 6.23: Summary of the predictive performance difference (%) for 10-fold cross validation of model CMB-TE19 repeated 10 times.

	Predictive performance difference (%) of each 'fold'											Summary measures of station absolute difference		
'Rpt'	1	2	3	4	5	6	7	8	9	10	Avg. pred. perf. (%)	Mean	Max	Sd
1	44.94	43.69	46.32	44.45	39.81	45.91	48.27	45.70	46.05	42.87	44.80	12.74	483	38.42
2	44.73	44.80	47.71	45.42	45.84	45.91	43.90	45.84	41.75	45.01	45.09	12.83	483	38.48
3	47.64	42.93	46.67	42.86	45.63	47.85	41.75	45.91	47.64	44.67	45.35	12.90	483	38.44
4	45.15	41.47	46.39	46.88	47.57	44.59	46.53	42.86	45.01	42.52	44.90	12.77	483	38.45
5	45.77	44.59	46.32	45.84	44.38	46.88	46.32	44.04	45.21	43.49	45.29	12.88	483	38.51
6	50.21	45.35	41.68	46.67	43.48	43.48	44.80	47.78	46.46	43.77	45.37	12.91	483	38.43
7	43.97	45.42	43.76	44.31	48.47	44.38	44.52	44.80	43.07	45.43	44.81	12.75	483	38.45
8	47.16	44.31	45.77	44.45	43.48	45.42	45.63	42.86	45.42	48.89	45.34	12.90	483	38.48
9	43.34	47.57	41.89	43.48	45.77	48.68	44.80	47.78	42.72	43.77	44.98	12.79	483	38.47
10	47.30	44.24	42.30	44.66	46.53	43.27	44.04	44.45	44.52	48.20	44.95	12.79	483	38.42
								Average of all 'repeats'			45.09	12.83	483	38.46

TABLE 6.24: Summary of the predictive performance difference (%) for the base model (probability of nearest station being chosen equals one) calculated for the same fold and repeat structure as the *k*-fold cross validation.

There is very low variance in the average predictive performance measure between repeats, with a maximum difference of 0.4 for both models, indicating a high level of model stability. Both models perform considerably better than the base model in terms of their overall predictive performance (base model: 45.1%) and in terms of the summary measures of station absolute difference, with the lower mean difference accompanied by a substantially smaller maximum difference and standard deviation. The predictive performance of the base model and CMB-TE24 for each individual station is shown in Figures 6.20 and 6.21 (based on the first cross-validation repeat with probabilities summed across the folds). The Exhibition Centre station in Glasgow is marked in these figures, providing an example of a station that was only chosen once in the dataset and substantially over-predicted by the base model; an issue largely corrected by model CMB-TE24.

6.7.4.2 Validation using an independent sample

The predictive accuracy of several models was assessed against a survey carried out by LATIS in 2013. The data from this survey were prepared along with the data from the 2014 and 2015 surveys, as described in Chapter 4. The validated dataset contained 1,190 choice situations and was based on interviews carried out in early February. While the interviews were conducted across Scotland, they were concentrated in the Highlands and Moray, areas that were under-represented in the 2014 and 2015 surveys (see Figure 6.22). Choice sets were prepared in the same manner as those for the WG and LATIS 2014 and 2015 datasets (as described in Section 6.3), both with and without the nearest major station being appended (if not already present).

In order to obtain an unbiased assessment of the predictive accuracy of the models, any choice situations where the chosen station was not present in the choice set were removed (the impact of missing chosen stations on model validity will be considered in due course). A summary of the predictive performance of the models when applied to the 2013 dataset, along with a comparator base model, is shown in the left-hand side of Table 6.25. Results based on choice sets compiled with and without the nearest major station appended are included. For the former, the predictive performance measures for models CMB-TE19 and CMB-TE24 are very similar (23.11% and 23.30% respectively) and a noticeable improvement over the estimate from the cross-validation exercise (around 28%). The base model has also performed much better against this dataset (30.48% compared to 45.09%). The improvement in the base model can be explained by the much higher proportion of choice situations where the nearest station was chosen (82% compared to 69% for the combined dataset), and this is also likely to account for the better performance of the other models. Nevertheless, it is reassuring that models CMB-TE19 and CMB-TE24 still out-perform the base model by seven percentage points when the proportion of observations choosing their nearest station is so high. When choice sets with the nearest major station appended are used for the assessment, there is a noticeable deterioration in predictive performance. This would be expected given

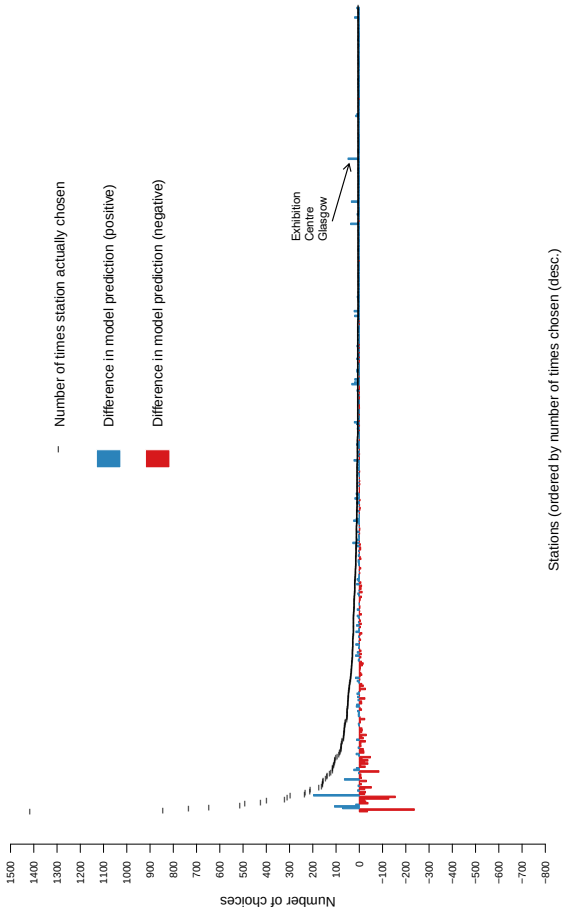


FIGURE 6.21: Predictive performance based on repeat 1 of the k-fold cross validation with station probabilities summed across all folds.

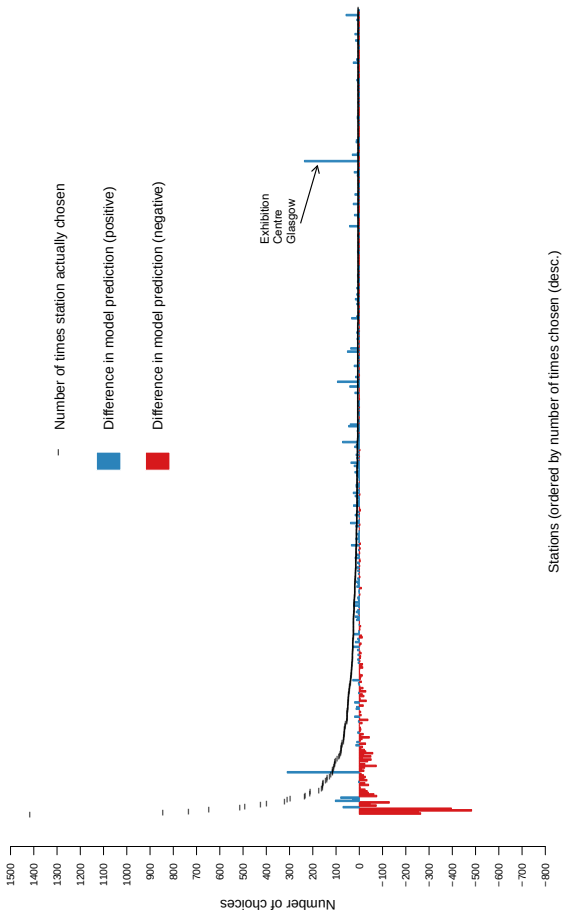
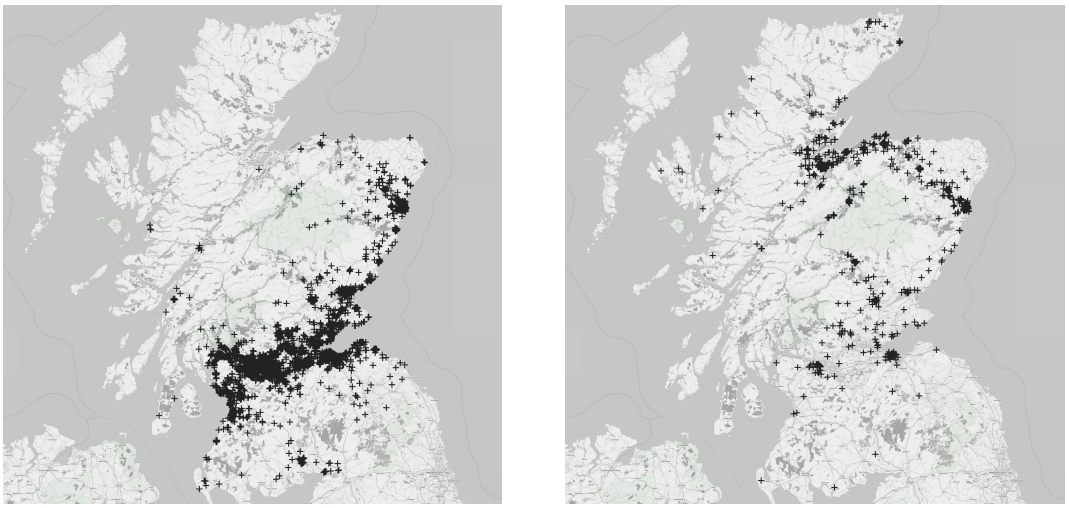


FIGURE 6.20: Model predictive performance - combined base model (nearest station probability = 1).



(A) 2014 and 2015 surveys (B) 2013 survey

FIGURE 6.22: Trip origins for LATIS surveys.

the poorer in-sample performance of model CMB-MN-TE12 (see Table 6.20) and the absence of variables related to the train-leg in this model, such as on-train time and waiting time, without which it will struggle to adequately account for long access journeys to board at a major station[12].

		Analysis includes only choice sets where chosen alternative present					Analysis includes all choice sets (absolute difference adjusted)				
		Major station not appended			Major station appended		Major station not appended			Major station appended	
		TE19	TE24	Base	TE12	Base	TE19	TE24	Base	TE12	Base
Choice situations		1073	1073	1073	1142	1142	1190	1190	1190	1190	1190
Measures of station absolute difference	Sum	248	250	327	372	465	447	443	521	439	521
	Mean	1.39	1.40	1.84	1.40	1.75	1.52	1.50	1.77	1.49	1.77
	Max	68	67	80	112	126	117	116	129	115	129
	Sd	5.80	5.75	6.76	7.59	8.46	7.81	7.75	8.57	7.76	8.57
Pred. Perf. (%)		23.11	23.30	30.48	32.57	40.72	37.56	37.23	43.78	36.89	43.78

TABLE 6.25: Summary of predictive performance of combined station choice models and comparator base models against 2013 LATIS survey.

For the analysis based on choice sets without the nearest major station appended, a total of 117 choice situations were removed because the chosen station was not in the choice set, representing 9.8% of the total. When the nearest major station was appended to the choice sets it was only necessary to remove 48 choice situations, representing 4% of the total. Table 6.26 summarises the chosen stations that were missing from the choice sets in each

[12]An important effect of appending the nearest major station without incorporating train-leg variables is to reduce the size of the negative parameter for access distance (from −2.265 in CMB-TE24 to −1.836 in CMB-MN-TE12). In comparison, when the train-leg variables were introduced into the separate LATIS and WG models (which have the nearest major station appended), the size of the negative parameter for access distance *increased* as the longer access journey could be better explained (e.g. LATIS-TE25 compared to LATIS-FM1 in Tables 6.10 and 6.11, as discussed in Section 6.4.2.1).

case. For choice sets without the nearest major station appended it can be seen that three major stations are responsible for the vast majority of missing chosen stations: Inverness, Perth and Glasgow Queen Street. When the nearest major station is appended to the choice sets, the issue is largely corrected for Inverness and Glasgow Queen Street, although Perth remains problematic with only a small reduction. While these findings support the decision to incorporate the nearest major station in the choice sets for the separate WG and LATIS model calibration, there is clearly a trade-off between accounting for a greater proportion of observed choice and a reduction in the predictive performance of the model (at least when train-leg variables are not present).

To get a fuller appreciation of the potential reduction in predictive accuracy of the models when applied to new data it is necessary to also assess the impact of these missing chosen alternatives. This was achieved by repeating the analysis with all choice situations included and in those cases where the chosen station was missing from the choice set, adjusting the absolute difference for the affected station. For example, if an individual chose Inverness but it was not in their choice set, the calculated absolute difference for Inverness (between the number of times it was actually chosen and the sum of its probabilities across the model) was incremented by one. This adjustment is equivalent to assuming that Inverness was in the individual's choice set but was assigned a probability of zero by the model. The results of this analysis are shown on the right-hand side of Table 6.25, with and without the nearest major station appended to the choice sets. As expected, there is a substantial reduction in predictive accuracy, with the performance difference measure increasing from 23% to around 37% for models CMB-TE19 and CMB-TE24. The performance of model CBM-MN-TE12 is also reduced, but to a lesser extent, reflecting the lower number of missing chosen alternatives. However, there is now very little difference between the three models, confirming that a trade-off exists between accounting for a greater proportion of observed choice and the predictive accuracy of a model that does not contain train-leg variables that can adequately explain these observed choices.

It should be noted that the LATIS 2013 dataset contains a higher proportion of choice situations where the chosen alternative is not present in the choice sets (when major station not appended) than the calibration datasets: 9.8%, compared to 7.9% and 5% for the LATIS (2014 & 2015) and WG datasets respectively. This probably reflects the higher proportion of trip origins located in remote parts of the Highlands, where passengers have preferred to make a very long access journey to Inverness rather than board at a more local station and take a slow service (where it would be necessary to change at Inverness in any case for their onward journey). The extent of this behaviour is illustrated in Figure 6.23, where the red markers indicate trip origins where Inverness was chosen as the boarding station. Given that this is likely to be a particular characteristic of the 2013 LATIS survey, the degree that model predictive accuracy has been penalised when the missing chosen alternatives are taken into account may be overstated and not indicative of the expected performance of the models more generally.

		Chosen alternatives missing from choice sets	
		Major station not appended	Major station appended
Station	Major	Number	Number
Inverness	Y	50	4
Perth	Y	37	34
Glasgow QS	Y	20	2
Huntly	N	3	3
Aberdeen	Y	2	2
Haymarket	Y	2	NA
Aviemore	N	1	1
Gleneagles	N	1	1
Stirling	Y	1	1
Total		117	48
% of choice situations		9.83	4.03

TABLE 6.26: Summary of chosen stations missing from choice sets for LATIS 2013 survey validation, with and without the nearest major station appended.

6.8 Conclusions

This chapter has shown that it is possible to calibrate station choice models, using two independent and geographically distinct datasets, that are suitable for integration into both trip end and flow rail demand models. The best MNL models had a very good fit as measured by adjusted rho-squared and predicted station choice substantially better than a base model where the nearest station was assumed to have a probability of one. There was reasonably good coincidence in parameter estimates for many of the explanatory variables across the two datasets, indicating that the models have the potential to be transferable. This was tested by applying the best WG calibrated models to the LATIS dataset and vice versa, with somewhat mixed results, although in all cases the predictive performance of these models was superior to the base model.

The trip end variant RPL models showed that individual variation in parameter estimates was only significant for the mode-specific access time variables; and for the LATIS dataset there was no significant variation once the train leg variables had been introduced. There was only a marginal difference between the predictive performance of the MNL and RPL models, and the RPL models do not, therefore, appear to offer sufficient improvement over the MNL models to justify the extra complexity and time that would be involved in simulating station

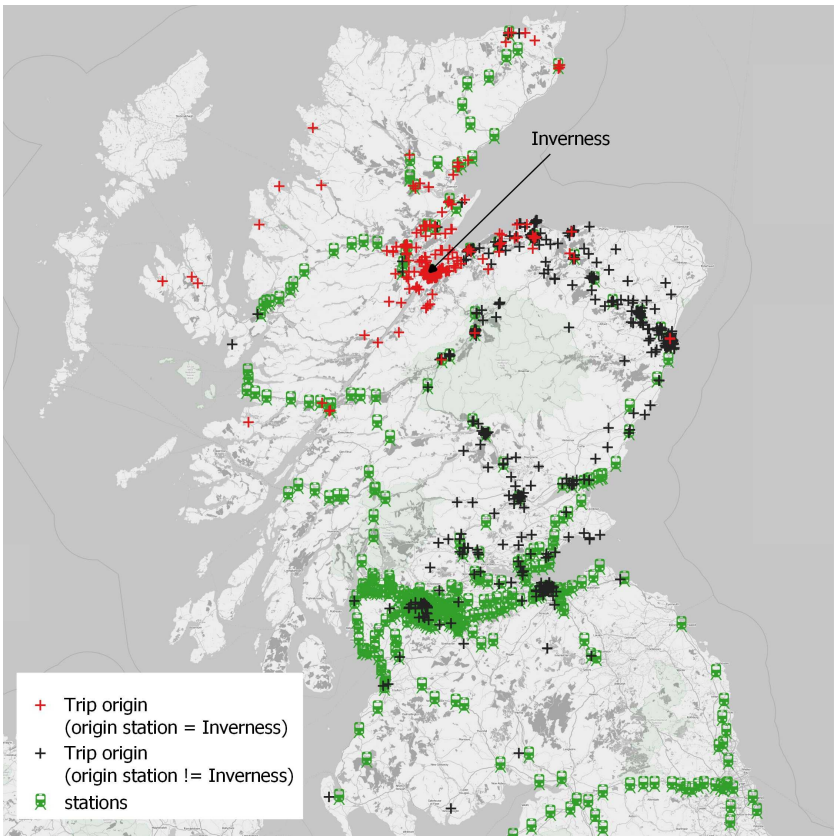


FIGURE 6.23: Trip origins for the LATIS 2013 survey. Red markers indicate those where Inverness was the chosen origin station.

probabilities for every unit postcode in GB (a requirement for calibrating a national-scale aggregate model).

The accessibility term, intended to account for spatial correlation between stations and potentially address the issue of proportional substitution, was found to have a significant and negative parameter in the trip end and flow variant MNL models, the flow variant LATIS model, and the combined dataset model. This indicates that there is a competition effect at play, and the closer a station is on average to other, and more 'attractive', stations, the less likely it is to be chosen. In the models where the parameter was positive, this may have been caused by the artificial spatial construct of choice sets with the 'nearest major' station appended. It remains to be seen, however, to what extent the estimated parameter will modify proportional substitution and what impact this might have on station abstraction forecasts.

The superior predictive performance of the station choice models compared to the base model suggests that they have the potential, through a more realistic representation of station catchments, to improve the aggregate models that are commonly used to assess proposals for new railway stations. The next chapter will focus on the development and application of a methodology to incorporate probabilistic station catchments, derived using

the combined dataset station choice model which was calibrated for this purpose, into a national-scale trip end model.

Chapter 7

Integrated trip end and station choice models

7.1 Introduction

This chapter is concerned with the development of enhanced trip end models to forecast demand for local railway stations in Great Britain. Section 7.2 sets out the background to this work, explaining how it builds on earlier research by incorporating probability-based station catchments; utilising much smaller-scale origin zones; and extending the calibration dataset to include Scotland. The preparation of the calibration dataset and derivation of the explanatory variables for the models are then described in Section 7.3. In Section 7.4, the proposed general model form that incorporates a probabilistic catchment is presented, and the key differences from the earlier research are discussed, including the estimation of decay functions from observed data. Section 7.5 explains the processes used to generate a choice set of railway stations for every unit postcode in mainland GB and to calculate the choice probabilities. The results of the model calibrations, which for comparative purposes include models which adopt either deterministic or probabilistic approaches to defining the station catchments, are presented in Section 7.6. The chapter then closes by summarising the work completed and drawing some conclusions (Section 7.7).

7.2 Background

Previous research carried out at the University of Southampton Transportation Research Group has successfully developed linear regression models to forecast the number of trips made to/from local railway stations in England and Wales (Blainey, 2010). In these trip end models, local stations were considered to be those assigned to Network Rail categories

E and F (otherwise known as 'small staffed' and 'unstaffed' stations). Station catchments were defined by allocating census output areas in England and Wales to their nearest station by road distance and applying a distance decay function to the population associated with each output area (from the 2001 Census), reflecting the expectation that the number of trips generated by the population of an output area will fall as the distance from the station increases. The best models were found to explain over 75% of variation in the observed data, and to better predict actual demand on the Ebbw Vale branch line (which opened in 2008) than the methods used in the feasibility study carried out prior to scheme approval. As part of consultancy work carried out for the Welsh Government, these models were subsequently re-calibrated using more recent data, including output area population from the 2011 census and station entries and exits (the basis of the dependent variable) from 2011/12 (Blainey, 2017).

These more recent trip end models have been taken as the starting point for developing new trip end models that incorporate probability-based catchments derived using the station choice models described in Section 6.7. These new models extend the earlier work in several key respects. Firstly, they are calibrated for stations in the whole of mainland GB, and not restricted to England and Wales. Secondly, unit postcodes are used to define catchment zones rather than census output areas; providing a much higher spatial resolution to the population data (there are some 1.5 million unit postcodes covering GB, compared to less than 0.25 million output areas). Thirdly, rather than assigning the population of each zone to its nearest station, the population is allocated to each station in a zone's choice set based on the probability that each station will be chosen, thus defining a probabilistic catchment.

7.3 Calibration dataset

In line with the earlier work carried out by Blainey (2017), the calibration dataset was defined as those railway stations assigned to Network Rail categories E and F. The categorisation of stations in England and Wales was last reviewed in 2009, with the revised categories published in a report commissioned by the Department for Transport (Green & Hall, 2009). This report was used as the definitive source for stations in England and Wales. Unfortunately, there does not appear to be an equivalent published list for stations in Scotland. Instead, a spreadsheet held within the Transportation Research Group containing this information was used[1]. Any station that opened after these lists were compiled was manually allocated to a category based on the category descriptions contained in Green and Hall (2009). The affected stations and the categories assigned are shown in Table D.2 in Appendix D.

Only stations that opened prior to 1 April 2011 were selected for inclusion in the calibration dataset. This date was chosen to ensure that all stations had been open for a full twelve

[1]The ultimate source of this information is not clear, but it certainly pre-dates the 2009 review of English and Welsh stations.

months when the annual station entries and exits data (used for the dependent variable) was compiled by ORR for the financial year 2011/12 (which ran from 1 April 2011 to 31 March 2012).

Some Category E and F stations were removed from the calibration dataset for several reasons. Those with no weekday service, restricted public access or located on the Isle of Wight were removed (i.e. any Category E or F station listed in Table D.5 in Appendix D). For ticketing purposes some stations (usually within the same town or city but on different lines) are grouped under a single common location, allowing passengers to travel to or from any station in a group (from or to any stations outside the group) using the same ticket. As a consequence, there is no accurate information available from the ticketing system on the number of trips made to or from these stations, and although the trips are apportioned to individual group stations in the data released by the ORR, this is likely to be unreliable. The groups were identified from the ATOC fares feed (further information is provided in Section D.1.3 in Appendix D), and any group stations were removed from the calibration dataset[2] (the station groups and member stations are summarised in Tables D.3 and D.4 in Appendix D). Following these removals, the final calibration dataset consisted of 1,792 stations.

7.3.1 Dependent variable

The basis of the dependent variable used in the trip end models was the total number of station entries and exits in the financial year 2011/12 as reported by the ORR (Office of Rail and Road, 2013).

7.3.2 Explanatory variables

The explanatory variables selected for inclusion in the models were based on those used to calibrate the previous trip end models (Blainey, 2017).

7.3.2.1 Workplace population

The number of usual residents aged 16 to 74 in employment the week before the 2011 census was obtained for each workplace zone in England and Wales from the NOMIS service (Nomis, 2014); and for each census output area[3] from Scotland's Census Data Warehouse (Scotland's Census, 2016). Each dataset was then merged in R with its corresponding population weighted centroids dataset obtained from the UK Data Service (UK Data Service, 2011), and then a combined GB dataset was exported into CSV format for use in subsequent ArcGIS analysis.

[2]Bicester Village station was not removed from the dataset as the Bicester North and Village group was not created until 28 July 2015.

[3]For Scotland, workplace population was not available using the new workplace zone geography.

Within ArcGIS, polygons were generated to represent the area accessible by road within one, two, three and four minutes drive-time of each category E and F station, using a 'New Service Area' analysis and the Open Roads network described in Section 7.5.2. Using a series of spatial joins, the workplace population within one, two, three and four minutes of each station was calculated by summing the population associated with any OA or workplace weighted centroid contained within each of the drive-time polygons. This information was then exported to DBF files, and subsequently imported into R during preparation of the trip end model data frame.

This approach has some limitations, as there will be instances where although part of a workplace zone falls within a travel time polygon, the zone centroid itself does not, and therefore no jobs will be included for that zone. A possible solution would be to distribute the workplace population within the zone, for example by creating a grid within each zone polygon and proportioning the population to each cell, potentially taking into account the placement of buildings, but unfortunately there was insufficient time to explore this further.

7.3.2.2 Train frequency

The train frequency at each station was obtained from train schedule information using a similar procedure to that used to derive this variable for the station choice models (see Section 5.4.2), with the data obtained in GTFS format, loaded into a series of PostgreSQL database tables, and then a suitable query run to obtain the train frequency for each station. The earliest suitable version of the schedule in GTFS format, dated 23 November 2013, was downloaded from the maintained archive[4] (see PotsgreSQL code segment B.2 in Appendix B).

7.3.2.3 Electric trains

The power type of trains is available in the schedule feed provided by Network Rail. Following a request to the 'openraildata-talk' Google Group, one of the group members provided a URL to retrieve all stations in the current timetable served by electric and electric multiple unit trains (the only electric power types recorded in the timetable at that time) (live-departures.info, 2017). Although this data source formed the basis of the variable, it needed to reflect the situation in 2011/12 (the base year for the model calibration), rather than 2017 when it was retrieved. Therefore, in addition to creating a boolean variable in the 'stations' database table to indicate those stations served by electric trains, an additional field was created to record, for all schemes completed since 2011, the date that electric services began. This information was manually collated from a variety of on-line governmental, news and reference sources.

[4]The archive for the weekly GTFS feed prepared by `http://www.gbrail.info/` is located at `http://transitfeeds.com/p/association-of-train-operating-companies/284`

This enabled a database query to be run to select only those stations served by electric trains as of 31 March 2011.

7.3.2.4 Travelcard boundary

Travelcard boundary stations were identified for schemes running in eight cities and regions of GB: Strathclyde (Roundabout Ticket), London (Zones 1–6 Travelcard), West Midlands (Centro supported area), Merseyside (Merseyrail Railpass area), Manchester (Greater Manchester ticketing boundary), West Yorkshire (METRO Zones 1–5), Tyne & Wear (Travelcard Zones 1–5); and South Yorkshire (PTE TravelMaster area). Where possible the boundary stations were identified based on the schemes that were running in 2011, and particular use was made of the collection of past rail schematic maps and diagrams provided by Project Mapping[5]. A total of 62 Category E and F stations were identified, and these are listed in Table D.1 in Appendix D.

7.3.2.5 Nearest Category A–D station

All category A, B, C, C1, C2, and D stations opened prior to 1 April 2011 were selected from the database along with their coordinates. These were then imported into ArcGIS, and an OD cost matrix analysis was carried out to find the nearest category A–D station by distance to each of the Category E and F stations in the calibration dataset, using the Open Roads network.

7.3.2.6 Terminus stations

This variable indicates whether or not a station forms the limit of passenger services on a particular line. To save unnecessary manual work, the data compiled during prior research carried out by Blainey (2017) was merged with the trip end model dataset. All stations where the terminus status was unknown, which included all stations in Scotland, were plotted in QGIS over a transport network base map to aid rapid identification of terminus stations.

7.3.2.7 Population

The resident population at the unit postcode level was obtained in CSV format from the NOMIS web service for England and Wales (Nomis, 2013) and from 'Scotland's Census' website for Scotland (Scotland's Census, 2013). For further information on the preparation of the postcode data see Section 7.5.1.

[5]See: `http://www.projectmapping.co.uk/rail_maps_diagrams.html`

7.4 Model form

7.4.1 Previous models

The starting point for the trip end model calibration was a model developed during previous work (Blainey & Preston, 2013b) and subsequently calibrated using more recent sources of the dependent and explanatory variables (Blainey, 2017). The model form when applied in forecasting mode is as follows:

$$\ln \hat{V}_i = \alpha + \beta \left(\ln \sum_{z}^{Z} P_z w_z \right) + \gamma \ln F_i + \delta \ln T_i + \epsilon \ln J_{it} + \zeta \ln Ps_i + \eta Te_i + \theta El_i + \iota B_i, \quad (7.1)$$

where $\hat{V}_i$ is the estimated annual passenger entries and exits for station i; P_z is the resident population of zone z; Z is all zones where the closest station by car travel time is station i; w_z is a distance decay function; F_i is weekday train frequency at station i; T_i is distance in km from station i to the nearest Category A–D station; J_{it} is the number of jobs within t minutes drive of station i, Ps_i is the number of parking spaces at station i, and Te_i, El_i and B_i are dummy variables that take the value of 1 if station i is a terminus station, served by electric trains or a travelcard boundary station respectively, and zero otherwise; and $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \eta, \theta$ and ι are the estimated parameters. The distance decay function w_z was specified as $(t+1)^{-3.25}$, where t is the road travel time from zone z to its closest station. The version of this model that gave the best model fit (as measured by adjusted R^2) specified the number of jobs within two minutes drive time of each station, and the reported results are summarised in Table 7.1.

Variable	Parameter	t-statistic
Intercept	3.992	24.660
Population	0.228	12.370
Employment (2 mins)	0.068	7.982
Train frequency	1.294	42.685
Distance to Cat A-D station	0.103	3.637
Car park spaces	0.157	14.018
Terminus dummy	0.767	7.701
Electrification dummy	0.238	4.914
Travelcard boundary dummy	0.490	4.166
Adjusted R^2	0.822	

TABLE 7.1: Results of trip-end model developed by Blainey (2017).

7.4.2 New models

While the new models follow a similar approach to those described above, they differ in several respects:

- In addition to allocating the population of each zone to its nearest station (for comparative purposes), the population of each zone is allocated to 10 (or more) alternative stations based on the probability of those stations being chosen.
- The most suitable distributions to represent both time and distance decay effects are identified, and function parameters estimated based on observed station access trips in the LATIS and WG datasets.
- The calibration dataset is larger (279 additional stations) as stations in Scotland are included.
- The zones used to define catchments are unit postcodes rather than census output areas.
- The road network from which distance or time related explanatory variables are obtained is based on the more detailed OS Open Roads dataset, rather than Meridian 2.

The first two items listed above will now be examined in more detail.

7.4.2.1 Trip decay functions

Appropriate trip decay functions, both time and distance-based, were obtained by analysing the access trips in the revealed preference survey data (LATIS and WG). Histograms of access time and access distance (as measured assuming car as access mode) were produced for the observations where the chosen station was designated Category E or F, as shown in Figures 7.1 and 7.2. One minute or 250 m bins were defined, and the number of bins was limited so that while nearly all observations were accounted for, a very large number of empty bins in the long right-hand tail of the distribution was avoided. The time and distance-based bins accounted for 99% and 98% of the 5,574 observations respectively. The histograms indicate that the decay does not begin until after the two-minute or 750 m bins, suggesting that a two-stage decay function would be appropriate, with no weighting applied to the population of any zone (i.e. postcode centroid) within either of these thresholds of a station.

The 'huff.decay' function of the MCI R package (Wieland, 2017) was used to estimate a time- and distance-based decay function using different function types (linear, power, exponential and logistic). As the observed decay does not begin until after the two-minute or 750 m bins, observations within those thresholds of their chosen station were removed prior to the

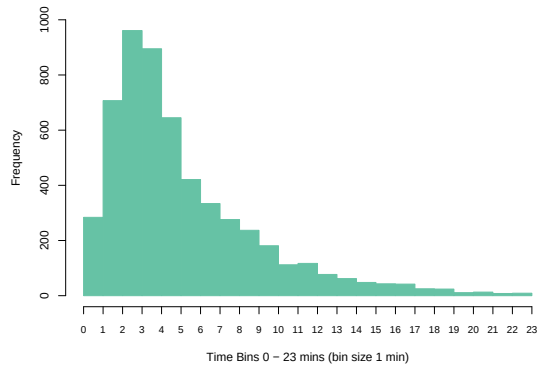


FIGURE 7.1: Histogram of access time for Category E and F stations (WG and LATIS data.)

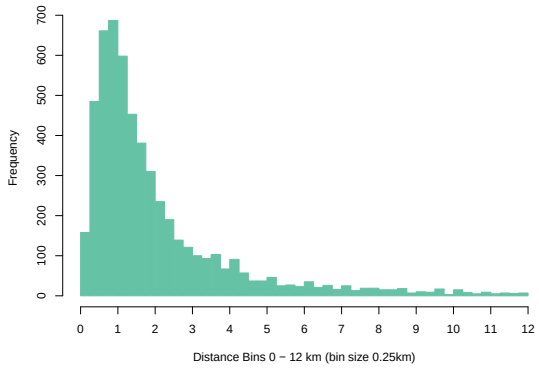


FIGURE 7.2: Histogram of access distance for Category E and F stations (WG and LATIS data.)

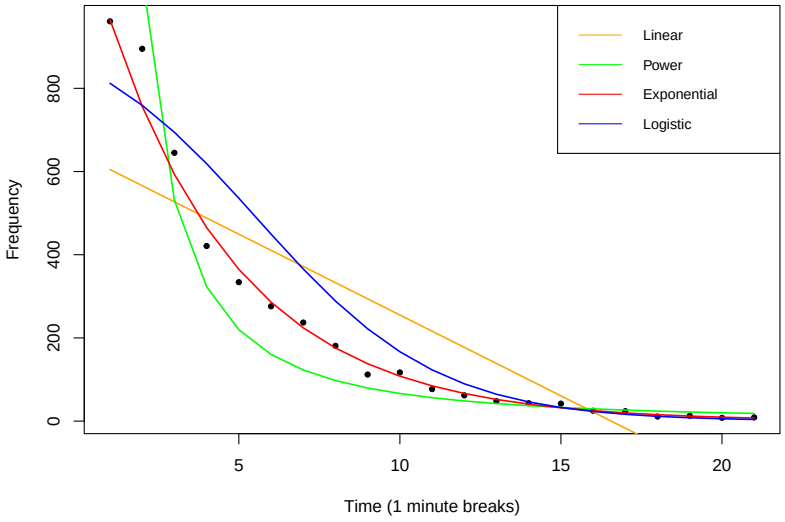


FIGURE 7.3: Output from the huff.decay() function using time bins.

relevant decay function being estimated. The results are shown graphically in Figures 7.3 and 7.4, and summarised in Tables 7.2 and 7.3.

The results show that an exponential function (slope −0.2432) gives the best fit to the access time data, with an adjusted R^2 of 0.99; while a power function (slope −1.5212) gives the best fit to the access distance data, with an adjusted R^2 of 0.91, slightly better than the exponential function with adjusted R^2 of 0.90. Figures 7.5 and 7.6 show a simulated decay for an initial population of 10,000 using the estimated power and exponential distance-based decay functions respectively. These support choosing the power function as the preferred model, as it appears to better represent the observed distribution (as shown in figure 7.2), with a deeper initial decay profile.

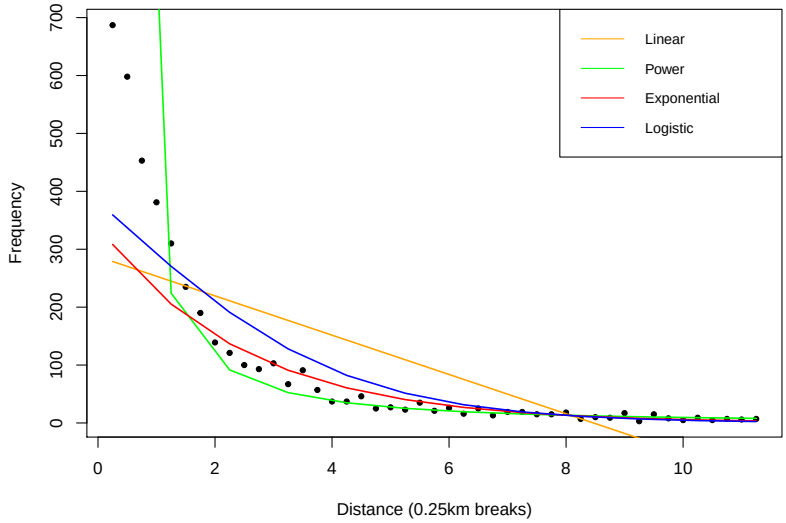


FIGURE 7.4: Output from the huff.decay() function using distance bins.

Model type	Intercept	p Intercept	Slope	p Slope	R-Squared	Adj. R-squared
Linear	643.6810	0.0000	-38.8584	0.0000	0.7021	0.6864
Power	3540.7886	0.0000	-1.7272	0.0000	0.8674	0.8604
Exponential	1230.4066	0.0000	-0.2432	0.0000	0.9907	0.9902
Logistic	-1.9904	0.0001	0.3563	0.0000	0.8613	0.8540

TABLE 7.2: Time decay function estimate.

Model type	Intercept	p Intercept	Slope	p Slope	R-Squared	Adj. R-squared
Linear	287.3283	0.0000	-33.9315	0.0000	0.5073	0.4959
Power	314.4850	0.0000	-1.5212	0.0000	0.9126	0.9106
Exponential	341.2106	0.0000	-0.4066	0.0000	0.9036	0.9014
Logistic	-0.2016	0.4931	0.5197	0.0000	0.7631	0.7575

TABLE 7.3: Distance decay function estimate.

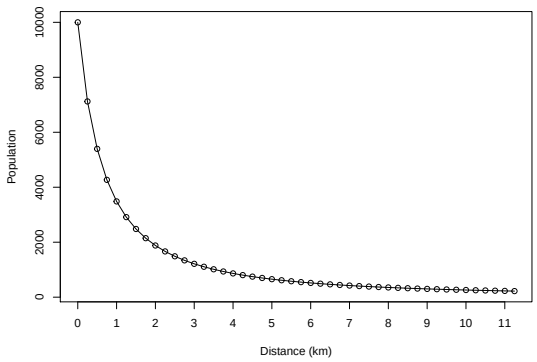


FIGURE 7.5: Simulated decay for population of 10,000 using power function (slope −1.5212).

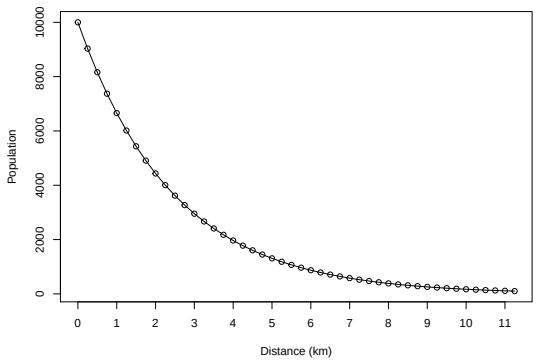


FIGURE 7.6: Simulated decay for population of 10,000 using exponential function (slope −0.4066).

7.4.2.2 Probabilistic catchment definition

To incorporate probabilistic station catchments into a trip end model, the model shown in 7.1 can be amended to the following form:

$$\ln \hat{V}_i = \alpha + \beta \left(\ln \sum_{z}^{Z} Pr_{zi} P_z w_{zi} \right) + \gamma \ln F_i + \delta \ln J_{it} + \epsilon \ln Ps_i + \zeta Te_i + \eta El_i + \theta B_i, \qquad (7.2)$$

where Pr_{zi} is the probability of someone located in zone z choosing station i; Z now consists of all zones which have station i within their choice set; and T_i has been removed from this model. T_i was incorporated to try to capture potential competition effects of nearby larger stations, something that should now be more adequately captured by the station choice component. An intuitive interpretation of the bracketed part of the equation is the trip generation potential of the population expected to use a station. This is the proposed general form of the model, with the nature of the zone being defined by the researcher. In the case of the models reported here, the zone is defined as the unit postcode, and the two-stage decay function w_{zi}, is either distance-based:

$$w_{zi} = \begin{cases} (d+1)^{-1.5212} & \text{if } d > 0.75 \\ 1 & \text{otherwise,} \end{cases} \qquad (7.3)$$

where d is the road distance in km from zone z to station i; or time-based:

$$w_{zi} = \begin{cases} e^{(-.2432 \times t)} & \text{if } t > 2 \\ 1 & \text{otherwise,} \end{cases} \qquad (7.4)$$

where t is road travel time in minutes from zone z to station i.

The next section will describe the process of generating station choice probabilities for every postcode in mainland GB, a level of detail needed to calibrate a national model. It should be noted that while the trip end model calibration dataset only contains Category E and F stations (the 'local' stations that the model will be used to forecast), all stations, of any category, are eligible to be included in the choice set of each postcode.

7.5 Generating station choice probabilities for Great Britain

In order to generate the station choice probabilities, it was necessary to first define a station choice set for every unit postcode in mainland GB. Then, for each choice set, the probability of each station being chosen could be calculated. The unit postcode represents the spatial level at which resident population will be weighted, both by the distance- or time-based decay function and the calculated choice probabilities, before being allocated to each station

in the model. The next three sections describe the preparation of the postcode data; the choice set creation process; and then the generation of probability tables.

7.5.1 Postcode data preparation

Only those postcodes that had resident population associated with them at the 2011 census were of interest. These postcodes, along with the population data, were obtained in CSV format from the NOMIS web service for England and Wales (Nomis, 2013) and from 'Scotland's Census' website for Scotland (Scotland's Census, 2013). The CSV files were imported into R, merged, and then written to a database table.

In the Scottish data some postcodes appeared twice with either an 'A' or 'B' appended to the the postcode, for example: 'AB12 3LPA' and 'AB12 3LPB'. Different population totals were associated with the two variants, suggesting that this might be connected with splitting postcode populations between census output areas, although no advisory information was provided with the dataset. These duplicated postcodes were identified using a regular expression matched against the last three characters of the postcode. In a valid unit postcode these should always be numeric, alpha, alpha. In any instances where this was not the case, one character from the right was removed from the postcode. This corrected 416 records, but left duplicated postcodes in the table with different population counts. To resolve this a new table was created using a select query that grouped records by postcode and summed the population field.

As several explanatory variables used in both the station choice and trip end models relate to the road network (for example station access distance), it was necessary to remove any postcodes that were isolated from the mainland road network. This included postcodes located on any island not connected by road to the mainland, and a few very remote postcodes in Scotland that are not connected to the public road network (for example, those only accessible via forest track[6]). Where possible postcode sectors unique to an island were identified using an interactive postcode district web map[7], which enabled all postcodes within those sectors to be readily removed. In other cases they were identified on an individual basis by visualising the postcode centroids in QGIS.

7.5.2 Deriving the choice sets

Initially it was intended to identify the ten nearest stations to each GB postcode using the same method adopted during development of the station choice models, as described in Section 6.3. However, while identifying the nearest 30 stations for each postcode by Euclidean distance would not be problematic, obtaining the actual distance to each of these stations in

[6]These were identified during the process of locating a postcode centroid to its nearest road segment.
[7]see `https://www.xyzmaps.com/maps/free-maps`.

Code List: RoadFunctionValue	
Code	**Description**
Motorway	A multi-carriageway public road connecting important cities.
A Road	A major road intended to provide large-scale transport links within or between areas.
B Road	A road intended to connect different areas, and to feed traffic between A roads and smaller roads on the network.
Minor Road	A public road that provides interconnectivity to higher classified roads or leads to a point of interest.
Local Road	A public road that provides access to land and/or houses, usually named with addresses. Generally, not intended for through traffic.
Local Access Road	A road intended for the start or end of a journey, not intended for through traffic but will be openly accessible.
Restricted Local Access Road	A road intended for the start or end of a journey, not intended for through traffic and will have a restriction on who can use it.
Secondary Access Road	A road that provides alternate/secondary access to property or land not intended for through traffic.

FIGURE 7.7: Road function codes within the Open Roads dataset. Note: Reprinted from 'OS Open Roads: User guide and technical specification', by Ordnance Survey (2017), p. 23. Image reproduced with permission of the rights holder, Crown copyright.

order to correctly rank them by road distance would have required over 40 million queries to the OTP API. To make this more manageable, it was planned to identify only the nearest 20 by Euclidean distance, and then carry out the API queries using multiple R clients running in a cloud client-server environment. However, initial tests indicated that certain geographical features, for example the River Thames and the Thames estuary, resulted in choice sets that did not accurately reflect the nearest stations on the road network. An alternative solution was therefore required that could directly identify the nearest x stations via the road network. The preferred option was to use the pgRouting extension for the PostGIS/PostgreSQL spatial database, in which the data was already held. However, as a suitable function to perform this task using pgRouting was not available, an OD Cost Matrix analysis using the ArcGIS Network Analyst extension was identified as the only viable option.

A network dataset was created in ArcGIS using the OS Open Roads dataset which was downloaded in the ESRI Shapefile format and imported into a file geodatabase (Ordnance Survey, 2016). The drive time of each network segment was assigned based on the identified road function and, where applicable, whether the segment was single or dual carriageway (See Figure 7.7 for the road functions defined in the Open Roads dataset, and Table 7.4 for the speed specified for each).

An OD cost matrix analysis requires origins and destinations to be loaded and located onto the nearest part of the road network. For this analysis the origins were the unit postcodes obtained from the 2011 census (as described above), and the destinations were the stations in operation prior to 2012[8]. Certain stations were excluded from the analysis, and these are summarised, along with the reason for their exclusion, in Table D.5 in Appendix D.

[8]The last station to open in 2011 was Buckshaw Parkway on 3 October 2011, midway during the 2011/12 financial year used by ORR to report annual station entries and exists — the dependent variable used in the trip end models.

	Assumed speed (mph)	
Road function	Single carriageway	Dual carriageway
Motorway	65	
A Road	45	50
B Road	40	45
Minor Road	30	
Local Road	25	
Local Access Road	20	
Restricted Local Access Road	20	
Secondary Access Road	15	

TABLE 7.4: Speeds applied to segments in the OS OpenRoads network dataset.

Predominantly this was because the station is not accessible to typical passengers, either because it is located on private property or a considerable distance from the public road network; or because the station does not offer any weekday service[9]. In addition, all stations on the Isle of Wight were excluded, as both the rail and road network are isolated from the mainland. Once all the origins and destinations had been loaded and located onto the road network a series of OD cost matrix analyses were run to find the nearest 15 stations by time, with distance also recorded[10]. In a small number of cases the analysis was unable to find any, or a sufficient number, of stations. This was due to some streets being orphaned from the rest of the road network. These issues were resolved by manually editing the network in ArcGIS to connect the orphaned sections to the rest of the network with reference to online mapping services, and then re-running the cost matrix analysis for the affected origins. The results of each analysis were exported from ArcGIS in DBF format and subsequently imported into R where they were merged into a single dataframe (of some 22 million records), processed, and then written to a PostgreSQL table. This table contained the nearest 15 stations to every postcode in GB, from which the choice set of the nearest 10 was selected. The same procedure was followed to identify the nearest major station to each postcode by distance, with the cost matrix destinations in this case consisting of all Category A, B and C1 stations.

7.5.3 Creating probability tables

Two probability tables were generated for the station choice component of the trip end models, one containing the nearest 10 stations to each postcode by distance, and the second additionally including the nearest major station to each postcode (if not already present). The tables were created by selecting the choice set records from the nearest 15 station tables

[9]Many of these stations are served by so-called 'parliamentary trains', a bare minimum service to avoid invoking the costly formal process of closing a station ("Why Do Some Stations", 2015).

[10]ArcGIS was unable to complete an OD cost matrix analysis with all the origins (1.45 million). The analysis was therefore run in seven batches of approximately 200,000 origins.

as required, and pulling in additional explanatory variables from the stations database table using joins.

7.5.3.1 Calculating the accessibility term

The calculation of the accessibility term was discussed in Section 6.4.1.4, and the modification of the weighting variable to enable it to be used in predictive models was outlined in Section 6.7.2. However, due to the size of the probability tables (in excess of 14 million records), the scripts that were previously used to identify unique station pairs (in order to look-up the distance between them) and perform the relatively complex calculation of the accessibility term would have taken several days to complete. To resolve this a block of procedural language (PL/pgSQL) code was written to generate the information directly using the PostgreSQL database, thus eliminating the processing overhead of a scripting language. The code used to identify the set of unique station pairs is shown in PostgreSQL code segment B.3 in Appendix B. A total of 47,520 unique station pairs were identified, and the distance between each of these pairs was obtained by querying the OTP API (specifying walk mode). The code used to calculate the accessibility term for every record in the probability table is shown in PostgreSQL code segment B.4 in Appendix B.

7.5.3.2 Generating probabilities

For each record in both probability tables, a field was populated with the exponentiated measured utility by applying the appropriate combined station choice model depending on choice set definition (CMB-TE19 and CMB-TE24 for nearest 10 stations and CMB-MN-TE12 for nearest 10 plus nearest major). Using a window function (with records partitioned by postcode), another column was then populated with the sum of the measured utility for all the alternatives in each choice set. Finally, a column was populated with the probability that each station was chosen.

7.5.3.3 A railway station choice predictor application

To enable the station probabilities for particular postcodes to be easily interpreted, and allow a sense-check of the performance of the predictive models across GB to be carried out, an application was developed using Shiny, an R package for creating interactive web applications (Chang et al., 2017). The user enters a postcode and selects the required station choice model and the application then queries the appropriate probability table in a PostgreSQL database and displays each station within that postcode's choice set and their respective choice probabilities. The application, which is hosted on a cloud server, can also generate a choropleth map showing the probabilistic catchment for any station. Screenshots of the web interface are shown in Figures 7.8 (probability table) and 7.9 (catchment map).

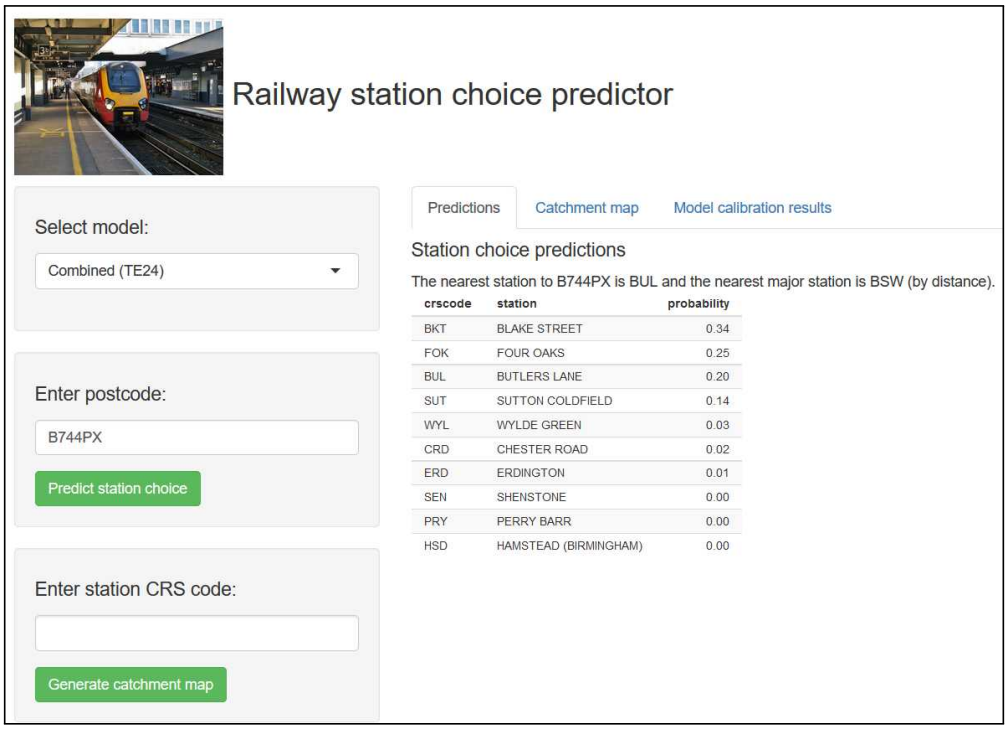


FIGURE 7.8: Interface of the station choice predictor web application showing the probability table for a postcode.

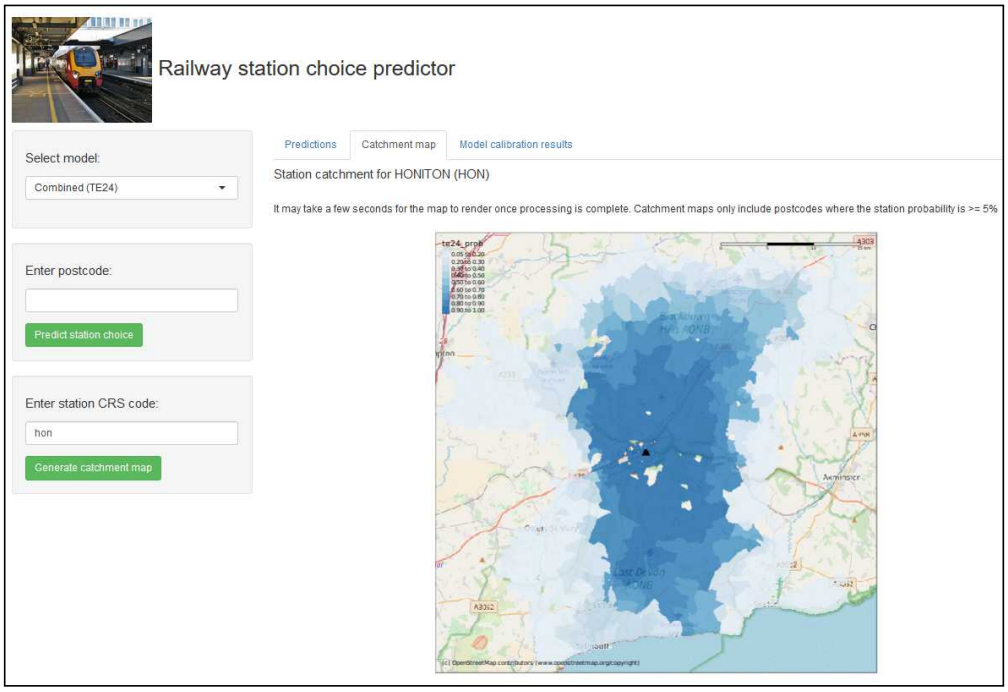


FIGURE 7.9: Interface of the station choice predictor web application showing the probabilistic catchment for a station.

7.5.3.4 Retrieving population totals

The population totals for each station in the calibration dataset were retrieved using SQL queries that pulled data from the relevant probability table and the postcode table, and applied probability and decay-function weightings on-the-fly. Example SQL queries using both simple and probability-based catchment definitions are provided in Section B.5 in Appendix B.

7.6 Trip end model results

The trip end models were estimated in R using the 'lm' function, and to enable comparisons to be made, models were estimated using deterministic catchments (model form shown in Equation 7.1) and probabilistic catchments (model form shown in Equation 7.2). The results are summarised in Tables 7.5 and 7.6. Although the adjusted R^2 goodness of fit measure is reported in the results tables, the AIC is considered to be a more appropriate in-sample measure for comparing the predictive accuracy of models, as it seeks to estimate how well a model will predict new or future data rather than how well it explains the current data (Sober, 2002). The preferred model is considered to be the model with the lowest 'headline' AIC value, and the difference between the AIC value of each model and the best performing model (known as the delta AIC) can be calculated as follows:

$$\Delta i(AIC) = AIC_i - \min(AIC). \tag{7.5}$$

This raises an important question: how much confidence can the researcher have that a model with a lower AIC value really is better than a model with a higher AIC value? And how big does the difference need to be to confidently discard a model when a single predictive model is being sought? These questions can be answered by calculating the Akaike weight of each model, which is the ratio of the delta AIC to the sum of the delta AICs of all (K) models:

$$w_i(AIC) = \frac{\exp(-\Delta_i/2)}{\sum_{k=1}^{K} \exp(-\Delta_k/2)}, \tag{7.6}$$

so that $\sum w_i(AIC) = 1$. $w_i(AIC)$ is then interpreted as the probability that model i is the best of the models under consideration. Furthermore, by calculating the ratio of the $w_i(AIC)$ of two models, known as the evidence ratio:

$$\frac{w_{m2}(AIC)}{w_{m1}(AIC)}, \tag{7.7}$$

it is possible to infer the extent to which model 2 is better than model 1, and, by expressing the evidence ratio as a normalized probability, the probability that model 2 is the better of

the two models[11]:

$$\frac{w_{m2}(AIC)}{w_{m1}(AIC) + w_{m2}(AIC)}. \tag{7.8}$$

The AIC value, delta AIC and the AIC weight are reported for each model in the results tables.

Initial models were estimated to determine whether assigning each postcode to its nearest station by time (models 1 to 3) or distance (models 4 to 6) resulted in better performing models when using deterministic catchments; and to identify which travel time threshold for workplace population (one, two, or three minutes[12]) performed the best. The models where postcodes were assigned by distance performed better than those where postcodes were assigned by time (see Table 7.5). Model 4, using a one-minute threshold for workplace population, was the preferred model, with the AIC weight indicating a > 99% probability that this was the best of the six models.

Results from subsequent models (7–10) are summarised in Table 7.6, with model 4 included for comparison purposes. In model 7 the postcode population was weighted using the distance-based decay function described in Section 7.4.2.1. This function was found to perform consistently better than the time-based decay function for both deterministic and probabilistic catchments (results from models estimated using this function are not reported here for reasons of brevity). Probabilistic station catchments were incorporated into models 8 to 10, with the postcode populations weighted by station probabilities derived using different station choice models (see Table 6.20). Model 8 used station choice model CMB-TE19, model 9 used CMB-TE24 which contains the accessibility term, and model 10 used CMB-MN-TE12 which includes the nearest major station in the choice sets (but not the accessibility term).

All the models fit the data very well, with model 9 the best fitting model (adjusted $R^2 = 0.8506$). Model 9 had the lowest AIC, and the AIC weights indicated an 80% probability that this was the best of the five models. Model 10 had the next lowest AIC, with an 18% probability of being the preferred model. While introducing the distance decay function into model 7 reduced the AIC by 64 units compared with model 4, the largest reduction in AIC (78 units) was observed between model 7 and model 8, with the incorporation of probabilistic station catchments. Model 8 was then further improved by the addition of the accessibility term in model 9. The difference in AIC between the best and worse performing models (between model 4 and model 9) was 149.

A standardized residuals[13] plot for model 9 is shown in Figure 7.10. The accuracy of the prediction is shown on the y-axis, with the prediction becoming less accurate as the distance from the zero line increases. Points above the line indicate that the prediction was too low, and points below the line indicate that the prediction was too high. In general the residuals

[11] This discussion about the use of AIC in assessing the performance of predictive models, and the notation used, is based largely on Wagenmakers and Farrell (2004).

[12] A four-minute threshold was also tested but for brevity these models, which performed worse than those with the three-minute threshold, are not shown in the summary tables.

[13] Standardized residual = $(observed - expected) \div \sqrt{expected}$

	Model 1			Model 2			Model 3			Model 4			Model 5			Model 6		
	Population assigned to nearest station (by time)			Population assigned to nearest station (by time)			Population assigned to nearest station (by time)			Population assigned to nearest station (by distance)			Population assigned to nearest station (by distance)			Population assigned to nearest station (by distance)		
Variable	*B*	*t*	*Sig*	*B*	*t*	*Sig*	*B*	*t*	*Sig*	*B*	*t*	*Sig*	*B*	*t*	*Sig*	*B*	*t*	*Sig*
Intercept	2.74	15.66	***	2.55	14.65	***	2.32	13.24	***	2.58	14.37	***	2.36	13.24	***	2.14	11.90	***
ln(population)[1]	0.22	14.58	***	0.21	13.73	***	0.23	14.54	***	0.23	15.06	***	0.23	14.52	***	0.24	14.99	***
ln(daily train frequency)	1.43	50.78	***	1.42	49.99	***	1.43	49.41	***	1.43	50.90	***	1.42	50.09	***	1.43	49.46	***
ln(dist. to Cat A-D station)	0.14	5.80	***	0.16	6.67	***	0.17	6.83	***	0.15	6.20	***	0.18	7.21	***	0.19	7.34	***
ln(work pop. 1 min)[1]	0.09	12.77	***							0.09	13.48	***						
ln(work pop. 2 mins)[1]				0.10	11.97	***							0.11	13.02	***			
ln(work pop. 3 mins)[1]							0.09	9.75	***							0.10	10.59	***
ln(car park spaces)[1]	0.13	13.23	***	0.13	12.97	***	0.13	13.03	***	0.13	13.43	***	0.13	13.14	***	0.13	13.25	***
Electric services	0.20	4.61	***	0.19	4.42	***	0.19	4.20	***	0.20	4.61	***	0.19	4.48	***	0.18	4.19	***
Travelcard boundary	0.31	3.23	**	0.32	3.37	**	0.31	3.23	**	0.31	3.30	***	0.33	3.45	***	0.32	3.32	***
Terminus	0.89	10.17	***	0.90	10.32	***	0.94	10.63	***	0.90	10.31	***	0.91	10.44	***	0.95	10.77	***
Adjusted R^2		0.8366			0.8349			0.8307			0.8378			0.8368			0.8318	
AIC		3924.5170			3942.7820			3988.2390			3911.7690			3922.7630			3976.2580	
Delta AIC		12.7480			31.0130			76.4700			0.0000			10.9940			64.4890	
Akaike weight		0.0017			0.0000			0.0000			0.9942			0.0041			0.0000	

Notes [1]log(variable + 1) used due to presence of zero values

TABLE 7.5: Summary of trip-end model calibration. Models 1–8, no weighting applied to population.

	Model 4			Model 7			Model 8			Model 9			Model 10		
	Population assigned to nearest station (by distance)			Population assigned to nearest station (by distance); distance decay function			Population probability-weighted; distance decay function			Population probability-weighted; distance decay function; accessibility term			Population probability-weighted; distance decay function; nearest main station incl.		
Variable	*B*	*t*	*Sig*	*B*	*t*	*Sig*	*B*	*t*	*Sig*	*B*	*t*	*Sig*	*B*	*t*	*Sig*
Intercept	2.58	14.37	***	2.37	13.53	***	3.67	38.50	***	3.65	38.30	***	3.77	40.96	***
ln(population)[1]	0.23	15.06	***	0.34	17.30	***	0.37	20.14	***	0.37	20.38	***	0.37	20.29	***
ln(daily train frequency)	1.43	50.90	***	1.36	48.40	***	1.14	41.47	***	1.13	41.21	***	1.12	39.96	***
ln(dist. to Cat A-D station)	0.15	6.20	***	0.21	8.64	***									
ln(work pop. 1 min)[1]	0.09	13.48	***	0.06	7.66	***	0.05	7.75	***	0.05	7.70	***	0.05	8.06	***
ln(car park spaces)[1]	0.13	13.43	***	0.15	15.44	***	0.13	14.14	***	0.13	14.07	***	0.12	13.60	***
Electric services	0.20	4.61	***	0.22	5.09	***	0.24	5.93	***	0.24	5.97	***	0.25	6.12	***
Travelcard boundary	0.31	3.30	***	0.30	3.26	**	0.30	3.29	**	0.30	3.26	**	0.32	3.53	***
Terminus	0.90	10.31	***	0.86	10.03	***	0.78	9.37	***	0.78	9.34	***	0.75	8.99	***
McFadden's adjusted R^2		0.8378			0.8434			0.8500			0.8506			0.8504	
AIC		3911.7690			3848.2150			3770.2630			3762.5480			3765.4980	
Delta AIC		149.2210			85.6670			7.7150			0.0000			2.9500	
Akaike weight		0.0000			0.0000			0.0169			0.8001			0.1830	
Mean Squared Error (MSE)		0.514			0.496			0.475			0.473			0.474	

Notes [1]log(variable + 1) used due to presence of zero values

TABLE 7.6: Summary of trip-end model calibration. Model 4 (best model from previous table) and Models 9–13.

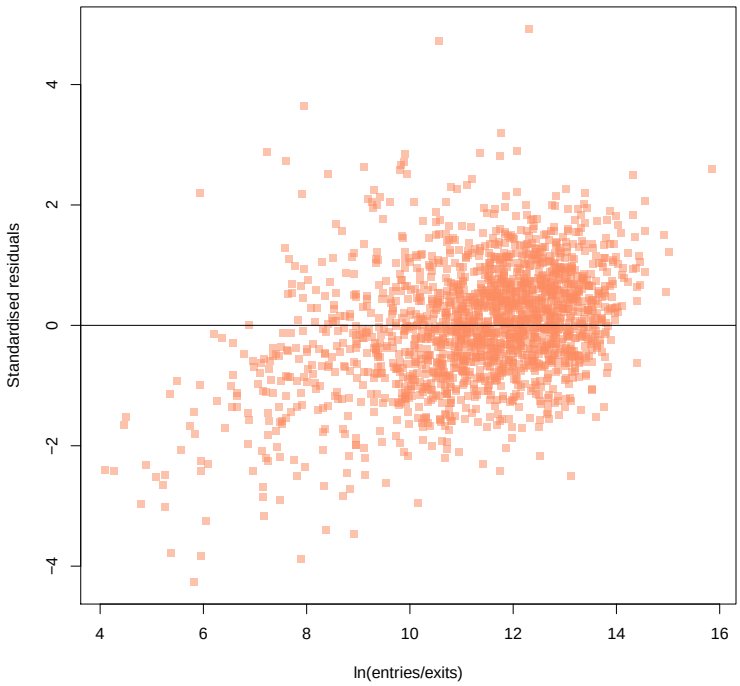


FIGURE 7.10: Standardised residuals plot for model 9.

are symmetrically distributed around zero, clustered toward the middle of the plot at lower values on the y-axis, and therefore consistent with random error. However, there does appear to be evidence of a systematic error at very low actual entries/exits (below eight on the logarithmic scale, or below around 3,000 annual entries/exits). At these lower values, the model systematically over-predicts, and the extent of over-prediction increases as the number of entries/exits becomes smaller. This indicates that the model is unable to account for unexpectedly low observed demand at some stations, given the predictor variables, and this could result in the model substantially over-forecasting demand for some new stations, if they were to share similar, but unknown, characteristics.

In Figure 7.11, the standardised residuals from model 9 have been plotted against the catchment population (weighted by probability and distance decay). This shows that under- and over-prediction becomes larger and more prevalent at low catchment populations. This effect can be seen in more detail in Figure 7.12, which only includes stations with catchment populations <= 5,000. The effect is particularly noticeable at catchment populations of around 100 and below, suggesting that station demand forecasts generated using this model should be treated with extra caution when the weighted catchment population is very low. Some stations with very low catchment population may have particularly strong attraction characteristics, such as a very large employer (for example, the station serving the Sellafield nuclear facility) or a large sports/entertainment arena. Demand at these stations is likely to be under-predicted by the model. Over-prediction when the catchment population is very low may be due to the 'fixed' elements of the model. For example, train frequency will generate

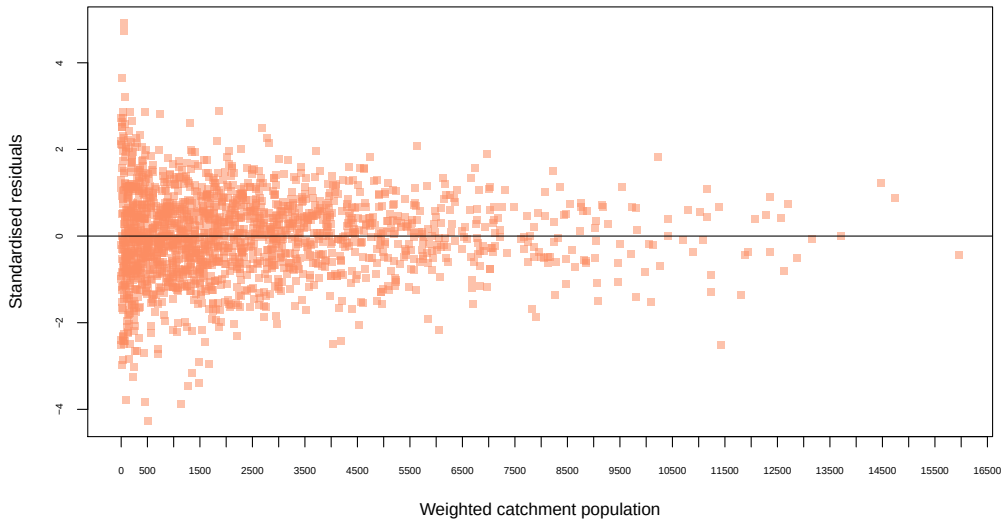


FIGURE 7.11: Standardised residuals against weighted population for model 9.

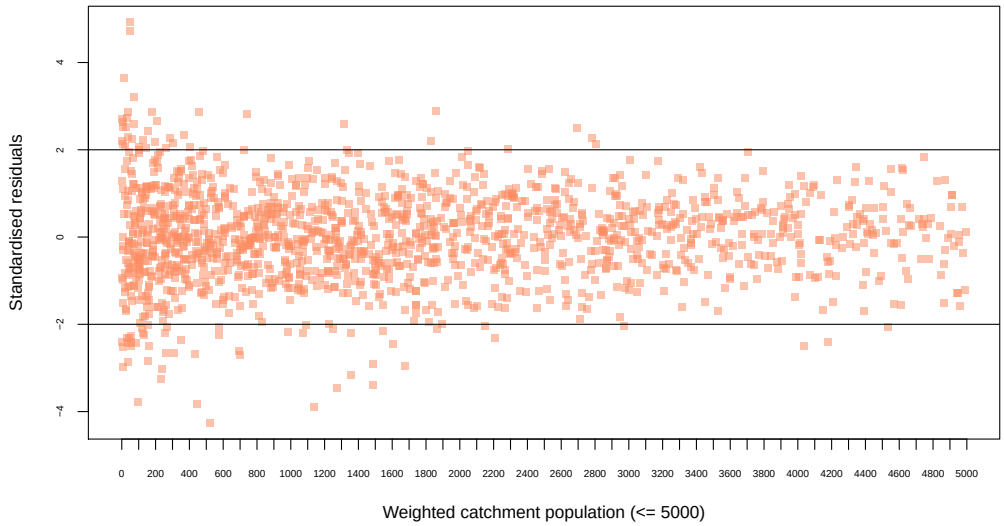


FIGURE 7.12: Standardised residuals against weighted population (< 5,000) for model 9.

trips even with zero population in the station catchment. This problem has to some extent been addressed by incorporating the probability-based catchments, as the weight attached to service frequency has been reduced relative to catchment population.

7.6.1 Examining geographic variation in model performance

In order to assess the performance of the trip end model on a geographic basis and identify any potential systematic bias at regional level, the standardised residual for each station was plotted on a map of GB, as shown in Figure 7.13. In this map the radius of each point is proportional to the size of the residual, although it should be noted that the points for stations with very small residuals are not visible at this scale. Overall, the map shows that under-prediction and over-prediction occurs in all regions of the country, and is present at a range of magnitudes. This suggests that the model performs similarly across the country, with

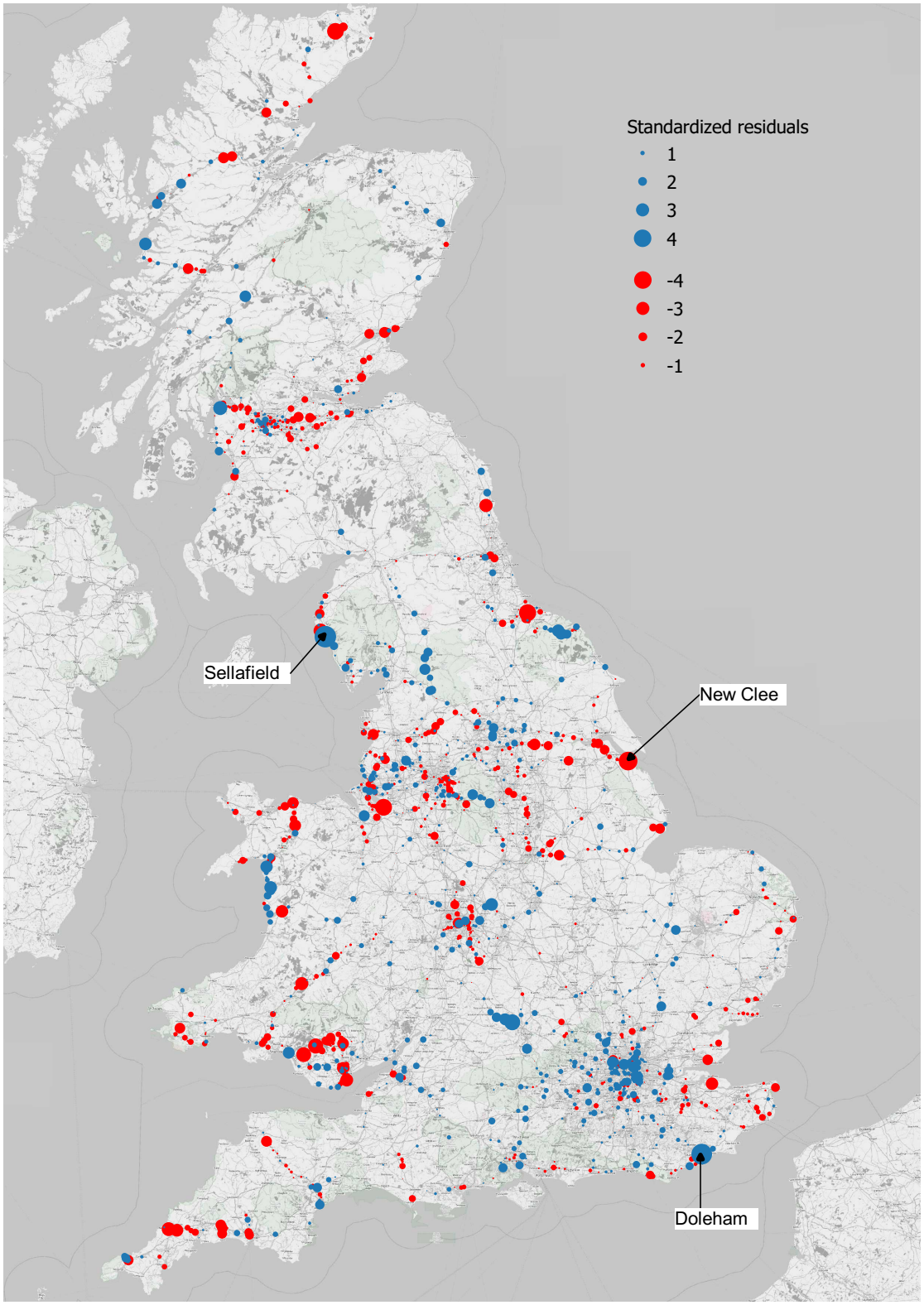


FIGURE 7.13: Standardised residuals (from model 9) for each station plotted on a map of GB. The radius of each point is proportional to the size of the residual with positive residuals (model under-prediction) shown in blue and negative residuals (model over-prediction) shown in red. The legend includes radii for example residual values.

no obvious regions where the model systematically under- or over- predicts station demand, and no regions where the standardised residuals appear systematically larger than in others. There is perhaps a tendency for under-prediction to dominate in the Greater London area. This would be expected given that there is no realistic alternative to public transport modes for travelling to/from central London and there is no variable that captures this additional generation effect within the trip end model.

The three stations with a standardised residual outside of the range ±4 are identified on the map. These are Sellafield, Doleham and New Clee. As previously mentioned, Sellafield is an example of a station with a very low weighted catchment population (51) but a high attraction factor due to the nearby nuclear facility. The centroid for the work population associated with this facility, some 12,000, fell outside the one-minute drive time threshold, exacerbating the degree of under-prediction. Doleham station also has a very low weighted catchment population (49), but is reported to have been very popular with weekend leisure travellers and walkers. A reduction in services appears to have caused a substantial fall in passengers at this station in recent years, with the number of entries/exits falling from 38,666 in 2011/12 (the calibration year) to 4,768 in 2016/17 — much closer to the model prediction for this station of 1,494. New Clee station is in a suburb of Grimsby and the model has substantially over-predicted demand for this station. It is a request-only stop with limited services and a small probability weighted catchment population of 520 (which reflects competition with nearby stations with better service provision). It is also likely that strong competition from a frequent bus service in this urban area has further suppressed demand at this station. The trip end model is not sensitive to competition from other modes, an issue discussed in the next chapter which addresses the application and appraisal of the model.

7.6.2 Comparison of parameter estimates

The parameter estimates for models 7 and 9 are compared in Figure 7.14, along with those from the model by Blainey (2017) which used census output areas as the zonal unit and was calibrated on English and Welsh stations only. Considering models 7 and 9, it is apparent that as the catchment definition is refined, the population parameter becomes larger and the daily frequency and terminus dummy parameters become smaller. The weighting attached to population is greatest in model 9, while the daily frequency and terminus dummy parameters are the smallest in this model. Wardman and Whelan (1999) note the importance of correctly specifying station catchments to avoid generation and attraction effects being falsely attributed to other variables, such as service levels. These results suggest that too much weight is being given to station service quality and characteristics in model 7, due to inadequacies in the catchment definition. It appears that model 9 can better account for differences in station usage that are explained by station catchments and their generation potential, and as a consequence this model should be more robust and transferable. It is also

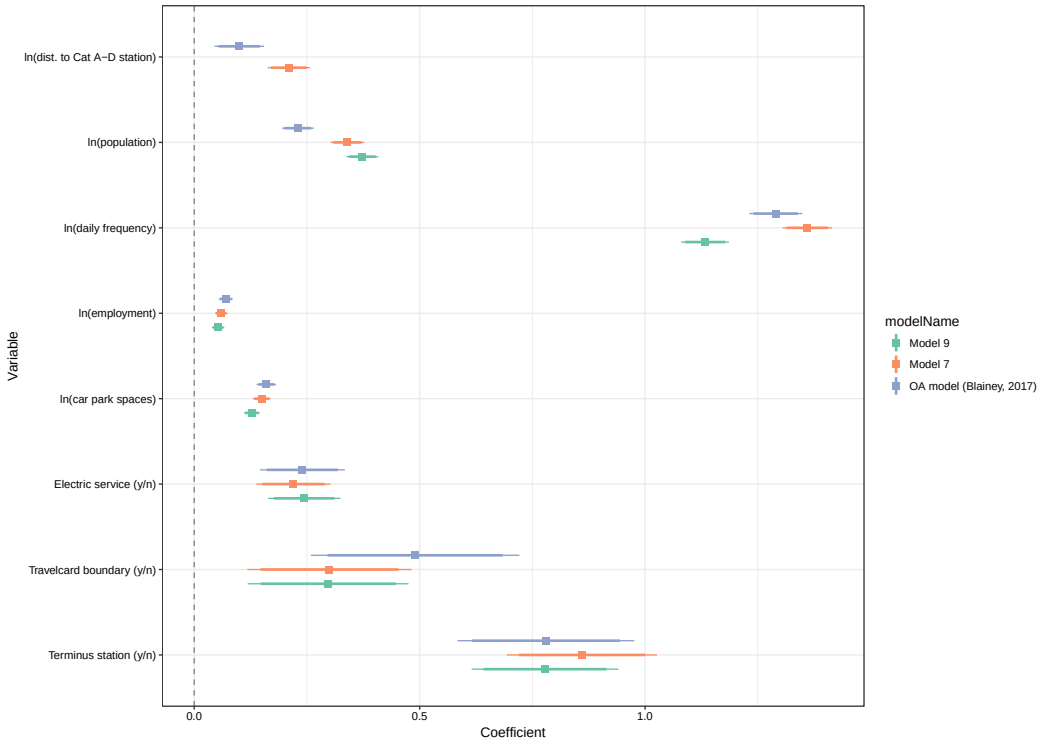


FIGURE 7.14: Comparison of coefficients (with 95% and 99% confidence intervals) estimated by the Blainey (2017) trip end model with OAs as zonal unit (deterministic), model 7 (deterministic), and model 9 (probabilistic).

interesting to note that in the Blainey (2017) model the population parameter is substantially smaller than in model 7, suggesting that the use of the higher spatial resolution zonal unit (postcode rather than census output area) has in itself improved the representation of the station catchment.

7.6.3 Assessing model predictive accuracy

To assess the predictive accuracy of model 9 a repeated *k*-fold cross-validation was carried using the `CVlm()` function from the R DAAG package (Maindonald & Braun, 2015). The predictive accuracy is expressed as the average mean squared error (MSE) of all the folds (see Section 6.7.4 for an explanation of the *k*-fold cross-validation technique). A 10-fold cross-validation was repeated ten times and the results for each fold and each repeat are shown in Table 7.7. The average estimate across all ten repeats was 0.478, representing a very small increase compared to the internal MSE of 0.473 for model 9 (see Table 7.6), suggesting that the model's predictive validity will hold when applied to new data. There is only a small variance in the cross-validation estimate across the repeats (the maximum difference is 0.002), indicating that the model has high stability. The results from the first repeat are plotted in Figure 7.15, where the points represent the dependent variable for each station (ln(entries/exists)) and the colours indicate the fold that each station was assigned

to. The lines join the cross-validation predicted values for each fold[14]. There is very close correspondence in the plotted lines for each fold, giving confidence that the model is stable and would be expected to perform consistently on new data.

	MSE of each fold										
Repeat	1	2	3	4	5	6	7	8	9	10	CV (mean)
1	0.568	0.374	0.434	0.572	0.506	0.473	0.414	0.369	0.586	0.490	0.478
2	0.442	0.631	0.434	0.435	0.606	0.490	0.380	0.485	0.476	0.401	0.478
3	0.460	0.547	0.549	0.439	0.536	0.451	0.452	0.426	0.468	0.453	0.478
4	0.516	0.496	0.453	0.484	0.612	0.451	0.435	0.439	0.403	0.481	0.477
5	0.378	0.467	0.437	0.545	0.588	0.465	0.441	0.432	0.543	0.488	0.478
6	0.402	0.540	0.475	0.508	0.520	0.362	0.413	0.437	0.528	0.596	0.478
7	0.549	0.528	0.480	0.525	0.475	0.502	0.462	0.349	0.468	0.453	0.479
8	0.536	0.519	0.457	0.456	0.550	0.385	0.439	0.466	0.429	0.550	0.479
9	0.448	0.532	0.538	0.410	0.426	0.476	0.473	0.489	0.374	0.623	0.479
10	0.483	0.411	0.503	0.422	0.657	0.457	0.507	0.504	0.426	0.413	0.478
										Average of all repeats	0.478

TABLE 7.7: Summary of the mean squared error (MSE) for 10-fold cross validation of trip end model 9, repeated 10 times.

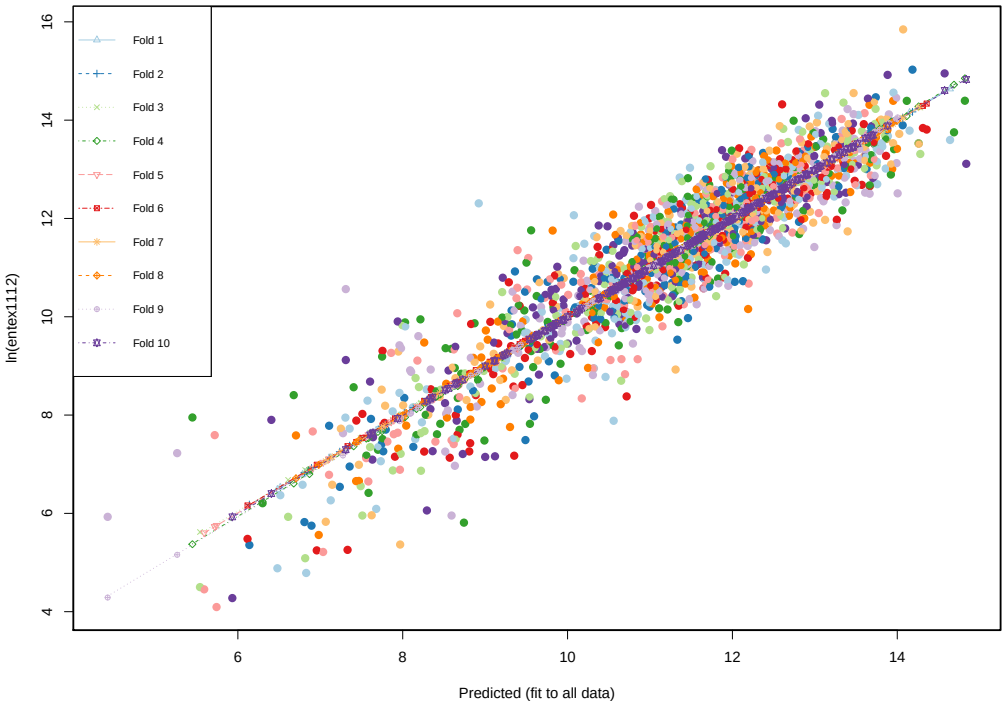


FIGURE 7.15: Plot of 10-fold cross validation completed for trip end model 9 (first repeat). The points represent the dependent variable for each station and the colour indicates the fold it was assigned to. The lines join the cross validation predicted values for each fold.

[14]As these values are not a linear function of corresponding overall predicted values the lines are approximate (Maindonald & Braun, 2015).

7.7 Conclusions

This chapter has described the development of national trip end models for local railway stations in Great Britain, which enhance models developed in earlier research in three key respects: by incorporating probabilistic station catchments; adopting zonal units of a higher spatial resolution; and extending the geographical scope to include stations in Scotland. A general model form has been proposed that allows trips at a station to be generated from any zone which has that station in its choice set, with the generation potential of the zone's population dependent upon the probability of the station being chosen and the distance of the zone from the station. Probability tables were generated which contained a choice set of ten[15] stations for every postcode in mainland GB, and the associated choice probabilities were calculated based on several estimated station choice models. Use was made of database queries and novel procedural code to enable efficient data processing and generation of model variables. A web application was developed to aid the interpretation of choice predictions for postcodes across GB.

An analysis of revealed preference survey data established that a power distance decay function (slope -1.5211) or an exponential time decay function (slope $-.2432$) applied to postcodes located more than 750 m or two minutes respectively from the chosen station best fit the observed trip data. For comparative purposes, models were calibrated using both deterministic and probabilistic station catchments. Initial model runs established that assigning each postcode to its nearest station by distance, rather than time, produced the best performing models with deterministic catchments; and a one-minute uncongested drive time for workplace population was found to be the optimum threshold. The power distance decay function performed consistently better than the exponential time decay function.

The models with probabilistic catchments performed better, in terms of R^2 and AIC, than those with deterministic catchments. The best model overall, model 9, was based on probabilities derived using the station choice model with the accessibility term included. Greater weight was given to the population variable in the models with probabilistic catchments, and reduced weight was given to variables related to station services and characteristics. This suggests that the more realistic representation of the catchment in these models, enables differences in number of trips to be better explained. As a consequence, these models should be more transferable and better suited for use as a national predictive model. The models developed here are the first to successfully incorporate probabilistic station catchments into a trip end model, and represent the only example of a national-scale trip end model that has defined the zonal unit at such a high spatial resolution. Furthermore, it is the first time that a trip end model has been calibrated using a dataset of this size and geographic scope, in that it incorporates nearly every local station in England, Wales and Scotland.

[15] For the probability table where the choice set includes the nearest major station, some choice sets will contain 11 alternatives.

While the in-sample model performance measures suggest that the enhanced models should perform better at predicting demand for new local stations, it is important that they are tested under real-world scenarios. The next chapter introduces a methodology that has been developed to generate the station choice and trip end model inputs under the changed circumstances that result from a proposed new station or new line, and then goes on to describe two case studies where the methodology has been applied to forecast demand for several recently opened stations and a newly constructed line.

Chapter 8

Model application and appraisal

8.1 Introduction

To investigate the predictive performance of the integrated trip end and station choice model described in the previous chapter, and to assess whether probabilistic catchments can produce more accurate estimates of station demand, the calibrated models were used to forecast demand at several recently opened stations. This chapter begins by considering how the integrated model would be applied in the context of the typical appraisal process used to assess new local rail schemes (Section 8.2). It then describes the methodology that was developed to generate the station choice and trip end model inputs under the changed circumstances that result from a new station or new line being introduced (Section 8.3). Two case studies where this methodology was applied are then presented. The first considers three individual stations that have opened since the calibration base year of 2011/12 (Section 8.5); and the second relates to a railway line that opened in 2015, consisting of seven new stations (Section 8.6). A proposed methodology to forecast abstraction of demand from existing stations is then outlined, and an example application relating to a possible new station in north-east Wales is presented (Section 8.7). The chapter then describes how the integrated station choice and trip end models, along with the forecasting methodologies, have been applied in a real-world assessment of 12 potential new railway stations sites carried out on behalf of the Welsh Government (Section 8.9). The chapter then closes by summarising the work completed and drawing some conclusions (Section 8.10).

8.2 Model application in the context of the scheme appraisal process

The integrated trip end and station choice model that has been developed is intended to be used as part of the appraisal process of local rail schemes for new stations or new lines, or

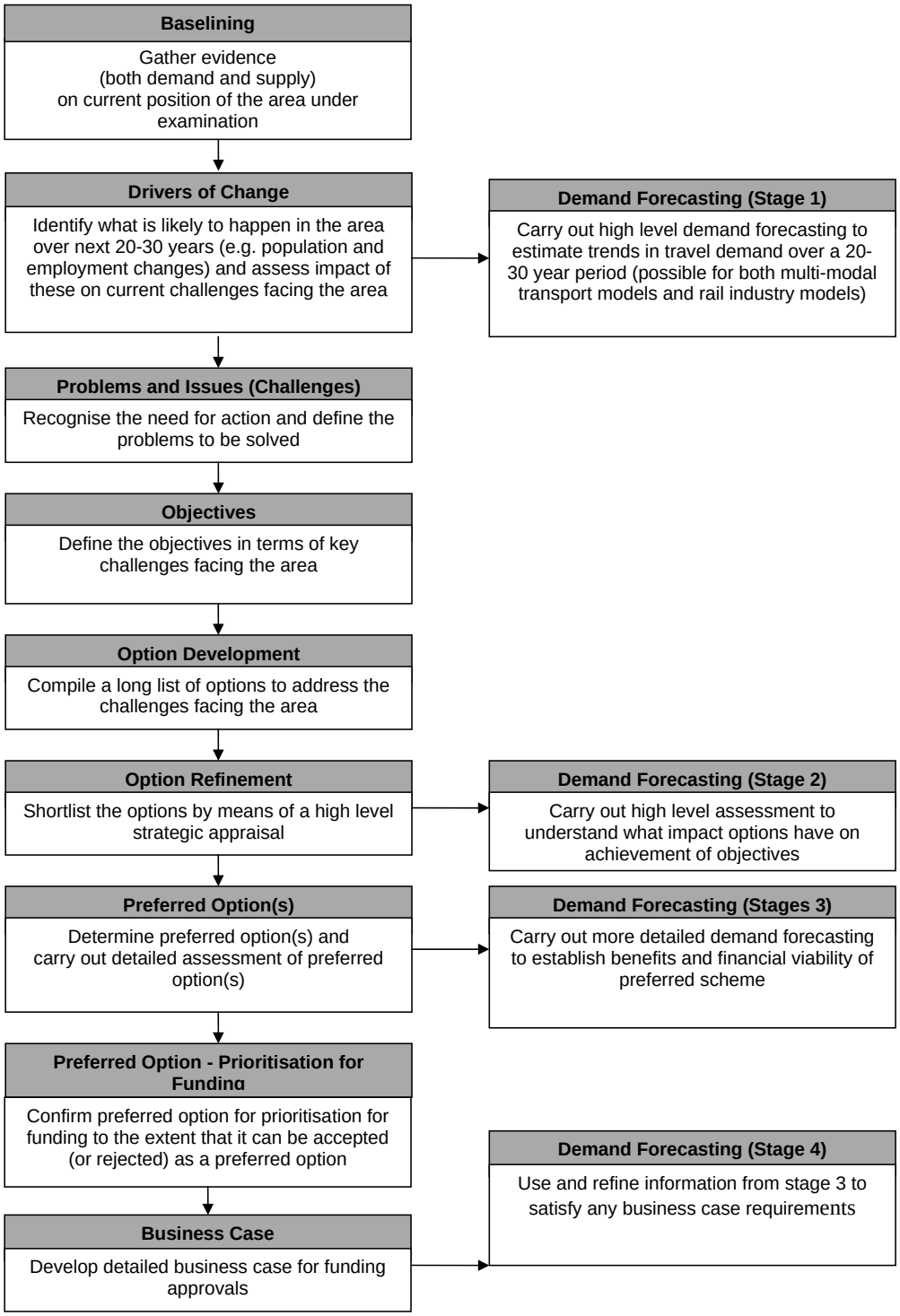


FIGURE 8.1: The process for a local rail scheme appraisal showing stages where demand forecasting should be carried out. Note: Reprinted from 'Guidance note on passenger demand forecasting for third party funded local rail schemes', by Department for Transport, 2011, p. 4. Reproduced under Open Government Licence v3.0.

where a non-incremental change to services at an existing station or stations is proposed. The generic components of the planning process of such a scheme, identifying the four stages that require a demand forecasting input, are shown in Figure 8.1 (Department for Transport, 2011). The integrated model is relevant to stages 2 – 4. These follow option development, once it has been established that rail is a feasible option to meet the scheme's objectives. There may be an early assessment of different options for new railway stations and a sifting process at stage 2, and the headline entries/exits forecast for each option may suffice at this stage, potentially alongside an analysis of abstraction from existing stations. For subsequent stages the entries/exits forecast would form a key input to the benefit-cost analysis, enabling the change in train operator revenue to be estimated. Crucially, the ability to generate probabilistic station catchments that can capture competition between stations should enable an estimate of abstraction from existing stations to be made (in contrast to current best practice discussed in Section 2.4). This would allow the net entries and exits resulting from a proposed station to be calculated as follows:

$$V_{net} = V_s + \sum_{r \in R} \Delta V_r, \tag{8.1}$$

where V_{net} is the forecast net entries and exits, V_s is forecast entries and exits for station s, R is the set of r stations at risk of abstraction by proposed station s, and ΔV_r is the estimated change in entries and exits at station r as a result of station s. This estimate of the net entries and exits resulting from the proposed station would be converted into train operating company revenue as part of the benefit-cost analysis.

In terms of the conventional business case framework of 'do nothing', 'do minimal', 'do something', the trip end model would provide a single input to the benefit-cost analysis of the 'do something' option. A range of other factors would need to be taken into account in quantifying the 'do something' scenario, but these are beyond the scope of this research project. The aim here is to improve this one key element. It should be noted that the 'do nothing' option must incorporate any assumptions that are made as part of the 'do something' option that remain valid, such as background growth in housing, jobs and rail passenger demand.

Having set out how the integrated trip end and station choice model is intended to be applied in assessing local rail schemes, the chapter will now go on to consider how well the model performs in the selected case studies and describe the development and assessment of a methodology for estimating abstraction from existing stations.

8.3 Methodology

The major consideration when seeking a workable methodology to apply the calibrated models is the process required to generate the station choice and trip end model inputs under

the changed circumstances that result from a new station being introduced. The previous chapter described the procedure used to define a set of ten alternative stations for each unit postcode in mainland GB, from which the choice probability of each station in each choice set was calculated and probabilistic catchments then derived. However, as these catchments were used to calibrate trip end models for station entries/exits in 2011/12, only stations which were open at that time were included in the universal set of stations from which the nearest ten were selected. It is therefore necessary to redefine the set of alternative stations available at each unit postcode when the models are applied, so that any recently opened stations, as well as the proposed new station(s), appear as available choices when appropriate. Given the computer processing overhead involved in creating the choice sets, generating predictor variables and calculating choice probabilities, it would not be practical to regenerate the nearest 10 stations for every postcode in mainland GB each time a new station needed to be modelled.

Analysis of the combined passenger survey dataset identified that only a very small number of reported station access journeys (0.83%) exceed 60 minutes, irrespective of the chosen access mode, as shown in Figure 8.2. This analysis excludes walk mode as any access journeys on foot > 60 minutes were previously removed from the dataset during trip validation (see Section 4.5.1.1). It was therefore decided that for any proposed new station the 'area of interest' could be justifiably limited to those unit postcodes within 60 minutes' drive time, as no meaningful demand would be generated from postcodes beyond this threshold.

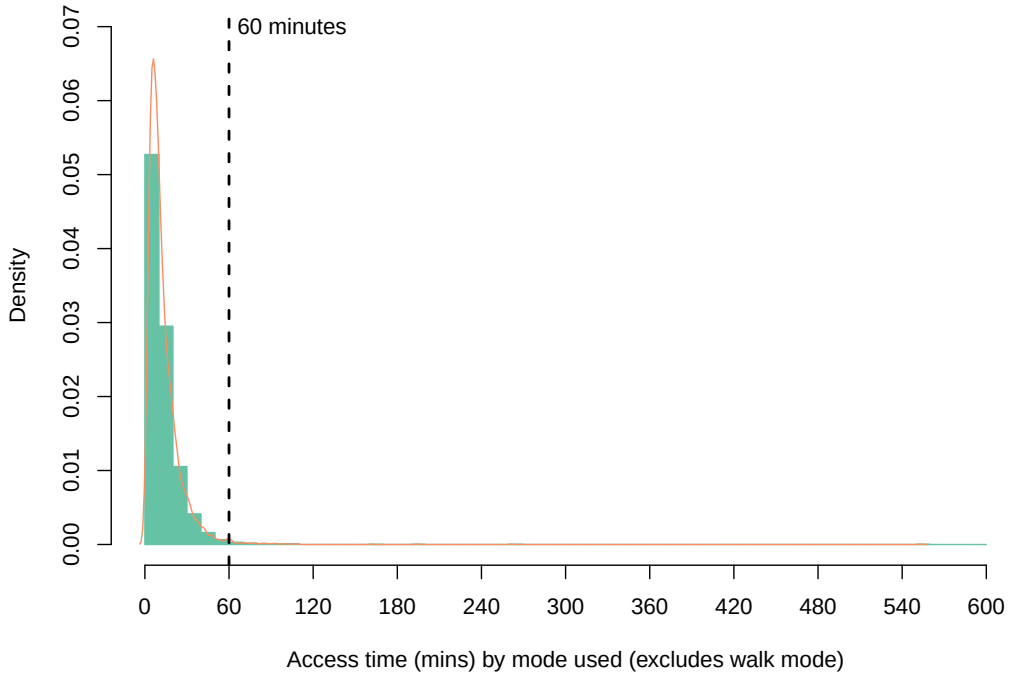


FIGURE 8.2: Histogram of access time to chosen station by reported mode (excluding walk mode) with kernel density plot.

Using this approach, the nearest 10 stations, selected from the universal set that now includes the proposed new station, to each of these postcodes can be readily generated. The proposed

station will *not* be present in the choice set of every postcode that is within the 60-minute threshold, as it will not always be amongst the nearest 10. Any postcode where the proposed station is not in the choice set can be discarded, as it will have no influence on the catchment definition. This further reduces the computing overhead involved in populating the probability database table and deriving the predictor variables.

The predictor variables required for the trip end and station choice model components will either be provided by the scheme proposer (for example, service frequency or parking spaces) or can be readily generated (for example, workplace population within one-minute drive-time). Calculating the accessibility term for each station in each choice set is a more time-consuming process, as it is necessary to generate all possible station pairs across the choice sets, eliminate any station pairs that are already known about (from the calibration exercise), obtain the distance between the remaining stations pairs using an OTP API lookup, and finally append them to the station-pairs database table.

Once the predictor variables for both model components have been obtained, a probability table can be created for the proposed station and the required trip end model run, with the weighted population input generated on-the-fly from the database as previously described in Section 7.5.3.4. The key steps involved in the proposed methodology are summarised in Figure 8.3.

In the case of a proposed railway line that consists of several new stations, each station catchment has the potential to be influenced by interaction with the other new stations. All the stations must therefore be modelled concurrently. In this situation the methodology can be streamlined using the following approach:

- When the 60-minute drive-time service area is generated for each station on the new line, the option to merge the polygons is selected. This creates a single polygon that encompasses the extent of all the individual service areas.
- All postcode centroids within the merged polygon are selected and the nearest 10 stations to each are obtained using an origin-destination matrix analysis.
- In R, any postcode where none of the proposed stations are present in its choice set is removed.
- A single probability table is then created and populated.
- When the trip end model is run to generate the demand forecast for a *specific station* on the new line, the database query that pulls the weighted population from the probability table only considers those postcodes that are within 60 minutes of that *specific station*.

This approach avoids creating a separate probability table for each station, and eliminates the duplication of postcodes, that are within 60 minutes of more than one station, across multiple tables.

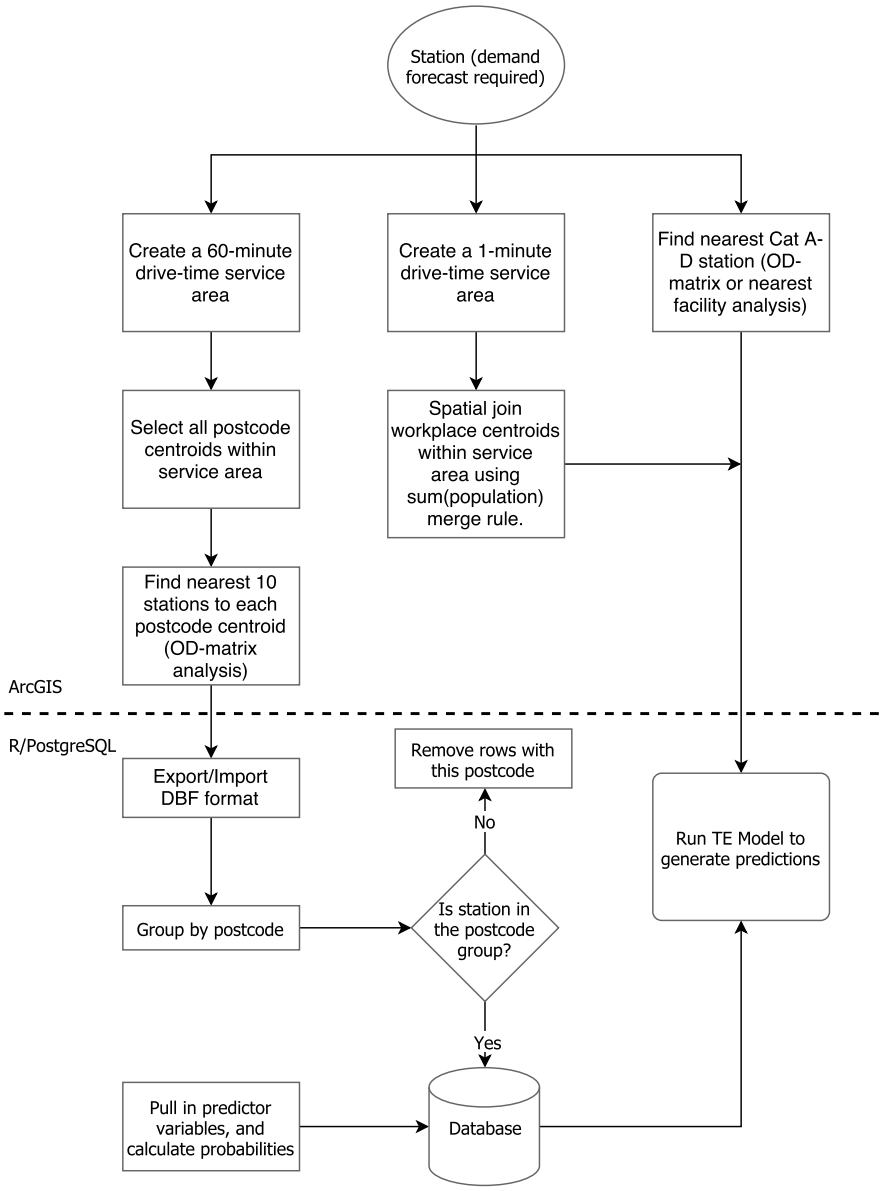


FIGURE 8.3: Proposed methodology for generating a demand forecast for a new station.

8.4 Demand forecast considerations

As the source of the postcode population data is the 2011 census, the number of station entries/exits in 2011/12 was used as the dependent variable in the trip end models. To account for growth in rail travel over recent years, it is possible to apply an uplift to the demand forecasts. In the case studies that follow, this was calculated separately for stations in Scotland and Wales. For stations located in Wales, the percentage change in the total number of entries/exits for Welsh stations in the calibration dataset between 2011/12 and 2015/16, calculated as 10.48%, was applied as the uplift. In the case of Scottish stations, the equivalent figure of 14.3% was used. This adjusted forecast should be treated with some caution, as part of the growth in journeys will have been driven by population growth over

this period. This population growth will depend on a range of regional and local factors, and will impact individual stations differently. As this aggregate change may not reflect the circumstances at the stations considered in the case studies, demand forecasts before and after applying the uplift are reported. For reasons of brevity the uplift is only applied to forecasts made using trip end model 9 (see Table 7.6).

When comparing forecast demand with actual demand it should be noted that the trip end models have been calibrated using stations that, with a few exceptions, are well-established and have been open for many years. There is evidence to suggest that, discounting for any general growth in rail journeys that might be occurring, it can take several years for a new station to reach its potential, as individuals adjust their behaviour over time. This might be through delayed mode change for existing trips (e.g. switching from car or bus), generation of additional trips as awareness grows of faster and less stressful journeys for work or shopping, or even by influencing decisions on where people live or work. Preston and Dargay (2005) found that this period may last for up to five years, while analysis by Blainey (2009) suggests that demand at new stations might increase relative to other stations in the surrounding area for up to six years, with this difference becoming smaller over time. However, Blainey also found a large variability in the effect between stations, and consequently a weak linear relationship between time (in years) and the growth difference (R^2 of 0.088). It is therefore difficult to predict the nature of this effect at a specific station with any confidence, although it should be borne in mind when comparing forecast demand with actual, especially in the initial years.

8.4.1 Catchment maps

Deterministic and probabilistic catchment maps have been produced for the case study stations, using postcode polygons from the OS 'Code-Point with polygons' dataset. These maps use a choropleth to indicate the probability that the proposed station will be chosen for each postcode within the station's catchment. It should be noted that this only indicates the probability of a station being chosen by someone located in a specific postcode if they were to choose to travel by rail; it does not indicate the likelihood of someone choosing to travel by rail over other modes. To aid clarity, a transparent fill is applied to those postcodes where the probability of the station being chosen is $< 1\%$. As only those postcodes included in the 2011 census population releases have been used in this work, gaps will occur in the catchment maps where corresponding data is not available for a particular postcode polygon[1]. It should also be noted that the scale of the catchment maps varies by station.

[1]These might be postcodes that have been introduced since 2011, or postcodes that were not present in the 2011 census resident population dataset as no resident population was assigned to them (e.g. a large user (business) postcode).

8.5 Case study A — new individual stations

The methodology was initially applied to forecast demand for three new stations that opened in 2012 (Fishguard & Goodwick) and 2013 (Conon Bridge and Energlyn & Churchill Park). The predictor variables entered into the trip end models for these stations, apart from weighted population, are summarised in Table 8.1. Demand forecasts were calculated based on simple (deterministic) station catchments (using model 7 in Table 7.6) and probabilistic stations catchments (using models 8 and 9 in Table 7.6). The demand forecasts obtained using the three trip end models are presented in Table 8.2, along with the weighted population input for each model, and the actual station usage data for 2015/16 obtained from ORR (the latest available at time of writing). The forecasts before and after applying the growth uplift are reported.

Station	work pop. (1 min)	Daily service freq.	Car park spaces	Nearest Cat A-D station (km)	Terminus station (0/1)
Conon Bridge	924	24	0	19.65	0
Energlyn & Churchill	0	56	18	1.97	0
Fishguard & Goodwick	876	14	0	24.60	0

TABLE 8.1: Predictor variables for stations (Case Study A).

8.5.1 Appraisal

8.5.1.1 Conon Bridge

All three models over-forecast demand at Conon Bridge, by around 60% before the growth uplift, with the probabilistic catchment models performing slightly worse than the deterministic model. However, all the models performed better than the reported original project forecast of 36,000 trips (Alderson & McDonald, 2017), which is more than double actual station usage in 2015/16. The deterministic and probabilistic catchments for Conon Bridge are shown in Figure 8.4. The probabilistic catchment indicates that Dingwall and Muir of Ord stations will attract passengers from many of the postcodes that are actually closer to Conon Bridge, especially those not in the immediate vicinity of the station. This is most likely due to the availability of car parking at Dingwall (12 spaces) and Muir of Ord (34 spaces), when there is no official station parking provision at Conon Bridge. There is also a slightly higher service frequency at these stations (one or two extra trains per day) and Dingwall is staffed on a part-time basis.

Conon Bridge is somewhat unusual as the impetus for building the station appears to have been to alleviate the effect of disruption to the road network during two five-month periods

Station	Weighted catchment population			Actual trips	Trip forecasts								
						Base year 2011/12						With uplift to 2015/16[b]	
	Distance decay	CMB-TE19 choice model	CMB-TE24 choice model	ORR 2015/16	Scheme forecast	Simple catchment	% diff from 15/16	Probability-based catchment (CMB-TE19)	% diff from 15/16	Probability-based catchment (CMB-TE24)	% diff from 15/16	Probability-based catchment (CMB-TE24)	% diff from 15/16
Conon Bridge	1249	859	856	15276	36000[a]	24453	60	25091	64	25090	64	28678	87.7
Energlyn & Churchill	3864	1183	1180	74206	unknown	73015	-2	75467	2	75329	2	83223	12.2
Fishguard & Goodwick	1992	1429	1416	19946	unknown	14345	-28	16317	-18	16387	-18	18104	-9.23

Notes: [a]Source: Alderson & McDonald, 2017; [b]uplift applied: Conon Bridge 14.3%; Welsh stations 10.48%

TABLE 8.2: Demand forecasts for three new stations and comparison with actual trips in 2015/16.

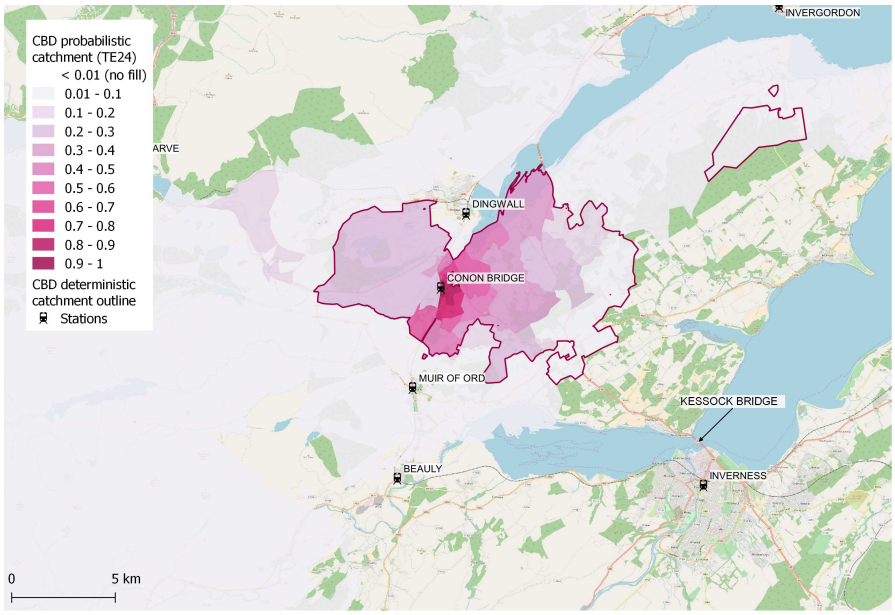


FIGURE 8.4: Deterministic and probabilistic catchments for Conon Bridge (CBD).

in 2013 and 2014 while the Kessock Bridge was repaired and resurfaced (BBC, 2012). As a result, station usage has actually fallen from 18,114 in 2013/14 (the first full reporting year) to 15,276 in 2015/16. Local media reports have highlighted service reliability issues at this station, with trains that are running late on this single-track line not stopping at Conon Bridge as scheduled, in order to make up time (North Star, 2014). It has also been suggested that passengers are preferring to drive to Dingwall station, where trains must stop due to signalling reasons, or may have abandoned rail altogether (The Inverness Courier, 2015). Given the competition with Dingwall identified by the probabilistic catchment, it is possible that some of the forecast demand at Conon Bridge has been drawn away or failed to materialise, as a result of these performance issues. This may go some way to explaining why actual demand is below forecast for this station.

8.5.1.2 Energlyn and Churchill Park

All three models produced a fairly accurate forecast for Energlyn & Churchill Park, within ±2% of actual trips before the growth uplift was applied. When adjusted for growth, the probabilistic model over-predicts actual demand by 12.2%. Actual demand grew by 6.94% between 2014/15 and 2015/16 at this station, which is substantially above the increase for the Welsh stations as a whole in the calibration dataset (1.82%). This may indicate the demand at this station is currently experiencing demand build-up, as discussed in Section 8.4, and if this effect was to continue into the third full reporting year (2016/17) and beyond, actual demand may prove to be closer to the model forecast than current data suggests. The deterministic and probabilistic catchments for Energlyn & Churchill Park are shown in Figure 8.5. The probabilistic catchment indicates that nearby stations will attract passengers from

many of the postcodes that are actually closer to Energlyn and Churchill Park, especially those not in the immediate vicinity of the station. This would be expected, as Aber, Caerphilly and Llanbradach stations all have substantially higher service frequency patterns (almost double the number of daily services), and Aber and Caerphilly have larger car parks, with 128 and 222 spaces respectively, compared to only 18 spaces at Energlyn & Churchill Park.

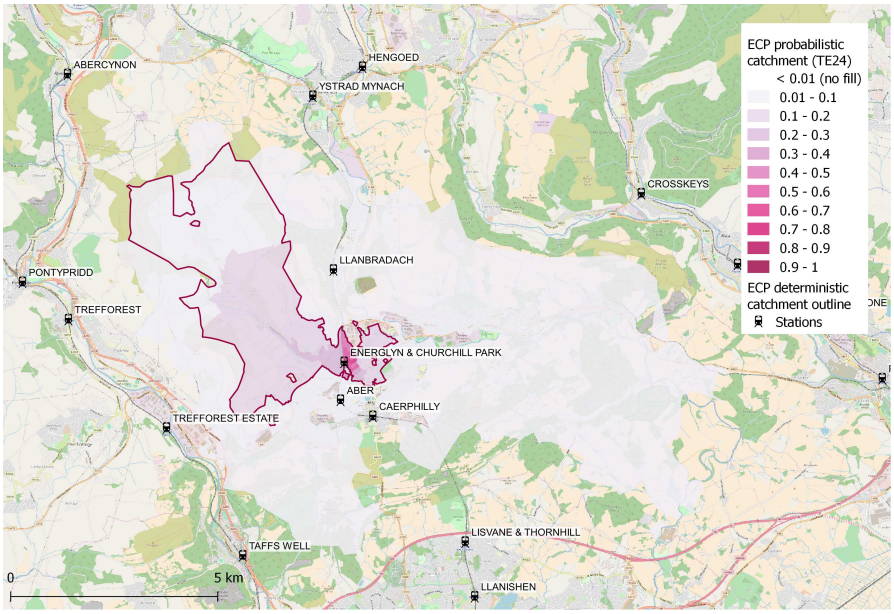


FIGURE 8.5: Deterministic and probabilistic catchments for Energlyn & Churchill Park (ECP).

8.5.1.3 Fishguard and Goodwick

Before applying the growth uplift, the probabilistic models under-forecast demand by 18% at Fishguard & Goodwick, although this represents a 10 percentage-point adjustment (in the desired direction) compared to the deterministic catchment model. Once the growth uplift is applied, the forecast is within 10% of actual demand. It is also worth noting that this station has been open longer than the other two, almost four full reporting years, and is likely to have reached its 'steady-state' demand.

The deterministic and probabilistic catchments for Fishguard & Goodwick are shown in Figure 8.6. The geographic placement of Fishguard Harbour station results in virtually all the postcodes being assigned to Fishguard and Goodwick station in the deterministic catchment. However, the probabilistic catchment indicates the likelihood of competition with Fishguard Harbour station in the area surrounding the two stations; as well as competition with more distant stations on the margins of the catchment.

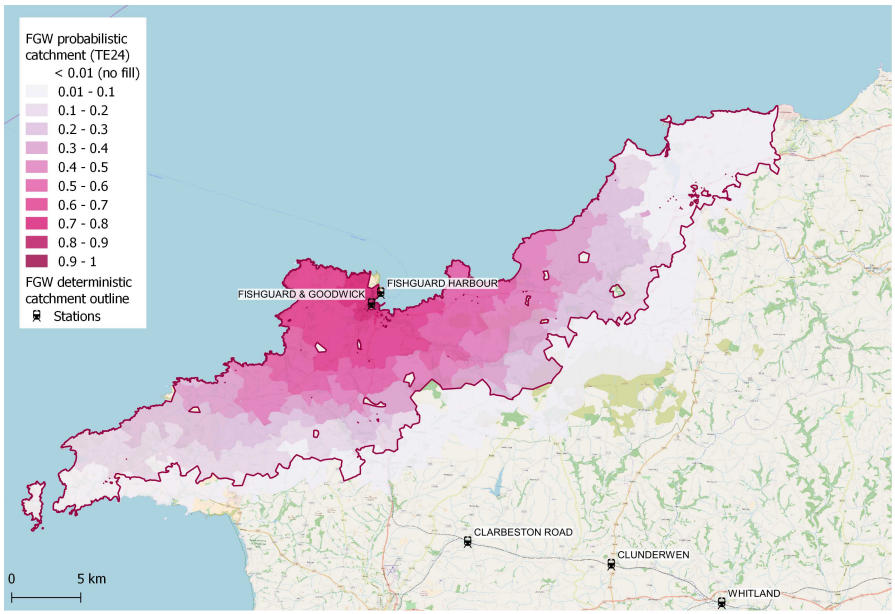


FIGURE 8.6: Deterministic and probabilistic catchments for Fishguard & Goodwick (FGW).

8.6 Case study B — a new railway line

The methodology was next applied to forecast demand for the seven stations that were built as part of the new Borders Railway in Scotland, which opened in September 2015 and runs from Edinburgh Waverley to Tweedbank (see Figure 8.7 (Wikipedia contributors, 2018)). The line passes through two pre-existing stations serving the Edinburgh suburbs of Brunstane and Newcraighall, and then the seven new stations, comprising four in Midlothian (Shawfair, Eskbank, Newtongrange, and Gorebridge) and three in the Scottish Borders (Stow, Galashiels and Tweedbank).

Station	Work pop. (1 min)	Daily service freq.	Car park spaces	Nearest Cat A-D station	Terminus station	Cat. F	Ticket mach.	Buses	CCTV
Tweedbank	1120	66	235	54.89	1	1	1	1	1
Galashiels	3746	66	0	50.62	0	1	1	1	1
Stow	718	47	33	39.08	0	1	1	1	1
Gorebridge	2330	66	73	16.58	0	1	1	1	1
Newtongrange	1965	66	56	13.17	0	1	1	1	1
Eskbank	819	66	248	11.39	0	1	1	1	1
Shawfair	0	66	59	10.16	0	1	1	1	1

TABLE 8.3: Predictor variables for Borders Railway (new stations only; trip-end and/or station choice models).

The predictor variables entered into the trip end and/or station choice models for each of these stations, apart from weighted population, are summarised in Table 8.3. Demand forecasts were calculated based on simple (deterministic) station catchments (using model 7 in Table 7.6) and probabilistic station catchments (using models 8 and 9 in Table 7.6). The

FIGURE 8.7: The Borders Railway. Note: Reprinted from 'Borders Railway' by Wikipedia Contributors, 2018, January 12. ©User:Pechristener, Wikimedia Commons, CC-BY-SA-2.0.

demand forecasts obtained using the three trip end models are presented in Table 8.4, along with the weighted population input for each model, station usage data for the first 12 months and for the 2016/17 reporting year[2], and the business case forecast for the first 12 months which was produced in 2012. Demand forecasts with the growth uplift to 2015/16 applied are shown in Table 8.5. These two tables are summarised using bar charts in Figures 8.8 and 8.9.

8.6.1 Appraisal

The forecasts before applying the growth uplift, show that model 9 (incorporating probabilistic catchments) has performed reasonably well across all seven stations and has, with the exception of Galashiels, produced more accurate forecasts than model 7 (using deterministic catchments). The forecasts for three of the stations are within 20% of actual trips, with Tweedbank and Eskbank +9%, and Stow +17%. This is substantially better than the performance of model 7, where the forecasts for these three stations are +70%, +37% and +46% respectively. Looking at the aggregate prediction for the seven stations, model 9 predicts a total of 1.50 million trips, slightly higher (+10%) than the 1.36 million actual trips in 2016/17. This compares favourably with the 48% over-prediction obtained using model 7. Despite some shortcomings, such as the large over-forecasts for Gorebridge and Shawfair, it is particularly encouraging that model 9 has performed substantially better than the business case forecast. This is most apparent for the three Scottish Borders stations (Tweedbank, Galashiels and Stow) where the business case projections severely under-estimated demand.

[2]At the time of writing the official station usage data for the 2016/17 reporting year had not been released by ORR. The trip data was read from graphs provided in the 'Borders Railway Year 1 Evaluation' report (Transport Scotland, 2017) and the figures used are therefore only indicative of actual values.

Station	Weighted catchment population			Actual trips		Trip forecasts							
								2011/12 base year					
	Distance decay	CMB-TE19 choice model	CMB-TE24 choice model	First year from opening	Lennon data[1] 2016/17	Business case forecast	% diff from 16/17	Simple catchment	% diff from 16/17	Probability-based catchment (CMB-TE19)	% diff from 16/17	Probability-based catchment (CMB-TE24)	% diff from 16/17
Tweedbank	2476	2426	2420	337864	474000	43242	-91	806146	70	520157	10	515919	9
Galashiels	4737	4520	4527	201666	342000	46862	-86	200381	-41	157217	-54	158062	-54
Stow	700	697	699	48282	66000	11686	-82	96263	46	77841	18	77351	17
Gorebridge	3189	2856	2874	74891	93000	180038	94	254489	174	226058	143	226256	143
Newtongrange	3538	2612	2604	96735	137000	105836	-23	239277	75	209621	53	209019	53
Eskbank	5230	2873	2830	133121	228000	261050	14	312784	37	250757	10	248872	9
Shawfair	1323	320	324	16853	21000	123720	489	106627	408	65467	212	64979	209
Totals				909412	1361000	772434	-43	2015967	48	1507118	11	1500457	10

Notes: [1]Trip data read from graphs provided in the Borders Railway Year 1 Evaluation report, therefore figures are only indicative of actual values

TABLE 8.4: Demand forecast for Borders Railway (new stations only) and comparison with actual trip data in 2016/17.

Station	Lennon data 2016/17	Trip forecasts Adjusted for growth to 15/16 TE model 7 (simple catchment)	% diff from 16/17	TE model 8 (probabilistic catchment (CMB-TE19))	% diff from 16/17	TE model 9 (probabilistic catchment (CMB-TE24))	% diff from 16/17
Tweedbank	474000	921747	94	594748	25	589902	24
Galashiels	342000	229116	-33	179762	-47	180728	-47
Stow	66000	110067	67	89003	35	88443	34
Gorebridge	93000	290983	213	258475	178	258701	178
Newtongrange	137000	273590	100	239681	75	238992	74
Eskbank	228000	357637	57	286715	26	284560	25
Shawfair	21000	121917	481	74855	256	74297	254
Totals	1361000	2305057	69	1723239	27	1715622	26

TABLE 8.5: Demand forecast for Borders Railway with growth uplift of 14.3% applied and comparison with actual trip data in 2016/17.

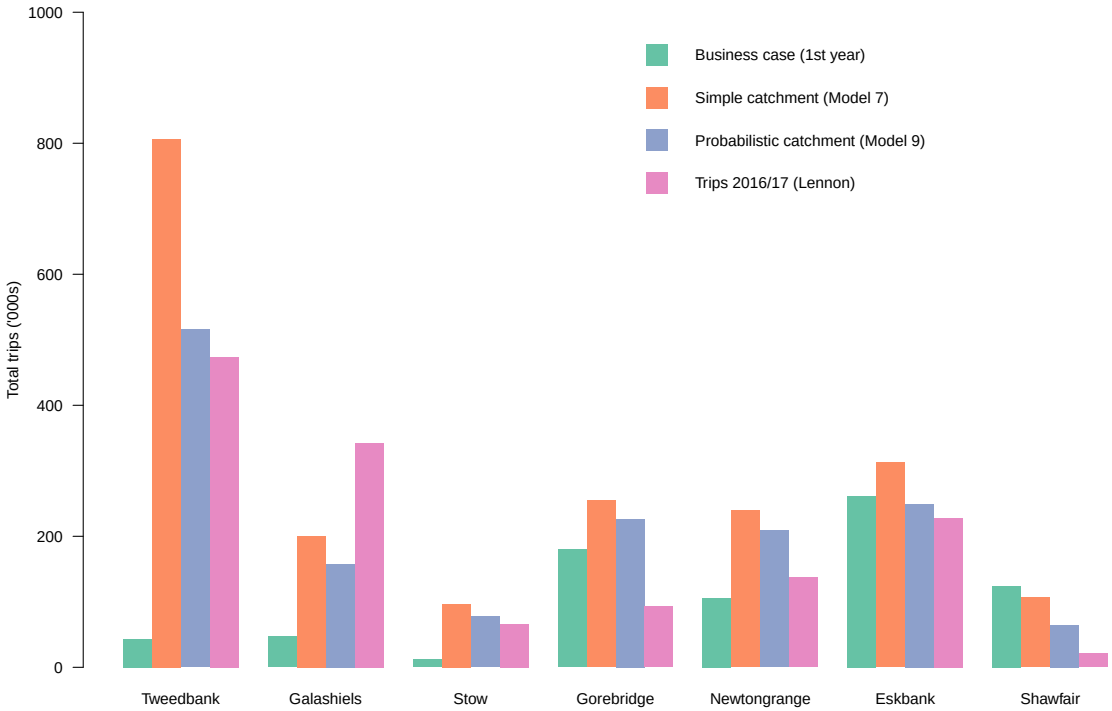


FIGURE 8.8: Comparison of demand forecasts (without growth uplift) and actual trips in 2016/17 for the new stations on the Borders Railway.

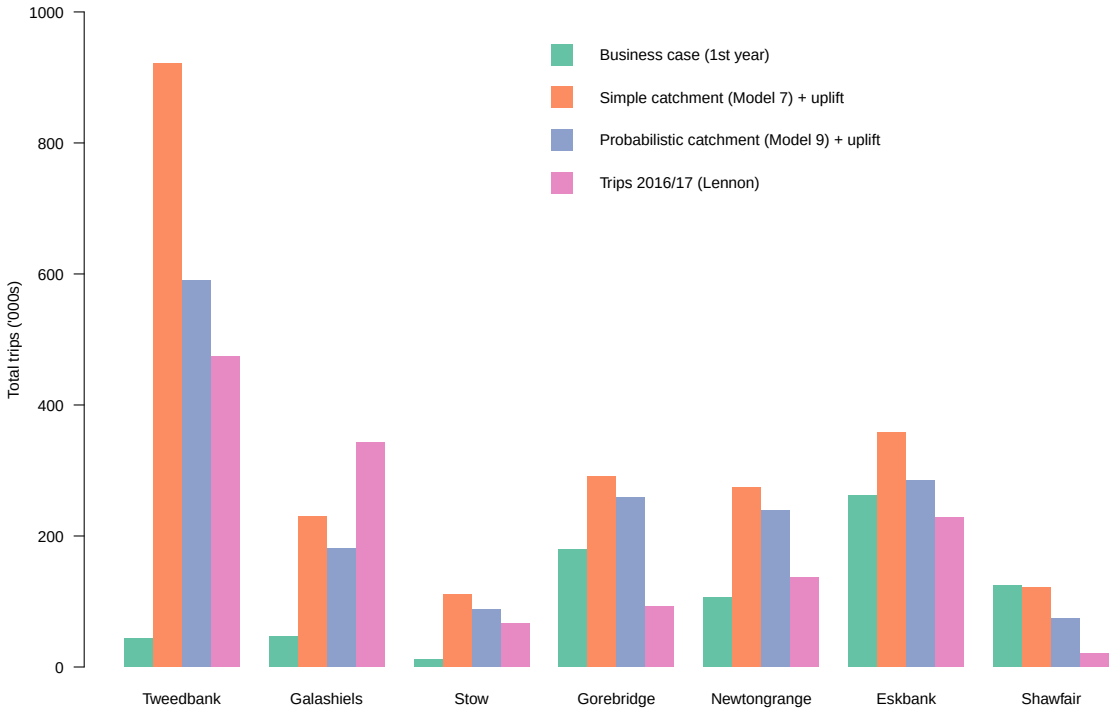


FIGURE 8.9: Comparison of demand forecasts (with growth uplift applied) and actual trips in 2016/17 for the new stations on the Borders Railway.

The impact of applying the growth uplift, with the exception of reducing the under-prediction at Galashiels, has been to raise the forecasts and so increase the difference from actual trips in 2016/17. However, it must be borne in mind that these stations are likely to be in the initial demand build-up stage. This is supported by the large difference in the number of trips in the first 12 months of operation (from September 2015) compared to the first full reporting year (from April 2016). There were 50% more trips in the latter period, indicating that demand build-up is taking place. If this continues into subsequent years then the 'steady-state' demand may be much closer to that predicted by the models.

The seven stations will now be considered on an individual or group basis, and some particular local circumstances that might have impacted the predictive performance of the models will be examined. In addition, the discussion will draw on information contained in a year-one evaluation of the new line carried out by Transport Scotland, which was informed by a survey of users and non-users of the line (consisting of 1,112 and 227 responses respectively) (Transport Scotland, 2017).

8.6.1.1 Tweedbank

Model 9 has noticeably corrected the large over-prediction for Tweedbank station produced by model 7, reducing it from +70% to +9% of actual trips (without growth uplift). Once the growth uplift has been applied, model 7 produces a forecast almost double actual demand (+94%), while model 9's forecast is +24%. In addition to the potential for further demand

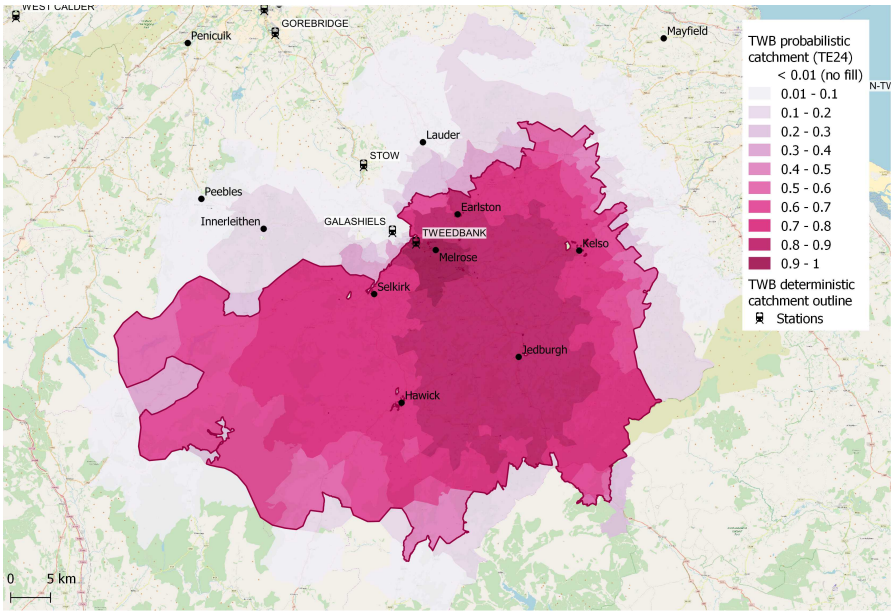


FIGURE 8.10: Deterministic and probabilistic catchments for Tweedbank (TWB).

build-up already mentioned above, there is some evidence that demand is being suppressed at Tweedbank, and this might also explain the apparent over-forecast. The first-year evaluation report (Transport Scotland, 2017) highlights capacity issues at the station car park, which required a temporary overflow car park to be provided, and survey responses indicate that some users, especially those from the Scottish Borders, have been discouraged from using the service due to reliability and capacity issues (these problems have been widely reported in the media, for example see Edinburgh Evening News (2015) and The Scotsman (2016)).

The deterministic and probabilistic catchments for Tweedbank are shown in Figure 8.10. The postcodes which make up the deterministic catchment all have moderate to high probability in the probabilistic catchment. The catchment is large with high probabilities maintained to the southern extent of the catchment[3], reflecting minimal competition from stations on other lines. However, the probabilistic catchment, along with those for Galashiels and Stow (see Figures 8.15 and 8.16), suggest that there is competition with Galashiels, and to a lesser extent Stow. The probabilistic catchment for Tweedbank also extends further north-west, encompassing Innerleithen and Peebles, and north-east beyond the extent of the deterministic catchment.

Figure 8.11 incorporates a map taken from Transport Scotland's first-year evaluation report of the Borders Railway, which plots the trip origins of surveyed passengers who boarded at each of the three Scottish Borders stations. This is useful empirical evidence that can be used to assess the realism of the probabilistic catchments generated by the station choice models. To aid interpretation, the map has been geo-referenced and the probabilistic and deterministic catchments for Tweedbank overlaid. While the observed trip origins of passengers boarding

[3]The reader is reminded that the extent of the catchment is limited to 60 minutes drive-time from the station as part of the demand forecasting methodology.

at Tweedbank (red dots) are predominantly located within the area of higher probability, as would be expected, they do also appear in Innerleithen and Peebles in the north-west, and east towards Berwick-upon-Tweed, outside of the deterministic catchment area, as predicted by the station choice model. In addition, there are trip origins for passengers who boarded at Galashiels station (blue dots) throughout the area of higher probability, in the towns of Selkirk, Hawick, Jedburgh, Earlston and Kelso. Again, this supports the station choice model, which has generally predicted a 10–25% probability of Galashiels station been chosen for trips originating from postcodes in these towns.

8.6.1.2 Galashiels

Prior to applying the growth uplift, model 9 has under-predicted demand at Galashiels by 54%, performing somewhat worse than model 7 (−41%), but still considerably better than the business case forecast. The performance of both models is improved when the growth uplift is applied (−47% and −33% respectively), although this gain is likely to be negated by further demand build-up before 'steady-state' demand is reached. There are, however, several factors that might help explain the poor performance of the model for this particular station, and these will now be considered below.

8.6.1.2.1 Car parking Unlike the other new stations on the line, Galashiels has no station car park. It is possible that the station choice model is penalising Galashiels excessively, attributing higher probabilities to Tweedbank and Stow than justified for some postcodes. The number of car parking spaces is also an important factor for generating trips in the trip end model, and the under-prediction by both the deterministic and probabilistic models may indicate that the trip end models generally perform less well in these circumstances. Alternatively, it could indicate that other parking opportunities are available that are not represented in the model but are being used by passengers boarding at Galashiels. This certainly appears a plausible explanation, as a new 'pay and display' car park with 43 spaces was built as part of the Galashiels Transport Interchange development (see Section 8.6.1.2.2) and is just a minute or two's walk from the station. Including these spaces in the trip end model would increase the demand forecast (after growth uplift) by some 120,000 annual trips[4], reducing the under-prediction to −12%. This should be treated with some caution as there are likely to be additional car parking facilities (including on-street parking) at many of the stations in the calibration dataset, and these were not accounted for in either the trip end or station choice models. However, the survey carried out as part of the first-year evaluation of the line, does show that 10% of passengers who boarded at Galashiels parked at the station (see Figure 8.12), suggesting that this might be a factor in the under-prediction, although this proportion is substantially lower than at the other stations.

[4]Note: this does not account for the impact of this car park on the station choice probabilities for this or other stations.

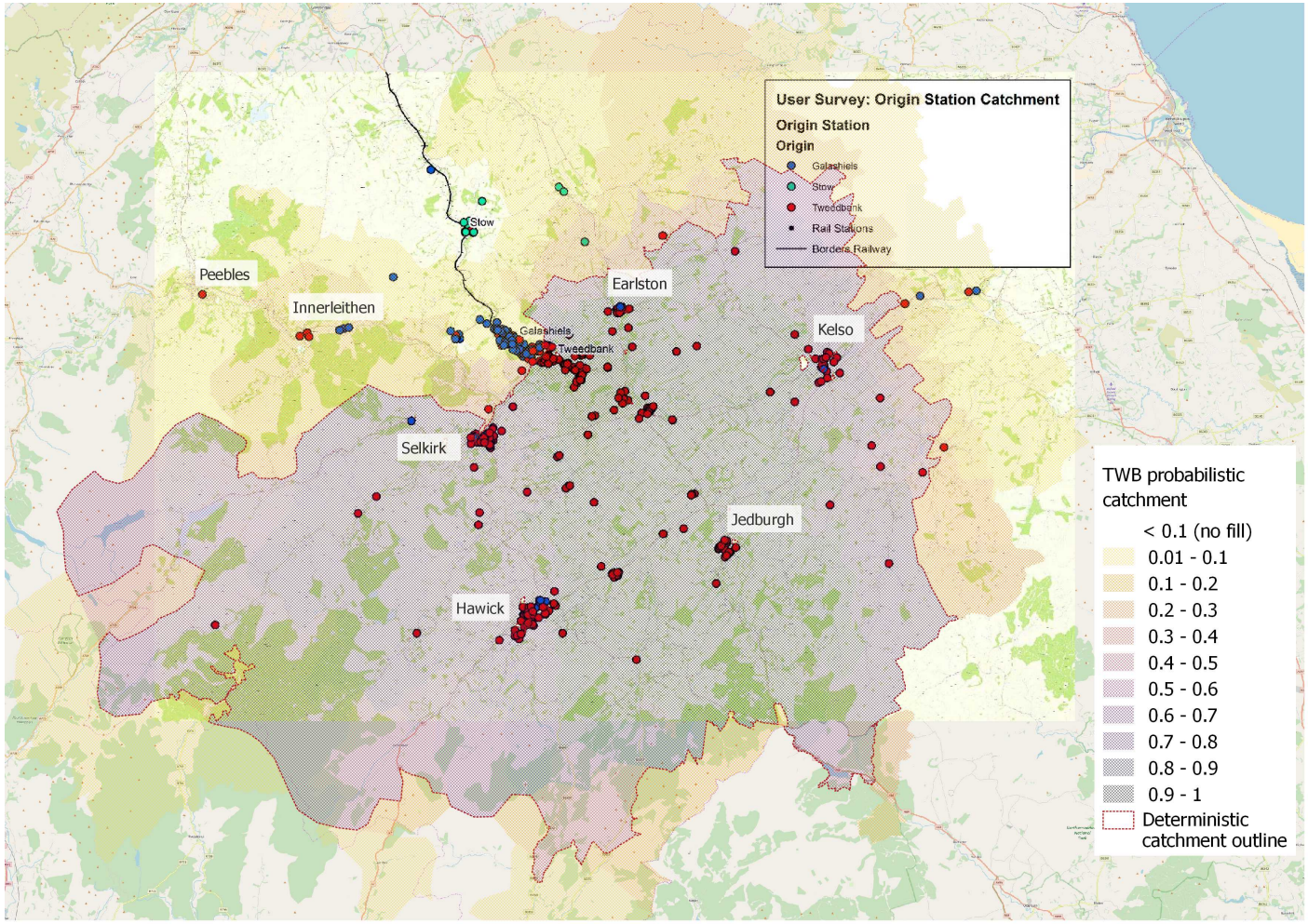


FIGURE 8.11: Observed origins of passengers boarding at each Scottish Borders station, from geo-referenced source map (Transport Scotland, 2017, p. 30) and overlaid with the Tweedbank probabilistic and deterministic catchments.

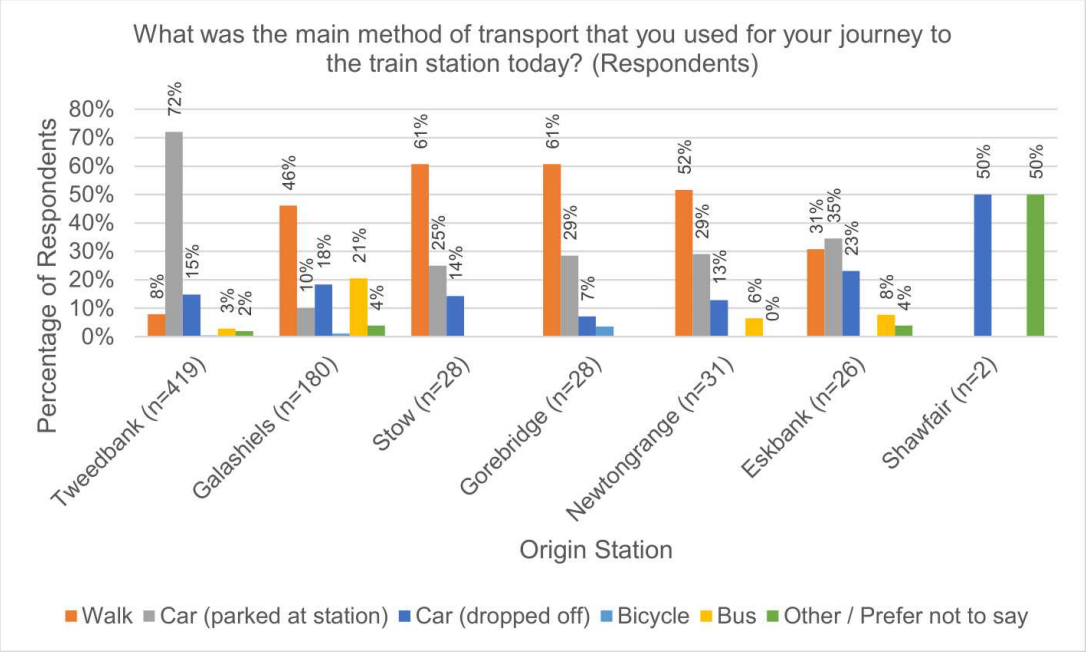


FIGURE 8.12: Reported station access mode for users of the Borders Railway. Note: Reprinted from '*Borders Railway Year 1 Evaluation*', 2017, p. 43. Reproduced under Open Government Licence v3.0.

8.6.1.2.2 Galashiels Transport Interchange As part of the Borders Railway project, a new Transport Interchange was built next to the new station at Galashiels, providing access to train and bus services (See Figure 8.13 (Wikipedia contributors, 2017)). The Transport Interchange, which has a range of facilities including a café, tourist information, showers and bike lockers, is being promoted as the 'gateway to the borders' and is a key hub for bus services in the region, with 1,400 bus departures in a typical week (Transport Scotland, 2016). Consequently, Galashiels is likely to be the preferred departure station for those using the bus for their access journey and travelling from many of the towns and villages in the Scottish Borders. This is supported by the first-year evaluation report, which found that 21% of survey respondents accessed Galashiels station by bus, compared with 3% for Tweedbank and 0% for Stow (See Figure 8.12 (Transport Scotland, 2017, p. 43)). The proportion accessing Galashiels station by bus is also particularly high when compared to the national average of 11% (Transport Focus, 2015a). Access by bus may explain why passengers who boarded at Galashiels originated from towns located south and east of Tweedbank, such as Selkirk, Hawick, Jedburgh, and Kelso, as indicated by the blue dots in Figure 8.14 which shows the Transport Scotland survey data overlaid with the probabilistic and deterministic catchments for Galashiels (also shown in Figure 8.15). These towns are nearer to Tweedbank station by road and Tweedbank has a large free car park, so it seems unlikely that many of those driving and parking would choose to board at Galashiels from these locations. The deterministic catchment for Galashiels is relatively small, primarily encompassing the immediate area around Galashiels itself and locations due west. The probabilistic catchment is far larger, and incorporates the towns from which passengers are known to have chosen to board at

FIGURE 8.13: The Galashiels Transport Interchange. Note: Reprinted from 'Galashiels' by Wikipedia Contributors, 2017, November 29. ©Walter Baxter, CC-BY-SA-2.0.

Galashiels. For example, the station choice model assigns a 25% probability of Galashiels station being chosen for postcodes in the centre of Hawick, and 14% for postcodes in the centre of Jedburgh. While the choice model does include a boolean variable indicating the presence or not of a bus interchange at a station, and this has a positive effect on utility, all the stations on the new line are recorded as having this facility. Therefore, the weighting attributed to bus interchange will be the same for Galashiels and Tweedbank, and the model may have under-estimated the likelihood of Galashiels being chosen. This suggests that a more nuanced measure of bus interchange may be preferable, as there is clearly a substantial difference between a dedicated bus interchange where multiple routes converge and a bus stop at a station which is served by a single service. Alternatively, if access mode choice could be adequately modelled, then mode-specific choice probabilities could be generated, which would be expected to increase the probability of Galashiels being chosen by those accessing a station by bus.

The number of car parking spaces is entered into both the station choice model and the trip end model. This can be justified on the basis that a station is more likely to be chosen if it has a car park (and if it has more spaces, as there is more likely to be a space available), and the larger a car park the more trips it is likely to generate. In the case of bus services, the increased likelihood of a station being chosen if it can be accessed by bus is captured in the station choice model (subject to the limitations already discussed), but there is no variable in the trip end model to capture the effect of more bus services generating more trips. This may be less important when access by bus is a relatively minor proportion of access trips (as is typically the case), but when it accounts for over 20%, as at Galashiels, the trip end model might under-estimate the number of trips. A possible solution could be to include bus service frequency, or a measure based on frequency and route diversity, into the trip end model.

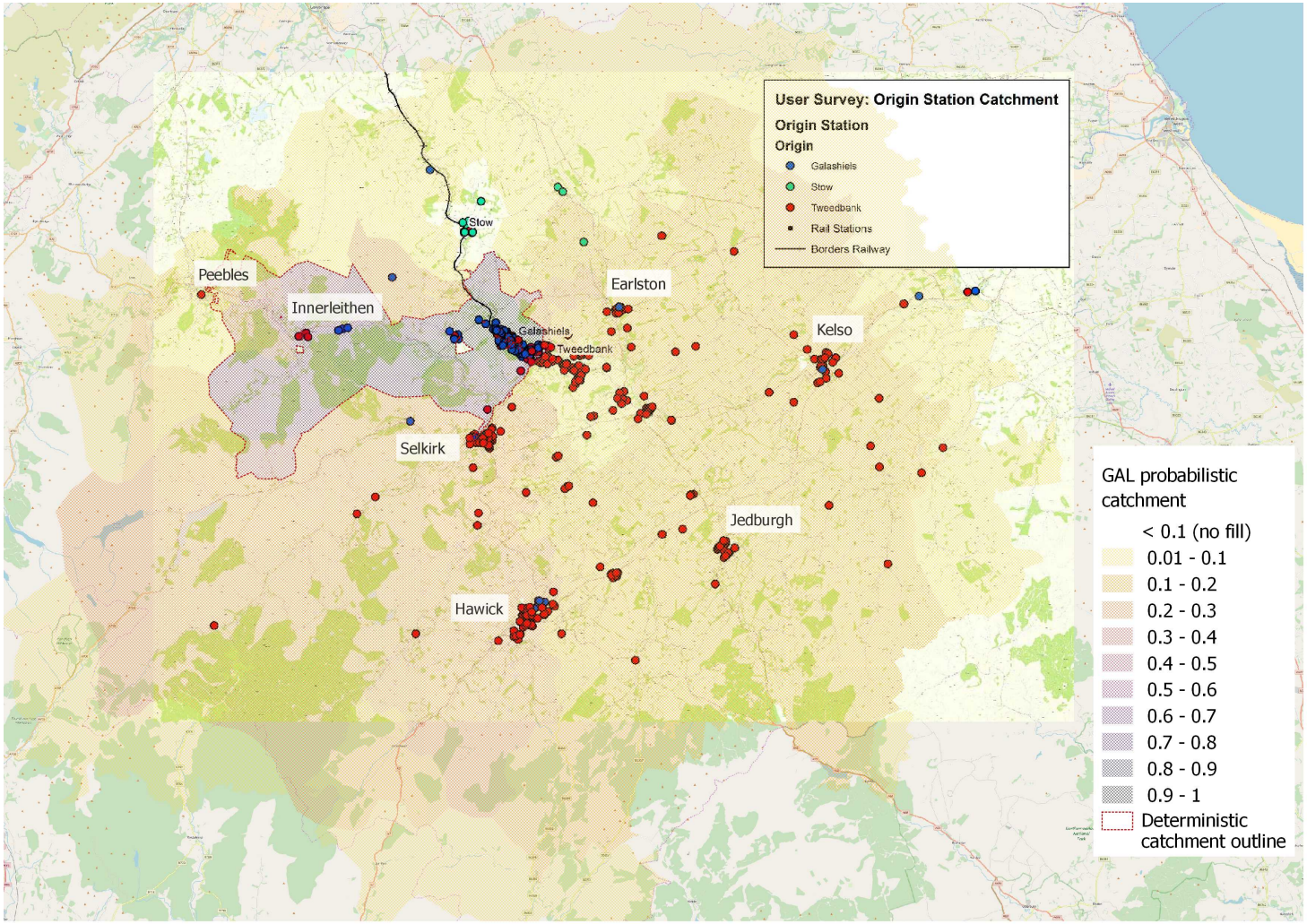


FIGURE 8.14: Observed origins of passengers boarding at each Scottish Borders station, from geo-referenced source map (Transport Scotland, 2017, p. 30) and overlaid with the Galashiels probabilistic and deterministic catchments.

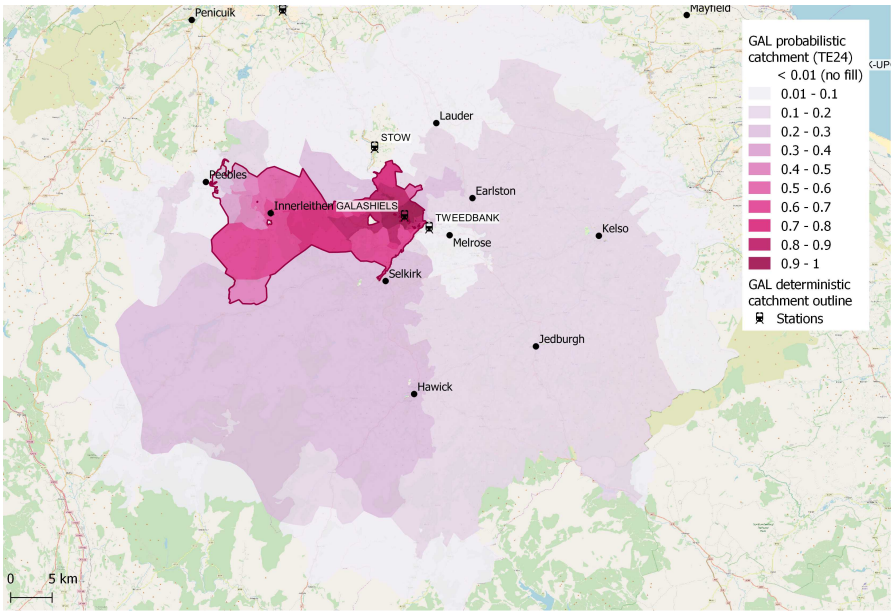


FIGURE 8.15: Deterministic and probabilistic catchments for Galashiels (GAL).

8.6.1.2.3 Tourism The final factor that may help explain the large under-prediction for Galashiels is the role of tourist trips. There is evidence that the opening of the new line has boosted tourism, particular in the Scottish Borders. The Scottish Tourism Economic Assessment Monitor (STEAM) statistics, which compared the number of visitor days in hotel and bed and breakfast accommodation in the first half of 2016 with the first half of 2015 (before the line opened), found an increase of 12.3% in Midlothian and 27% in the Scottish Borders (Midlothian Council, 2017). Of the passengers who responded to the first-year evaluation survey, which was completed in November/December and so outside of the peak tourist season, 39% said the purpose of their journey was either a tourist day trip or overnight stay. Once reported trip frequency is taken into account, this equates to 15% of annual single trips. While the majority of these were tourist day trips and overnight stays to/from Edinburgh (60% and 9% respectively), a significant proportion were day trips and overnight stays to the Scottish Borders (20% and 9% respectively). In contrast, only 2% of tourist trips were to Midlothian, with no overnight stays. Given that Galashiels is being promoted as the 'Gateway to the Borders', and bus services for onward travel are concentrated here, it is a reasonable assumption that a large proportion of tourist trips to the borders will be via Galashiels station. However, the only 'attraction' variable included in the trip end model is the workplace population within a one-minute drive of the station. The model is therefore likely to under-estimate demand at stations, such as Galashiels, where tourist trips form an important component of demand. The trip end model might be improved by incorporating a variable that could account for trips generated by visiting tourists. One possibility would be to include a measure of the number of available beds within a certain distance of the station, which could include hotels, bed and breakfast establishments and holiday rental properties.

8.6.1.3 Stow

Model 9 has produced a reasonably accurate demand forecast for Stow, bearing in mind the potential for demand build-up, with predicted trips +17% and +34% of actual trips before and after applying the growth uplift respectively, performing rather better than model 7 (+46% and +67% respectively). The deterministic and probabilistic catchments for Stow are shown in Figure 8.16. While the postcodes with the highest probability of choosing Stow also form the deterministic catchment, the probabilities suggest some competition with other stations, and this is supported by the presence of a few origins of passengers who boarded at Galashiels and Tweedbank, from the first-year evaluation survey, within the deterministic catchment (see Figure 8.17). This might partly reflect the difference in service frequency, as 19 fewer trains serve Stow on a typical weekday. The extent of the observed catchment for Stow is likely to be less reliable than that for Tweedbank or Galashiels, as relatively few passengers in the survey boarded at Stow.

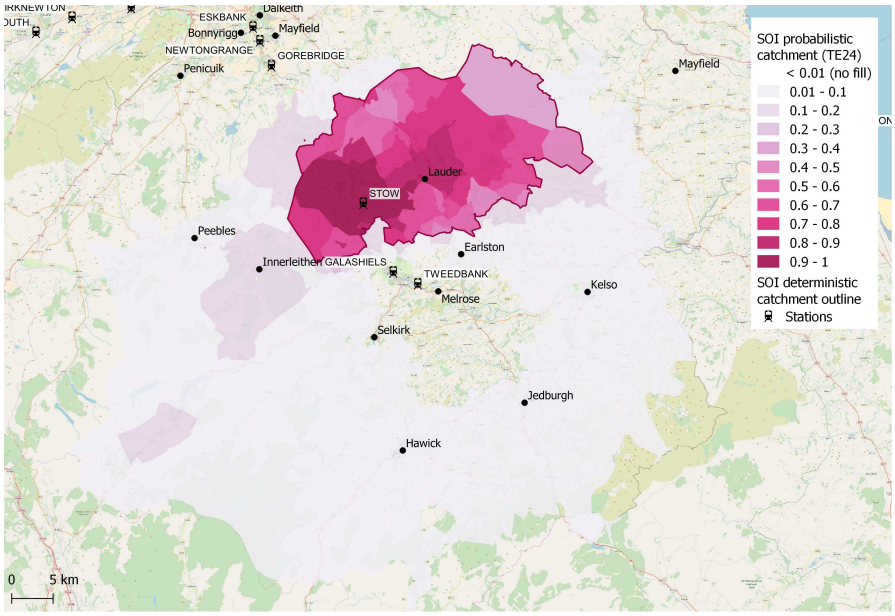


FIGURE 8.16: Deterministic and probabilistic catchments for Stow (SOI).

8.6.1.4 Midlothian stations

The Midlothian stations will be considered together, as a common factor may have contributed to the over-forecasting of demand at several of these stations. Prior to applying the growth uplift, model 9 substantially over-forecast demand at Gorebridge and Shawfair stations, by 143% and 209% respectively, although this does represent a considerable improvement over model 7, which over-forecast these stations by 174% and 408% respectively. Model 9 performed rather better for Newtongrange (+53%) and forecast demand at Eskbank to within 10% of the actual number of trips (+9%).

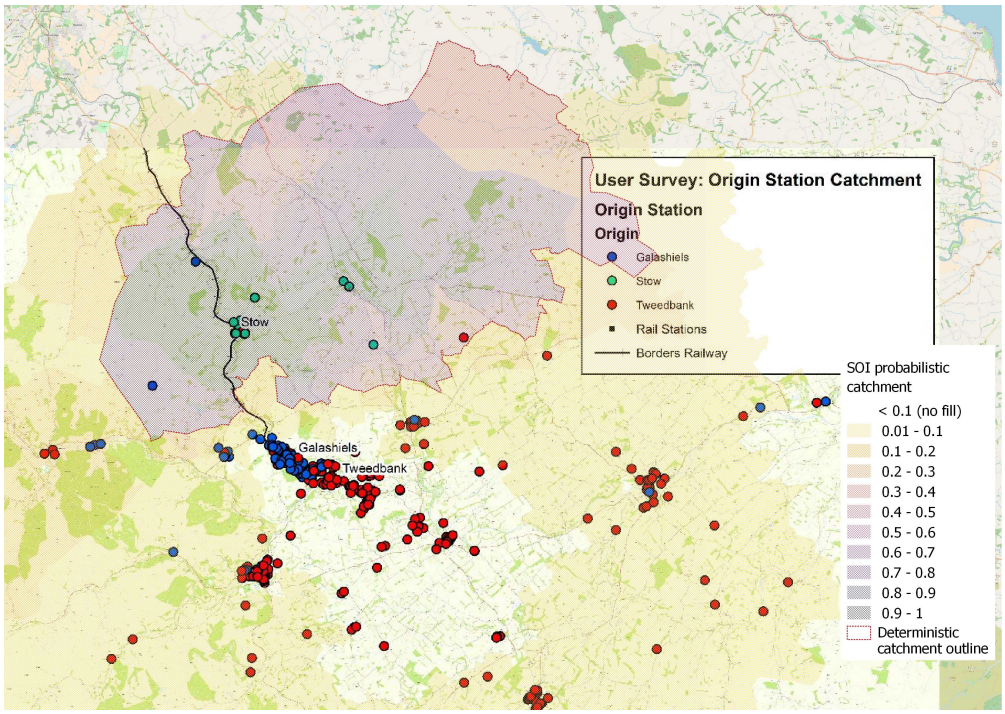


FIGURE 8.17: Observed origins of passengers boarding at each Scottish Borders station, from geo-referenced source map (Transport Scotland, 2017, p. 30) and overlaid with the Stow probabilistic and deterministic catchments.

The first-year evaluation study asked non-users and one-off users of the Borders Railway line to indicate their major and minor reasons for not using the line, or for not using it more frequently. The responses reveal a marked difference between residents of the Scottish Borders and Midlothian. In the Scottish Borders, the fact that buses were cheaper, more convenient and allow the use of the National Entitlement Card[5] were identified as major reasons by 19%, 12% and 16% of respondents respectively. In contrast, these were identified as major reasons by 46%, 37% and 27% of respondents from Midlothian respectively. A potential reason for this difference is the flat-rate single fare of £1.60 (at time of writing) offered by Lothian Buses, which is valid as far as Gorebridge. This is extremely competitive when compared to the cost of travelling by train from Gorebridge to central Edinburgh. The single train fare is £5.50, almost 3.5 times the bus fare. The bus fare also compares favourably to the single train fare from Newtongrange, Eskbank and Shawfair stations (£4.80, £4.50 and £3.40 respectively). By way of contrast, the train fare from Galashiels to Edinburgh is £9.60, which is only 1.4 times the cost of the bus fare (£6.90), and potentially offset by the greater travel time savings. It is therefore possible that competition from local bus services is suppressing demand at the Midlothian stations, and as the trip end model is a rail-only model it is unable to take account of competition from other modes. However, this does not explain why the model's forecast for Eskbank is reasonably accurate. One possible explanation may relate to differences in the socio-demographics of the station catchment

[5]The National Entitlement Card provides free bus travel throughout Scotland for those aged over 60 or with eligible disabilities, and reduced fares for those aged 16–18.

population. For example, Eskbank appears to be a more affluent area, with an average house price of £318k, compared to £167k and £173k for Gorebridge and Shawfair[6], and may have a greater proportion of workers commuting to Edinburgh who are less sensitive to price.

It may be possible to improve the trip end model by incorporating a variable representing the differential in fare between bus and train, although it would be a challenge to identify the most appropriate destination to use for this comparison, and other factors will be important determinants of bus patronage, such as journey time and service frequency. In view of these difficulties, this issue might be better addressed through the use of flow models. Previous work by Blainey and Preston (2010) attempted to calibrate flow models that could capture the impact of bus competition on rail demand, by including a relative bus journey time variable. As the data was obtained manually from an online journey planner, only a small subset of flows could be included in these models, and a counter-intuitive parameter sign was also obtained, suggesting that as bus journeys become faster relative to the train, rail demand increases. Unfortunately, bus fare was not included in these models, due to the lack of available data. This approach could be revisited, using the framework that has been developed as part of this research (see Section 5.3) to automatically generate the necessary journey data. Bus fares remain problematic, as there is no national dataset containing this information. However, fares are now provided on the Traveline Scotland journey planner, and these could potentially be obtained and used to calibrate a flow model using Scotland as a case study.

The deterministic and probabilistic catchments for the Midlothian stations are shown in Figure 8.18 for Gorebridge, Figure 8.19 for Newtongrange, Figure 8.20 for Eskbank; and Figure 8.21 for Shawfair. The observed origins of passengers who boarded the Midlothian stations, obtained from the first-year evaluation survey are shown in Figure 8.22 (Transport Scotland, 2017, p. 29). As substantially fewer passengers in the evaluation survey boarded at the Midlothian stations compared to the Scottish Borders stations (8% of survey respondents were residents of Midlothian, and 60% were residents of the Scottish Borders), these origins are less likely to capture the full extent of the actual catchments. This is particularly the case for Shawfair, where there only appears to be a single recorded origin.

8.7 Forecasting abstraction from existing stations

In addition to generating a demand forecast for a proposed new station, it is important to assess the potential effect of the new station on demand at existing stations. A new station may abstract passengers from one or more existing stations, and the net change in demand across the new and existing stations could be substantially lower than the gross forecast for a new station alone might suggest. If the scheme appraisal process does not adequately account for abstraction from existing stations, it could result in a new station being built that

[6] Average house price data was obtained from `http://www.rightmove.co.uk/house` on 28/11/2017.

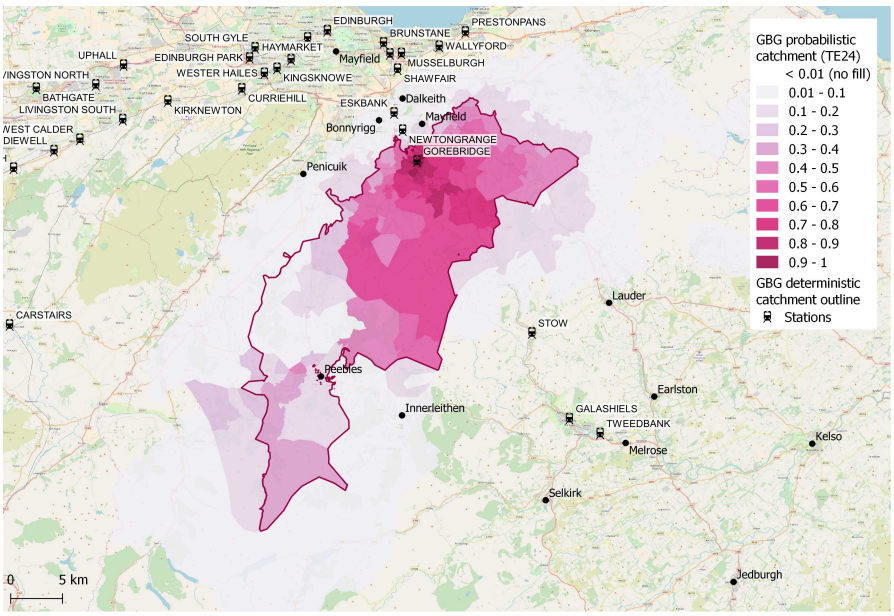


FIGURE 8.18: Deterministic and probabilistic catchments for Gorebridge (GBG).

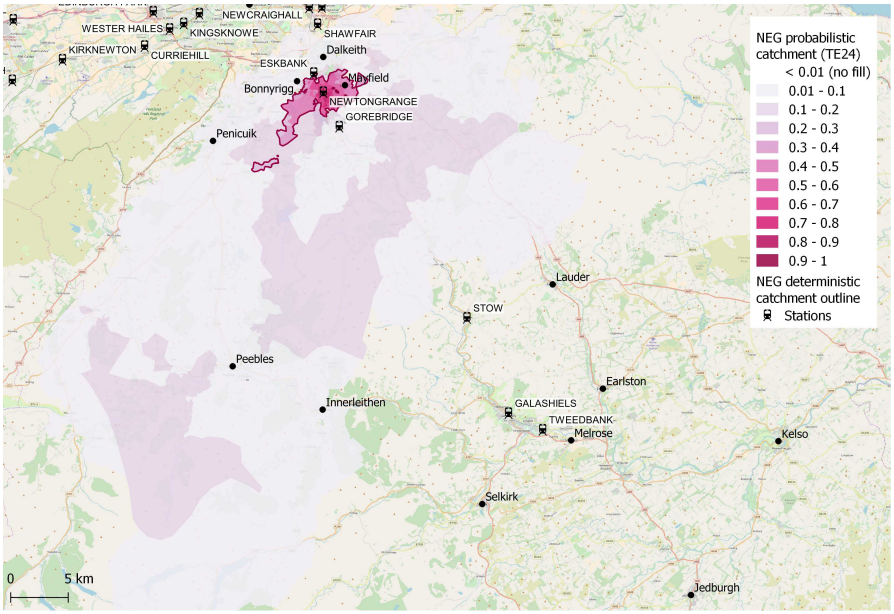


FIGURE 8.19: Deterministic and probabilistic catchments for Newtongrange (NEG).

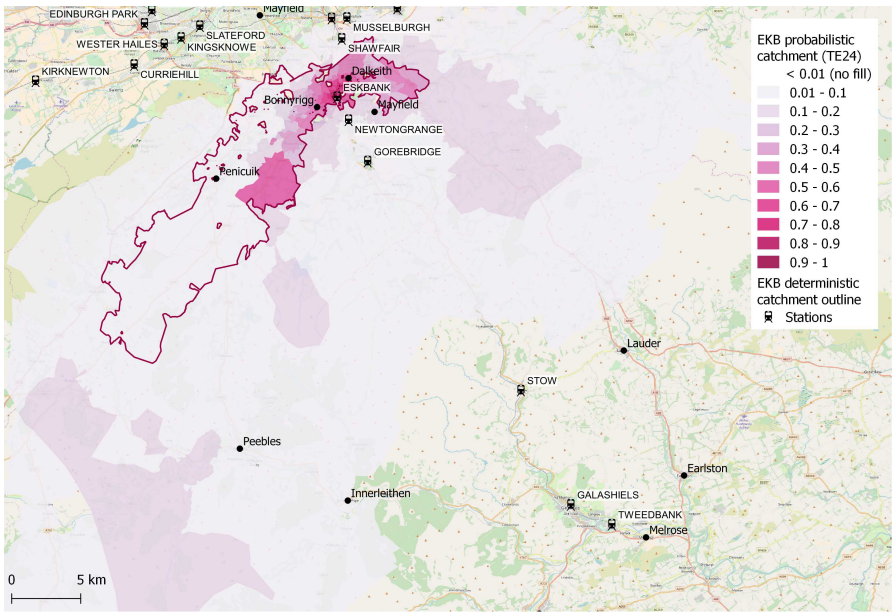


FIGURE 8.20: Deterministic and probabilistic catchments for Eskbank (EKB).

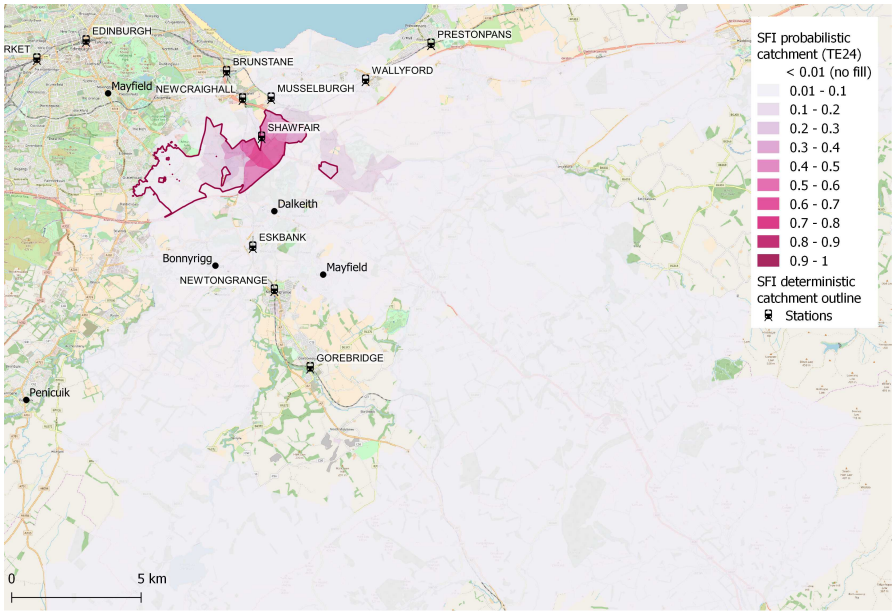


FIGURE 8.21: Deterministic and probabilistic catchments for Shawfair (SFI).

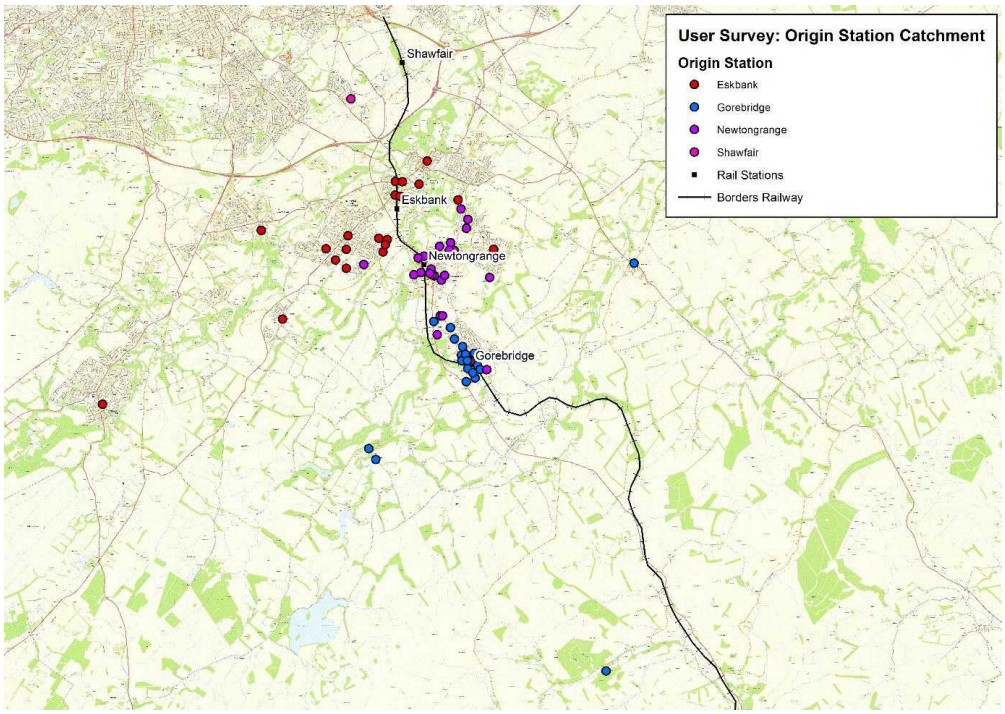


FIGURE 8.22: Observed origins of passengers boarding at each Midlothian station. Note: Reprinted from '*Borders Railway Year 1 Evaluation*', 2017, p. 29. Reproduced under Open Government Licence v3.0.

fails to deliver the expected economic and societal benefits. A methodology was therefore developed that can assess the potential extent of abstraction, based on the changes that occur to the probabilistic catchments of the affected station(s). The methodology consists of the following key steps:

- Identify the unit postcodes within 60 minutes drive-time of the station(s) identified as being 'at risk' of abstraction.
- For each 'at risk' station generate a 'before' choice set (selecting from current stations only) and an 'after' choice set (selecting from current stations *plus* the proposed new station) for each postcode.
- Create and populate separate probability tables for the before and after situation.
- Obtain the weighted population (applying the probability and distance weightings) for the 'at risk' station, in both the before and after situation and calculate the percentage change.
- Assume an elasticity of one between weighted population and the number of entries/exits, and apply the percentage change to the most recently reported annual entries/exists, thus giving an estimate of the abstraction effect. This is based on evidence in the PDFH relating to the external environment, and forecasting framework assumptions that

the population elasticity is equal to one for the number of trips originating in a zone (Association of Train Operating Companies, 2013, Chapter C1)[7].

This methodology was applied to assess the extent of abstraction that might result from several potential new stations in Wales, which formed part of a piece of consultancy work carried out for the Welsh Government (See Section 8.9 for more details). The example application presented here examines the abstraction effect of a proposed new station known as 'South Wrexham' (actually located in Rhosymedre), on the existing stations at Ruabon and Chirk. The results of the abstraction analysis, summarised in Table 8.6, show a substantial abstraction from Ruabon and, to a lesser extent, from Chirk. As the demand forecast for South Wrexham station was [redacted] annual entries/exits, the abstraction analysis suggests that over [redacted] of these trips [redacted] would be abstracted from Ruabon and Chirk. The effect of the new station on the probabilistic catchment for Ruabon station can be seen by comparing Figures 8.23 and 8.24, which show the catchment before and after the new station. While the proposed methodology has been successfully applied and appears promising, further work is needed to validate this approach, if possible against real-world observed abstraction effects.

Station	Weighted population (before new station)	Weighted population (after new station)	% change	Trips 2015/16	Adjusted trips	Change (trips)
Ruabon	3839	2312	-40	92986	55792	-37194
Chirk	1620	1381	-15	68444	58177	-10267

TABLE 8.6: Results of abstraction analysis for a new station at South Wrexham.

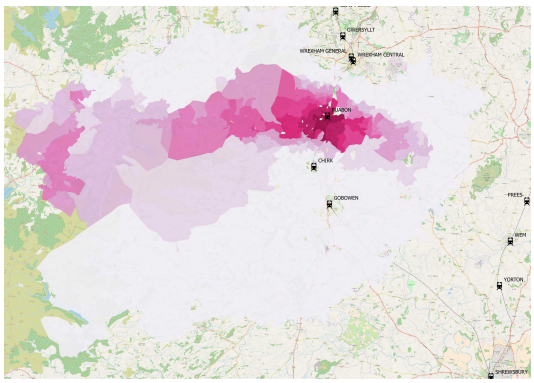


FIGURE 8.23: The existing probabilistic catchment for Ruabon station.

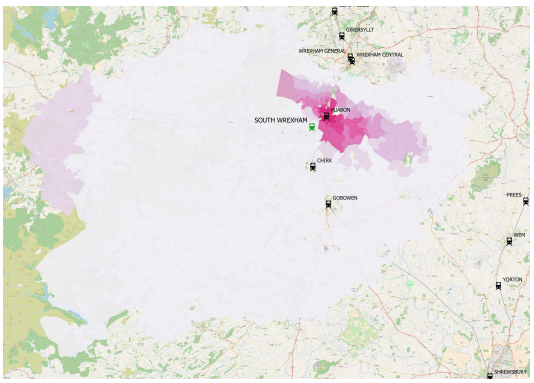


FIGURE 8.24: The probabilistic catchment for Ruabon station if South Wrexham station was opened.

[7]Although PDFH elasticities are intended to be applied to flows, the unitary elasticity assumes that only origin population is allowed to influence growth in rail demand. There is a lack of evidence on the appropriate elasticity to use if both origin and destination population changes are considered.

8.8 Impact of the accessibility term on demand forecasts and abstraction analysis

Two demand forecasts based on probabilistic station catchments have been reported for each of the case study stations, calculated using trip end models 8 and 9. The difference between these two models is the inclusion of the accessibility term in the station choice model component of model 9. As discussed in Section 6.7.2, the negative estimated parameter for the accessibility term is indicative of a competition effect, suggesting that the closer a station is on average to other, and larger, stations within a specific choice set, the less likely it is to be chosen. The trip end model results presented in Table 7.6 show that model 9 has the lower AIC value and the Akaike weight indicates a high probability that this is the better model. To assess the impact of including the accessibility term in the station choice model, the difference between the forecast entries/exits produced by the two models for the Borders Railway stations are summarised in Table 8.7, along with their performance against actual station usage in 2016/17. The differences between the forecasts produced by the two models are very small, with the largest adjustment made to Tweedbank station, with the number of forecast trips reduced by 4,238 (0.81%); and the smallest adjustment made to Gorebridge station with a increase of just 197 trips (0.09%). However, it is interesting to note that with the exception of Gorebridge, the number of trips has been adjusted in the required direction to produce a more accurate forecast, with an increase for Galashiels and a reduction for the other five stations.

Station	Model 8 - % difference from 16/17	Model 9 forecast less Model 8 forecast	% change in forecast from model 8	Model 9 - % difference from 16/17
Tweedbank	9.74	-4238	-0.81	8.84
Galashiels	-54.03	845	0.54	-53.78
Stow	17.94	-490	-0.63	17.20
Gorebridge	143.07	197	0.09	143.29
Newtongrange	53.01	-603	-0.29	52.57
Eskbank	9.98	-1885	-0.75	9.15
Shawfair	211.75	-488	-0.75	209.42

TABLE 8.7: Analysis of the effect of including the accessibility term in the station choice model on demand forecasts for the Borders Railway stations.

The issue of spatial correlation and the proportional substitution behaviour of the MNL model is of particular relevance to the abstraction analysis, as in the developed methodology an MNL model is specifically run before and after the proposed new station is added to the choice set of affected postcodes. To assess the extent to which the accessibility term can alter the proportional substitution effect, an analysis was carried out for two postcodes affected

by the proposed South Wrexham station. The results of the analysis are shown in Tables 8.8 and 8.9, and a map of the station and postcode locations is shown in Figure 8.25.

Station	Probability	Proportion	From Penyffordd	To S. Wrexham	Expected probability (if proportional)	Probability forecast by model	% change in probability
Hope	0.000	0.000	0.000	0.000	0.000	0.000	-92.20
Cefn-y-Bedd	0.000	0.000	0.000	0.000	0.000	0.000	-92.26
Caergwrle	0.000	0.000	0.000	0.000	0.000	0.000	-92.24
Gwersyllt	0.000	0.000	0.000	0.000	0.000	0.000	-92.29
Gobowen	0.005	0.005	0.000	0.005	0.000	0.000	-92.39
Wrexham Cent.	0.005	0.005	0.000	0.005	0.000	0.000	-92.31
Chirk	0.040	0.040	0.000	0.038	0.002	0.003	-92.42
Wrexham Gen.	0.111	0.111	0.000	0.106	0.005	0.009	-92.31
Ruabon	0.838	0.838	0.000	0.802	0.036	0.031	-96.29
Penyffordd (-)	0.000						
S. Wrexham (+)					0.956	0.956	
Totals	1.000	1.000	0.000	0.956	1.000	1.000	

TABLE 8.8: Analysis of the impact of the accessibility term on proportional substitution when South Wrexham station is added — choice set for postcode LL14 3BJ.

Station	Probability	Proportion	From Penyffordd	To S. Wrexham	Expected probability (if proportional)	Probability forecast by model	% change in probability
Hope	0.001	0.001	0.000	0.001	0.000	0.000	-70.61
Gwersyllt	0.001	0.001	0.000	0.001	0.000	0.000	-70.96
Cefn-y-Bedd	0.002	0.002	0.000	0.001	0.000	0.000	-70.85
Caergwrle	0.003	0.003	0.000	0.002	0.001	0.001	-70.78
Wrexham Cent.	0.013	0.013	0.000	0.011	0.003	0.004	-71.03
Gobowen	0.023	0.023	0.000	0.018	0.005	0.007	-71.34
Chirk	0.114	0.114	0.000	0.091	0.024	0.033	-71.47
Wrexham Gen.	0.287	0.287	0.000	0.228	0.059	0.083	-71.03
Ruabon	0.554	0.555	0.001	0.441	0.114	0.078	-86.02
Penyffordd (-)	0.002						
S. Wrexham (+)					0.794	0.794	
Totals	1.000	1.000	0.002	0.794	1.000	1.000	

TABLE 8.9: Analysis of the impact of the accessibility term on proportional substitution when South Wrexham station is added — choice set for postcode LL20 8AN.

In both examples, Penyffordd station was removed from the choice set of the nearest ten stations, and South Wrexham was added. Two adjustments are necessary to calculate the expected probabilities: the probability of Penyffordd has to be allocated to the remaining nine stations in proportion to their probabilities[8]; and the probability of South Wrexham has to be drawn from the nine stations, also in proportion to their probabilities. The tables show the expected probability of each station being chosen assuming proportional substitution,

[8]In these examples the probability of Penyffordd station being chosen was extremely low so its removal has negligible impact on the analysis.

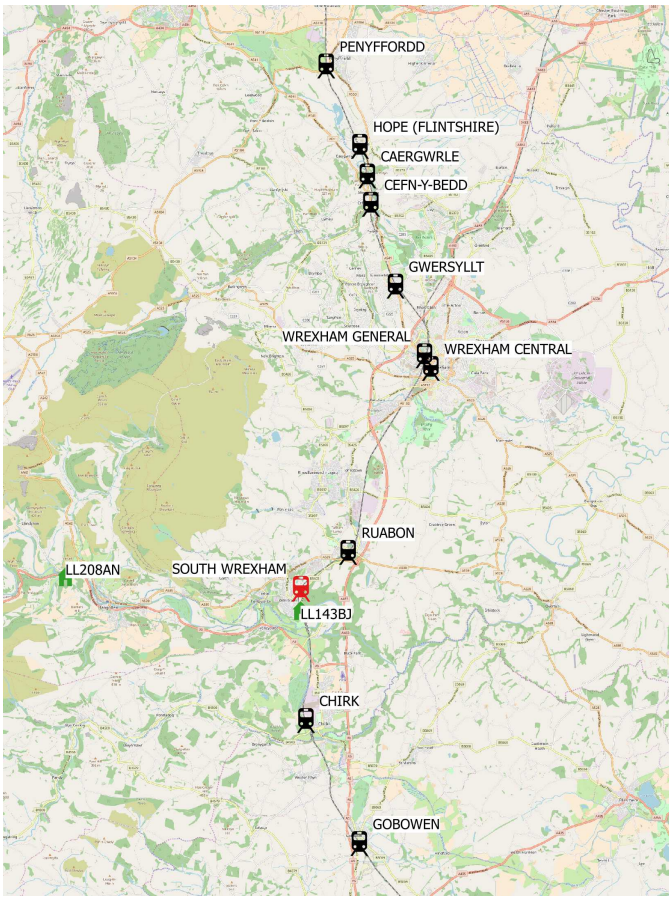


FIGURE 8.25: Map showing the location of postcodes and stations relevant to the analysis of the accessibility term's effect on proportional substitution behaviour.

alongside the probability forecast by the model (with accessibility term). In the case of both postcodes, the probability of Ruabon station being chosen is lower than that expected from a proportional substitution pattern. For LL208AN the chance of Ruabon station being chosen is reduced from 11.4% to 7.8%, and for LL143BJ it is reduced from 3.6% to 3.1%. In contrast the probabilities of the other stations being chosen are slightly higher than would be expected under proportional substitution, for example the chance of Wrexham General being chosen for LL208AN increases from 5.9% to 8.3%. It can be seen that the percentage reduction in probability caused by the introduction of South Wrexham is not the same for each station; it is noticeably higher for Ruabon, the closest station to South Wrexham, and then it gradually falls for the remaining stations as their distance from South Wrexham increases. This is the effect that would be intuitively expected, with a new station abstracting more passengers from closer stations than more distant stations. However, further work would be needed to assess whether this altered substitution pattern was a realistic representation of the abstraction behaviours resulting from competition between stations.

8.9 Real-world application as a forecasting tool for the Welsh Government

The methodologies described in this chapter, incorporating the station choice and trip end models that were calibrated as part of this research project, have been used to produce demand forecasts for 12 proposed new stations in Wales, and to assess the likely abstraction effects on five existing stations. This work was commissioned by the Welsh Government as part of the Welsh National Travel Plan, and forecasts were produced using both deterministic and probabilistic station catchments. Table 8.10 summarises the predicted annual entries/exits for each of the 12 stations produced using the two approaches, and shows the percentage difference between the two forecasts. For eight of the stations the demand forecast is higher when using probabilistic catchments, ranging from +0.7% to +35.7%, while for the remaining four stations the forecast is lower, ranging from −12.7% to −28.5%. The difference between the two forecasts is greater than ±20% for half of the stations. While the accuracy of these forecasts cannot yet be assessed, they do affirm the earlier case study findings that meaningful differences occur between forecasts produced using models with deterministic or probabilistic station catchments. These differences are large enough to potentially affect the viability of proposed new station schemes or to impact planned levels of infrastructure, services and facilities. The full report that was compiled for this consultancy work can be found in Appendix C.

	Predicted annual entries/exits 2015/16		
Potential station	Deterministic catchment	Probabilistic catchment	Difference (%)
Cockett	[redacted]	[redacted]	29.6
Ely Mill/Victoria park	[redacted]	[redacted]	-16.7
Llanwern	[redacted]	[redacted]	3.4
Newport Road/Rover Way	[redacted]	[redacted]	-28.5
Landore	[redacted]	[redacted]	22.8
St. Clears	[redacted]	[redacted]	0.7
Deeside Industrial Park	[redacted]	[redacted]	35.7
North Wrexham	[redacted]	[redacted]	14.4
South Wrexham	[redacted]	[redacted]	20.3
Llangefni	[redacted]	[redacted]	1.4
St. Mellons/Cardiff Parkway	[redacted]	[redacted]	-26.8
Carno	[redacted]	[redacted]	-12.7

TABLE 8.10: Summary of station demand forecasts for potential new station locations in Wales, showing the difference between forecasts produced using deterministic or probabilistic station catchments.

8.10 Conclusions

This chapter has described the development and application of a methodology to forecast demand for new stations using a trip end model that incorporates probabilistic station catchments based on station choice modelled at the unit postcode level. The performance of the trip end model was assessed for ten recently opened stations, and for all but three of the stations the model with probabilistic station catchments produced more accurate forecasts than the model with deterministic catchments. For two of the stations there was very little difference in the forecasts produced by the two types of model, while in the remaining case the deterministic catchment model produced a more accurate forecast.

For several of the case study stations, the probabilistic catchment model resulted in an adjustment to the demand forecast produced by the deterministic catchment model of over 30%. For example, the forecast for Tweedbank was reduced by 36% (some 330,000 trips) and the Shawfair forecast was reduced by 39% (some 48,000 trips). In both cases the revised forecast was more accurate. A difference of this magnitude could result in a change to the benefit-cost ratio of a proposed new station that alters the assessment made of its viability. This signifies the potential importance of using a trip end model based on probabilistic station catchments, that better represent the complexity of real-life station catchments. The assessment of scheme viability has the potential to be further enhanced by the proposed method for estimating abstraction from existing stations.

The trip end model with probabilistic station catchments has also performed well when compared to the official forecasts produced during scheme appraisal. This is particularly the case for the three Scottish Borders stations on the new Borders Railway line, where the final business case forecast produced by Transport Scotland massively under-predicted demand. However, for two stations on this line (Gorebridge and Shawfair) the model over-predicted demand by more than 100%, although this was considerably better than the forecast produced by the model with deterministic catchments, and under-forecast demand at Galashiels by almost 50%. In those cases where the model performed less well, there appear to be several contributing factors that highlight weaknesses in the trip end model that are not related to how the station catchment is defined. These include the inability to account for competition from other modes; not representing tourism as an attraction variable; and only allowing the 'quality' of access by one motorised mode (i.e. parking spaces for car users) to generate additional trips. Possible solutions to these issues have been outlined.

The potential of this practical and workable methodology, when combined with the underlying trip end and station choice models, to produce more robust station demand forecasts has already been recognised by transport practitioners at the national level, with work commissioned by the Welsh Government to assess 12 proposed station locations.

Chapter 9

Conclusions

9.1 Introduction

The overall aim of this research project was to determine whether the performance of the aggregate rail demand models commonly used in GB to forecast demand for new railway stations could be improved by defining probabilistic station catchments; and six key objectives that needed to be met in order to achieve this aim were set out in the introduction to this thesis. The research question arose from two connected concerns: that the catchments defined using existing methods were not adequately capturing the complexities of real-world station catchments; and that this might have contributed to some erroneous station demand forecasts over recent years. An alternative approach was suggested, where catchment zonal units would be assigned to several 'competing' stations, with the population of each zone allocated proportionately to each station based on probabilities derived using a station choice model. This chapter will identify how the research objectives were met and set out how the project has advanced knowledge in the area of rail demand forecasting. It will also consider the professional practice and policy implications of the research, discuss potential limitations, and make some proposals for future work.

9.2 What did we know before?

Trip rate or trip end models are the most common type of model used to forecast demand for new railway stations in GB, but these have not always perform well, with examples of substantial under- or over-prediction. The models are typically developed and applied on a local basis, reflecting guidance from the UK DfT (Department for Transport, 2011) and the rail industry (Association of Train Operating Companies, 2013), which both consider the appraisal of new stations to be a special case requiring bespoke models. There has been some previous research to develop nationally applicable trip end and flow models for local

stations (Blainey, 2010; Blainey & Preston, 2013a). However, in common with aggregate models generally, these rely on simplistic methods of defining station catchments that assume station choice is a deterministic process. There is an increasing body of evidence that real station catchments are far more complex entities, and failure to account for this may have contributed to the poor performance of the models used to appraise some recent schemes.

A review of prior station choice research found that MNL and NL have been the most commonly applied models, primarily used to explore station choice alone or combined access mode and station choice. As applied, these models have ignored the spatial nature of railway station choice, and their proportional substitution behaviour is problematic: a new station would be expected to abstract proportionately more passengers from existing stations that are closer to it. While recent research has developed a ML model to account for spatial correlation between station pairs (Weiss & Habib, 2016), this has not been tested in a demand forecasting scenario, and it remains unclear whether it can produce a realistic pattern of substitution. A failure to move beyond simply explaining station choice behaviour is a general criticism that can be made of most prior research. The only previous work to take a broadly similar approach to that proposed for this project was the unsuccessful attempt by Wardman and Whelan (1999) to calibrate a flow model that defined probabilistic catchments by apportioning the population of postal sectors to one of five competing stations.

9.3 Research summary — what do we know now?

9.3.1 Models of station choice

The first part of the research project, relating to objectives 1, 2 and 3, was concerned with the development of station choice models suitable for integration into either trip end or flow models of rail demand. This involved obtaining and preparing observed station choice data, generating the potential station choice predictor variables, and then model calibration and appraisal.

Data from on-train passengers surveys were obtained from the WG and Transport Scotland's LATIS service, and several novel techniques were developed to validate these datasets and maximise their usefulness. These included the estimation of trip origins from incomplete address information, and the automated identification of illogical trips. Analysis of the trip data showed that most postcodes were located within the observed catchments of multiple stations, and there was little evidence to support the notion of stations having discrete non-competing catchments. Objective 1, to obtain, process and validate suitable survey datasets able to reveal observed station choice behaviour, ideally covering more than one region of GB, was therefore achieved.

A range of potential station choice predictor variables were derived from open transport data sources, with a focus on ensuring a realistic representation of components of the access

journey and train leg that would have influenced passengers' choice of station on the day and at the time that they travelled. A processing framework based around OTP, R and PostgreSQL was implemented to manipulate the large amount of data in a reproducible manner, and an API wrapper was written to query OTP and parse the planner response. Objective 2, to derive candidate predictor variables for the station choice models, with a particular focus on maximising the potential of open transport data sources, was therefore achieved.

MNL models were calibrated separately for the WG and LATIS datasets, with the choice set for each observation defined as the ten nearest stations (plus the nearest major station, if not present). In-sample predictive performance of the best MNL models was substantially better than a comparator base model, where the nearest station was assumed to have a probability of one. There was also reasonably good concurrence in the parameter estimates for many of the predictor variables across the two datasets, indicating a degree of transferability. This was tested by applying the best models to the alternative dataset, and while the WG models performed rather better on the LATIS dataset than vice-versa, in all cases the predictive performance was superior to the base model. RPL models were also calibrated, and while there was some evidence of individual taste variation with respect to mode-specific access time, the marginal difference in predictive performance did not justify the extra complexity and computational time that would be involved in simulating station probabilities for every unit postcode in GB (a requirement for calibrating a national aggregate model).

A model intended for incorporation into a national trip end model was then calibrated using the combined dataset, thus maximising the information available to the model. An accessibility term was introduced to account for spatial correlation between stations, and a significant negative parameter was estimated, indicating the presence of a competition effect, with a station less likely to be chosen the closer it is on average to other (and more attractive) stations. Using a fixed attractiveness weighting in the accessibility term based on station category was found to be a suitable proxy for total entries/exits, enabling the term to be used when choice sets contain a proposed new station. The best performing model, and the one used in subsequent trip end model calibration, is shown in Equation 6.11. Objective 3, to calibrate station choice models appropriate for integrating into aggregate rail demand models, and assess their predictive performance and transferability, was therefore achieved.

$$V_{nik} = \exp(\beta N_k + \gamma \sqrt{D_{ik}} + \delta U_k + \epsilon \ln F_k + \zeta C_k + \eta Ps_k + \theta T_k + \iota B_k + \kappa \ln A_k). \quad \text{(6.11 revisited)}$$

9.3.2 A national trip end model

The second part of the research project, relating to objectives 4, 5 and 6, was concerned with the calibration, application and appraisal of national trip end models for GB.

A model form was proposed where a station's trips are generated by the population of each postcode which has that station in its choice set, with the generation potential dependent

upon the probability of the station being chosen and, by way of a two-stage decay function, the postcode's distance from the station. Choice sets were constructed for every postcode in mainland GB, consisting of the ten nearest stations, and the associated choice probabilities were calculated by applying the combined dataset station choice model. Trip end models were calibrated for Category E and F stations in mainland GB, using both probabilistic and deterministic catchment definitions. The models with probabilistic catchments were found to perform better, in terms of adjusted R^2 and AIC, than those with deterministic catchments. Importantly, greater weight was given to the population variable in the models with probabilistic catchments, while reduced weight was given to variables related to station services and characteristics. This indicates that the more realistic representation of the catchment in these models enables differences in the number of trips to be better explained through the population variable, and as a consequence they should be more transferable and better suited for use as a national predictive model. The population parameter in the model with *deterministic* catchments was substantially higher than that found in similar models calibrated by Blainey (2017), where the zonal unit was of lower spatial resolution (census output area). This suggests that the use of postcodes as the zonal unit has in itself been important in defining more realistic station catchments. The form of the trip end model with probabilistic station catchments is shown in Equation 7.2. Objective 4, to develop a methodology to incorporate probability-based station catchments into aggregate demand models and apply this methodology to calibrate a national-scale model for local railway stations in GB, was therefore achieved.

$$\ln \hat{V}_i = \alpha + \beta \left(\ln \sum_{z}^{Z} Pr_{zi} P_z w_{zi} \right) + \gamma \ln F_i + \delta \ln J_{it} + \epsilon \ln Ps_i + \zeta Te_i + \eta El_i + \theta B_i \quad \text{(7.2 revisited)}$$

A methodology was developed to apply the calibrated trip end models to forecast demand for new stations, and to estimate abstraction effects from existing stations; thus achieving objective 5. The models were then applied to several case studies, and their predictive performance was assessed for ten recently opened stations, including seven on a newly built railway line. For all but three stations, the model with probabilistic catchments produced a more accurate forecast than the model with deterministic catchments. For several of the stations, the probabilistic catchment model adjusted the demand forecast by more than 30% in the desired direction, highlighting the potential importance of using a trip end model that better represents real-life station catchments. The model also performed well when compared to the official forecasts produced during scheme appraisals, particularly for stations on the new Borders Railway line. A methodology developed to assess the extent that a new station might extract demand from existing stations was tested for a proposed new station in Wales, and an analysis of the impact of the accessibility term showed an appropriate adjustment to the MNL proportional substitution pattern, with the new station abstracting proportionately more demand from closer existing stations. Objective 6, to apply the demand forecasting methodology to several case studies, and carry out a performance appraisal, including an

assessment of models with either deterministic or probabilistic station catchments, was therefore achieved.

9.3.3 Summary of contribution to knowledge

This research project has made the following empirical or methodological contributions to knowledge in the field of rail demand forecasting and related fields:

- A national trip end model for new local railway stations that incorporates probabilistic station catchments derived from a station choice model applied at the unit postcode level, and which has superior predictive performance and transferability when compared to models based on simple deterministic station catchments. This is the first known example of successfully incorporating probabilistic station catchments into an aggregate rail demand model, and is an important advancement of the previous national models developed by Blainey (2010), which were based on deterministic station catchments.
- MNL station choice models suitable for integration into either trip end or flow models of rail demand where the choice decision is modelled at high spatial resolution (unit postcode level) and that can account for spatial correlation between stations through incorporation of an accessibility term based on Fotheringham's CDM. The CDM has not previously been applied in the context of station choice modelling, and in the combined trip end variant model revealed the presence of a competition effect.
- A methodology for applying the trip end model with probabilistic station catchments to forecast demand for new individual stations or new railway lines, which includes the assessment of abstraction from existing stations.
- Two novel methods to process and validate OD survey data. The first maximises the usability of OD survey data by estimating the coordinates of an origin or destination based on incomplete address information; and the second identifies the two most common errors in this type of data ('reversed trips' and 'substantial backtracks') by calculating ratios based on information inherent to the reported trip.
- A framework to automatically generate variables for transport-related models from open transport data using open source tools, supported by a set of functions to query the OTP routing API.

9.4 Practice and policy implications

Forecasting demand for new railway stations is considered by the rail industry to be a 'special case' requiring bespoke models to be developed and applied in a local context for the specific

scheme being appraised. As established in Chapter 2, this is primarily achieved through the use of trip rate/end models that have not always performed well. Given the background of growing passenger demand and increasing interest in opening new stations and lines, there will be an ongoing need to assess proposed schemes. The national trip end model that has been developed during this research project has the potential to remove the need for scheme proponents, such as local and regional government or transport authorities, to commission bespoke studies. In cases where it was still considered prudent to apply local models, the national model could be used as a sense-check tool. For example, if demand forecasts produced by the local and national models differed by orders of magnitude, it would be a clear warning that the local models may not be reliable. Given that the level of station usage is a key driver of the benefit-cost ratio upon which investment decisions are made, identifying a potential problem with the demand forecast at an early stage of a project would be hugely beneficial.

Ideally, advice contained in the rail industry's demand modelling 'bible', the PDFH, would be updated to highlight the approach adopted in this research project and the potential benefits of a national model. However, to derive maximum benefit from the work already completed, attention should be given to how the knowledge already gained can be transferred to industry practitioners. Access to the model could be provided on a consultancy basis, as has already happened in the case of work completed for the Welsh Government to assess 12 station locations as part of the Welsh National Transport Plan. An alternative and more sophisticated solution would be to incorporate the model and associated data into the new Data and Analytics Facility for National Infrastructure (DAFNI)[1]. A potential implementation would enable a DAFNI user to specify (or select on a map) potential new station locations, provide the variables required by the underlying models (for example, service frequency or number of car parking spaces) and then submit a batch job. Forecasts and visualisations, such as station probabilistic catchments, would then be prepared and the user notified upon completion. There would also be the potential to approach the problem from an alternative perspective, with the model asked to identify potential optimum locations for new stations within a particular area subject to specified criteria. This would be a charged-for service for non-academic users of DAFNI. Whatever mechanism was adopted, it would be necessary to regularly re-calibrate the station choice and trip end models to ensure their temporal transferability, for example by incorporating revised population data and taking into account new access and egress modes (such as on-demand ride-sharing services and autonomous vehicles).

[1] `http://www.itrc.org.uk/dafni-data-and-analytics-facility-for-national-infrastructure/`

9.5 Research limitations and potential solutions

9.5.1 Data-related issues

9.5.1.1 Station facility variables

There is some doubt about the accuracy of information contained in the NRE Knowledgebase station feed, which was the source of the facilities variables used in the station choice and trip end models. It was noted in Section 6.7.1 that the staffing level information was not reliable for stations in England, and the variable was changed in the combined station choice model. It also appears that the data on car parking, i.e. whether a car park is present and/or the number of parking spaces, may not be reliable. This became apparent when data was collated for the appraisal case studies, with Energlyn and Churchill Park station reported to have no car park, while a review of Google satellite and Street View imagery revealed an official station car park with approximately 18 spaces. In view of these findings, it is reasonable to assume that other information relating to station facilities within the NRE Knowledgebase is either incomplete or incorrect. It would not be practicable to manually verify this information for every station in Britain, and it would be preferable if a concerted effort was made by the rail industry to ensure that this information is both accurate and based on the application of consistent definitions (for example, provision to contact remotely located staff is not the same as a station having full-time staff). This would be of benefit to the rail industry generally as the knowledgebase is also used to provide customer-facing information via the NRE website. The impact of these data quality issues on the station choice and trip end models is difficult to assess, as the extent of the problem is unknown. However, assuming that the data is correct for most stations, it is likely that the models would have performed somewhat better had this data been more accurate. For example, with car parking spaces being an important driver of trip generation in the trip end model, the number of trip entries/exits will have been under-predicted for any station where a car park is present but not recorded as such.

9.5.1.2 OpenTripPlanner edge traversal issue

A problematic issue with using OTP, which was discussed in Section 5.4.1, occurs when the nearest edge to an origin does not have traversable permissions for motorised vehicles. This was resolved when deriving the access variables for the station choice models by manually adjusting the affected origins. However, this was not a feasible solution when the station access variables needed to be obtained for every postcode in mainland GB. To resolve this issue, ArcGIS was used with a street network created from the Ordnance Survey Open Roads dataset. This network was much less sophisticated than that generated by OTP using OSM data. For example, one-way roads, pedestrianised streets, and turn restrictions were not

represented. Consequently, the drive distances used to calibrate the station choice models will have been more realistic than those used when the models were applied, and in some instances this will have affected identification of the nearest station by distance, the relative distances to alternative stations, and ultimately the choice probabilities. An alternative solution would have been to use walk mode to generate the distances. While this would have generated more realistic distances for stations that are likely to be walked to, by allowing routing via pedestrian pathways and ignoring restrictions on motorised traffic, the distances would be less realistic for longer access journeys, for example by not allowing traversal of motorways. A better long-term solution would be to amend the OTP source code to allow the option of a walk component at the start or end of a car trip when it is not possible to reach the origin or destination by motorised vehicle, with the shortest path to/from the nearest edge traversable by motorised vehicles selected.

9.5.2 Station choice model limitations

9.5.2.1 Revealed preference surveys

The revealed preference surveys used to calibrate the station choice models were obtained from interviews with rail passengers in Wales and Scotland. While the findings suggest reasonable transferability of the models between these two regions, it has not been possible to rigorously assess how well they might predict station choice in England. Attempts were made at the beginning of the project to obtain survey data from train operating companies and passenger transport executives operating in England, but this was not successful. Using the station choice predictor tool described in Section 7.5.3.3, the performance of the station choice models was assessed for several locations in England based on local knowledge of the researcher, and the probabilities were considered reasonably realistic for the locations checked. However, it would be preferable if additional survey data for regions in England could be obtained and separate choice models calibrated and then compared with the Welsh and Scottish models.

A potential problem arising from survey respondents being asked at which station they boarded and would alight from the *current train*, rather than requesting their ultimate boarding and alighting station, was discussed in Section 6.4.2.1. To ensure that the ultimate origin and destination stations were correctly identified, any observations where the access or egress mode was recorded as 'another train' were excluded from the analysis. This should have limited the observations to direct journeys only, but this was not the case, suggesting that some passengers did not interpret the question as intended. While this enabled models to be calibrated that incorporated the number of transfers and waiting time, the estimated parameters need to be treated with some caution. This limitation does not affect the trip end model, as it only relates to flow variant station choice models. However, if future work was to seek to incorporate probabilistic catchments into flow models then ideally new station choice

models should be calibrated based on surveys that ensure ultimate origin and destination stations are captured for all indirect train journeys.

9.5.2.2 Access mode

The best performing station choice models calibrated on the WG and LATIS datasets were those with mode-specific access time parameters. However, as choice of access mode was not modelled, only a single parameter was estimated in the combined model used to define probabilistic catchments in the trip end model. This deficiency was offset to an extent by using a square root transformation of the access distance variable, thus imposing a proportionately higher disutility on the shorter access journeys that are most likely to be walked. Other issues remain, such as access by bus not being a realistic or even possible option for some or all stations in a choice set. However, given that access by bus usually only accounts for a very small proportion of access journeys (with the notable exception of London), this may not be a major cause for concern in most cases. The separate models also included a dummy variable for car as access mode, allowing the impact of certain factors, for example number of car parking spaces, to only be estimated against those observations where a car was actually used to access the station. While modelling access mode would allow these issues to be addressed, the role of car ownership/availability at the individual household level in determining whether car is a valid access mode choice presents a significant challenge, which may be more suited to an agent-based modelling approach. This would create a model of greater complexity that would be more difficult to implement and potentially less likely to be adopted by transport planners who currently rely on simple implementations of trip rate/end models to forecast demand for new stations.

9.5.2.3 Spatial correlation

While the inclusion of an accessibility term based on Fotheringham's CDM was successful in the combined station choice model, with a negative parameter indicating the presence of a competition effect, the adjustment made to proportional substitution behaviour in the case study stations was fairly subtle. It is possible that a combination of both agglomeration and competition effects is actually present within the data, and although the competition effect dominates it is correspondingly small. The accessibility term is also a measure of average proximity of a station to all the other stations in the choice set, when the spatial correlation between pairs of stations is likely to be more important in obtaining realistic abstraction effects. Furthermore, this approach precludes the addition of the nearest major station to each choice set, which would otherwise be desirable given the observed choice behaviour. Although several promising spatial models were identified, these could not be implemented as the model forms are not available in either proprietary or open source statistical software. The potential for future work in this area is considered in Section 9.6 below.

9.5.3 Trip end model limitations

The national trip end model was based on the log-log model calibrated by Blainey (2017), with the catchment definition component modified to incorporate probabilistic station catchments. This model was chosen as it is an established model that has been used to forecast demand for new stations as part of consultancy work for a number of clients, and therefore served as a robust comparator model. However, given the improved representation of station catchments in the new model, it is possible that the base model is no longer optimal and variables that were rejected when that model was calibrated may now be relevant to improving the model's predictive performance, and vice versa. In addition, the appraisal of the case study forecasts discussed in Chapter 8 identified several potential limitations of the trip end model: trip attraction resulting from tourism; additional trip generation due to an unusually high proportion of passengers accessing a station by bus; and the impact of competition from competing modes, in particular when a frequent, reliable and lower-priced bus service is available. Potential additional variables that could be investigated to address these issues include: a measure of tourist accommodation within a certain travel distance of a station; the frequency of bus services serving a station; and the difference in generalised journey time to the nearest major employment centre by rail compared to bus.

9.6 Programme of future work

It would be a natural extension of the research already completed to develop a national flow model based on probabilistic station catchments. In a flow model, rather than forecasting total trips at a station, the number of trips on each flow (OD station pair) is forecast, and previous work using deterministic catchments has shown that such models have the potential to more accurately forecast station demand (Blainey & Preston, 2010). Station choice models suitable for incorporation into a flow model were calibrated as part of this research project, but time constraints and difficulties obtaining suitable flow data prevented further progress being made. However, this thesis has shown that including elements of the train leg as predictor variables in the station choice models (for example, on-train time or number of transfers) can improve their predictive performance. In turn, this should enable more realistic stations catchments to be defined, which could ultimately result in a more robust and transferable flow model. To calibrate such a flow model, information on the number of trips on each flow would need to be obtained from the LENNON ticketing database, and while this has proved very difficult to obtain in the past, Transport Scotland has recently indicated their willingness to provide access to the Scottish data. The methods for incorporating station choice into the flow models, and for applying the calibrated model to generate demand forecasts, will be more complex than those already developed for the trip end models. For example, rather than a postcode having a single choice probability for each alternative station, it would have a separate probability for each flow for each station; and each station would have a potentially

different probabilistic catchment for each flow. In addition, there are known deficiencies with the LENNON data that may be problematic, such as missing trips in travelcard areas and difficulties accurately assigning trip direction.

The second potential area for future research, is to better address the proportional substitution behaviour that is a characteristic of the MNL models. Several promising spatial choice model forms have been identified and these were discussed in Section 3.3.3. However, the functionality to run them is not present in proprietary or open-source software packages and it is therefore necessary to define the likelihood function programmatically, using a matrix programming language such as GAUSS. A possible way forward would be to collaborate with researchers who have established expertise in this area, and one possibility is to work with academics based at the University of Toronto who have been exploring spatial choice models in the context of transit station choice (for example, see Weiss and Habib (2017)) and have already expressed an interest in carrying out collaborative research.

A final potential area of future work would be to develop a comprehensive API wrapper to query the OTP route planner, which could be released as an R package. This would build on the set of functions that were written as part of this research project, preventing duplicated work amongst the research community and making the functionality available to those who lack the necessary knowledge, skills or time to develop a solution themselves. This could be of benefit to researchers worldwide and across disciplines.

9.7 Concluding remarks

The evidence from the empirical models that have been developed, and their practical application to real-world case studies, supports the conclusion that the aggregate models used to forecast demand for new local railway stations can be improved, both in terms of their transferability and predictive performance, by incorporating probabilistic station catchments derived using station choice models. The trip end model that has been developed is the only known example of a national-scale aggregate rail demand model to incorporate probabilistic station catchments. It is also the first to define the catchment zonal unit at such a high spatial resolution and to be calibrated on a dataset of this size and geographic scope, in that it incorporates nearly every local station in England, Wales and Scotland. The model has already been applied commercially to assess proposed new station locations on behalf of a national government. This serves to highlight its potential role in providing decision makers with more accurate demand forecasts and guidance on expected abstraction effects, thus maximising the likelihood that new local railway stations will in future deliver the economic and societal benefits expected of them.

Appendix A

R code segments

A.1 OTP API wrapper

```
# This is a set of functions used to query the OTP API — the beginnings of a comprehensive API wrapper for OTP

# Load the required libraries
library(curl)
library(httr)
library(jsonlite)

# otp connect function

otpConnect <-
 function(hostname = 'localhost',
          router = 'default',
          port = '8080',
          ssl = 'false')
 {
  return (paste(
   ifelse(ssl == 'true', 'https://', 'http://'),
   hostname,
   ':',
   port,
   '/otp/routers/',
   router,
   sep = ""
  ))
 }

# Function to return distance for walk, cycle or car — desn't make sense for transit (bus or rail)
otpTripDistance <-
 function(otpcon,
          from,
          to,
          modes)
 {
  # convert modes string to uppercase — expected by OTP
```

```
35    modes <- toupper(modes)
36
37    # need to check modes are valid
38
39    # setup router URL with /plan
40    routerUrl <- paste(otpcon, '/plan', sep = "")
41
42    # Use GET from the httr package to make API call and place in req - returns json by default
43    req <- GET(routerUrl,
44               query = list(
45                 fromPlace = from,
46                 toPlace = to,
47                 mode = modes
48               ))
49    # convert response content into text
50    text <- content(req, as = "text", encoding = "UTF-8")
51    # parse text to json
52    asjson <- jsonlite::fromJSON(text)
53
54    # Check for errors - if no error object, continue to process content
55    if (is.null(asjson$error$id)) {
56      # set error.id to OK
57      error.id <- "OK"
58      if (modes == "CAR") {
59        # for car the distance is only recorded in the legs objects. Only one leg should be returned if mode is car and we pick that
          ↪ - probably need error check for this
60        response <-
61          list(
62            "errorId" = error.id,
63            "duration" = asjson$plan$itineraries$legs[[1]]$distance
64          )
65        return (response)
66        # for walk or cycle
67      } else {
68        response <-
69          list("errorId" = error.id,
70               "duration" = asjson$plan$itineraries$walkDistance)
71        return (response)
72      }
73    } else {
74      # there is an error - return the error code and message
75      response <-
76        list("errorId" = asjson$error$id,
77             "errorMessage" = asjson$error$msg)
78      return (response)
79    }
80  }
81
82  # Function to make an OTP API lookup and return trip time in simple or detailed form. The parameters from, to, modes, date
        ↪ and time must be specified in the function call other parameters have defaults set and are optional in the call.
83  otpTripTime <-
84    function(otpcon,
85             from,
86             to,
87             modes,
88             detail = FALSE,
89             date,
90             time,
91             maxWalkDistance = 800,
```

```
92          walkReluctance = 2,
93          arriveBy = 'false',
94          transferPenalty = 0,
95          minTransferTime = 0)
96  {
97    # convert modes string to uppercase — expected by OTP
98    modes <- toupper(modes)
99
100   routerUrl <- paste(otpcon, '/plan', sep = "")
101
102   # Use GET from the httr package to make API call and place in req — returns json by default. Not using numItineraries due
        ↪ to odd OTP behaviour — if request only 1 itinerary don't necessarily get the top/best itinerary, sometimes a
        ↪ suboptimal itinerary is returned. OTP will return default number of itineraries depending on mode. This function
        ↪ returns the first of those itineraries.
103   req <- GET(
104    routerUrl,
105    query = list(
106     fromPlace = from,
107     toPlace = to,
108     mode = modes,
109     date = date,
110     time = time,
111     maxWalkDistance = maxWalkDistance,
112     walkReluctance = walkReluctance,
113     arriveBy = arriveBy,
114     transferPenalty = transferPenalty,
115     minTransferTime = minTransferTime
116    )
117   )
118
119   # convert response content into text
120   text <- content(req, as = "text", encoding = "UTF-8")
121   # parse text to json
122   asjson <- jsonlite::fromJSON(text)
123
124   # Check for errors — if no error object, continue to process content
125   if (is.null(asjson$error$id)) {
126    # set error.id to OK
127    error.id <- "OK"
128    # get first itinerary
129    df <- asjson$plan$itineraries[1,]
130    # check if need to return detailed response
131    if (detail == TRUE) {
132     # need to convert times from epoch format
133     df$start <-
134      as.POSIXct(df$startTime / 1000, origin = "1970-01-01")
135     df$end <-
136      as.POSIXct(df$endTime / 1000, origin = "1970-01-01")
137     # create new columns for nicely formatted dates and times
138     #df$startDate <- format(start.time, "%d—%m—%Y")
139     #df$startTime <- format(start.time, "%I:%M%p")
140     #df$endDate <- format(end.time, "%d—%m—%Y")
141     #df$endTime <- format(end.time, "%I:%M%p")
142     # subset the dataframe ready to return
143     ret.df <-
144      subset(
145       df,
146       select = c(
147        'start',
```

```
148           'end',
149           'duration',
150           'walkTime',
151           'transitTime',
152           'waitingTime',
153           'transfers'
154         )
155       )
156       # convert seconds into minutes where applicable
157       ret.df[, 3:6] <- round(ret.df[, 3:6] / 60, digits = 2)
158       # rename walkTime column as appropriate - this a mistake in OTP
159       if (modes == "CAR") {
160         names(ret.df)[names(ret.df) == 'walkTime'] <- 'driveTime'
161       } else if (modes == "BICYCLE") {
162         names(ret.df)[names(ret.df) == 'walkTime'] <- 'cycleTime'
163       }
164       response <-
165         list("errorId" = error.id, "itineraries" = ret.df)
166       return (response)
167     } else {
168       # detail not needed - just return travel time in seconds
169       response <-
170         list("errorId" = error.id, "duration" = df$duration)
171       return (response)
172     }
173   } else {
174     # there is an error - return the error code and message
175     response <-
176       list("errorId" = asjson$error$id,
177            "errorMessage" = asjson$error$msg)
178     return (response)
179   }
180 }
181
182 # function to return isochrone (only works correctly for walk and/or transit modes - limitation of OTP)
183 otpIsochrone <-
184   function(otpcon,
185            from,
186            modes,
187            cutoff,
188            walkspeed,
189            batch)
190   {
191     # convert modes string to uppercase - expected by OTP
192     modes <- toupper(modes)
193
194     routerUrl <- paste(otpcon, '/isochrone', sep = "")
195     # need to check modes are valid
196     # Use GET from the httr package to make API call and place in req - returns json by default
197     req <- GET(
198       routerUrl,
199       query = list(
200         fromPlace = from,
201         mode = modes,
202         cutoffSec = cutoff,
203         walkSpeed = walkspeed,
204         batch = batch
205       )
206     )
```

```
207     # convert response content into text
208     text <- content(req, as = "text", encoding = "UTF-8")
209
210     # Check that geojson is returned
211
212     if (grepl("\"type\":\"FeatureCollection\"", text)) {
213       status <- "OK"
214     } else {
215       status <- "ERROR"
216     }
217     response <-
218       list("status" = status,
219            "response" = text)
220     return (response)
221   }
```

A.2 Parse NRE Knowledgebase XML feed

```
1   # This script parses the ATOC Stations XML Feed to extract information on station services and facilities and then updates a
        ↪ PostgreSQL table.
2
3   # Load the required libraries
4   library(curl)
5   library(httr)
6   library(XML)
7   library(RPostgreSQL)
8
9   # initialize errors dataframe
10  errors <- data.frame(crsCode = character(),
11                       statusCode = character(),
12                       bodyEmpty = logical())
13
14  # define namespaces vector
15  ns <-
16    c(x = "http://nationalrail.co.uk/xml/station", y = "http://nationalrail.co.uk/xml/common", z =
        ↪ "http://www.govtalk.gov.uk/people/AddressAndPersonalDetails")
17
18  # get CRS codes from the full NRE stations xml list. I initially used CRS codes from Naptan, but Naptan does not have all
        ↪ stations. Having downloaded the full list it would have been better to parse that. But the code below requests the feed
        ↪ for each station individually.
19
20  stations_xml <-
21    xmlParse("C:/PhD/Analysis/r/stations_xml_feed/stations.xml")
22
23  allcrs <-
24    xmlToDataFrame(getNodeSet(xml_doc, "//x:Station/x:CrsCode", namespaces = ns),
25                   stringsAsFactors = FALSE)
26
27  # start loop
28  for (i in 1:nrow(allcrs)) {
29    # set crs code
30    # sleep so don't bombard the NRE server
31    Sys.sleep(2)
```

```
32  crsCode <- allcrs$text[i]
33
34  # set up the feed URL for current crs code
35  feedUrl <-
36   paste("http://internal.nationalrail.co.uk/xml/30/station-",
37         crsCode,
38         ".xml",
39         sep =
40           "")
41
42  # API calls are made via a cloud server acting as proxy. This is because access to the feed requires registration of a static IP
        ↪ address.
43
44  # timeout set (in seconds) to prevent R hanging if no response
45  req <- GET(feedUrl, use_proxy("95.85.54.43", 3128), timeout(10))
46
47  # check that staus_code is 200 and content body is not empty
48  if (req$status_code == 200 && paste(req[6]) != "raw(0)") {
49   # parse the response
50   xml <- xmlParse(req)
51
52   # initialise list
53   services <- vector("list", 21)
54   # set list names
55   names(services) <-
56    c(
57      "crscode",
58      "name",
59      "longitude",
60      "latitude",
61      "staffingLevel",
62      "cctv",
63      "ticketMachine",
64      "waitingRoom",
65      "stationBuffet",
66      "toilets",
67      "cycleStorage",
68      "cycleSpaces",
69      "cycleShelter",
70      "cycleCctv",
71      "freeCarPark",
72      "carSpaces",
73      "taxiRank",
74      "busServices",
75      "metroServices",
76      "carHire",
77      "cycleHire"
78    )
79
80   services$crscode <- crsCode
81
82   # get staffingLevel — mandatory fullTime, partTime or unstaffed
83   services$name <-
84    xpathSApply(xml,
85                "/x:Station/x:Name",
86                namespaces = ns,
87                fun = xmlValue)
88
89   # get staffingLevel — mandatory fullTime, partTime or unstaffed
```

```
90  services$longitude <-
91   xpathSApply(xml,
92               "/x:Station/x:Longitude",
93               namespaces = ns,
94               fun = xmlValue)
95
96  # get staffingLevel - mandatory fullTime, partTime or unstaffed
97  services$latitude <-
98   xpathSApply(xml,
99               "/x:Station/x:Latitude",
100              namespaces = ns,
101              fun = xmlValue)
102
103 # get staffingLevel - mandatory fullTime, partTime or unstaffed
104 services$staffingLevel <-
105  xpathSApply(
106    xml,
107    "/x:Station/x:Staffing/x:StaffingLevel",
108    namespaces = ns,
109    fun = xmlValue
110  )
111
112 # get CCTV status - mandatory TRUE or FALSE
113 services$cctv <-
114  xpathSApply(
115    xml,
116    "/x:Station/x:Staffing/x:ClosedCircuitTelevision/x:Overall",
117    namespaces = ns,
118    fun = xmlValue
119  )
120
121 # get ticketMachine status - optional true/false - if tag missing assume false as per schema
122 xpath <- "/x:Station/x:Fares/x:TicketMachine/x:Available"
123 if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
124  services$ticketMachine <-
125    xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
126 } else {
127  services$ticketMachine <- "false"
128 }
129
130 # get waitingRoom status - optional can have either available or open tags
131 # define xpaths
132 xpath.available <-
133  "/x:Station/x:StationFacilities/x:WaitingRoom/y:Available"
134 xpath.open <-
135  "/x:Station/x:StationFacilities/x:WaitingRoom/y:Open"
136
137 # Check for available tag first - if present get value
138 if (length(xpathSApply(xml, xpath.available, namespaces = ns)) > 0) {
139  services$waitingRoom <-
140    xpathSApply(xml,
141                xpath.available,
142                namespaces = ns,
143                fun = xmlValue)
144  # then check for open tag -  if tag present assume there is a waiting room
145 } else if (length(xpathSApply(xml, xpath.open, namespaces = ns)) > 0) {
146  services$waitingRoom <- "true"
147 } else {
148  services$waitingRoom <- "NA"
```

```
149   }
150
151   # get stationBuffet status − optional true/false/unknown
152   xpath <−
153    "/x:Station/x:StationFacilities/x:StationBuffet/y:Available"
154   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
155    services$stationBuffet <−
156     xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
157   } else {
158    services$stationBuffet <− "NA"
159   }
160
161   # get toilets status − optional true/false/unknown
162   xpath <− "/x:Station/x:StationFacilities/x:Toilets/x:Available"
163   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
164    services$toilets <−
165     xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
166   } else {
167    services$toilets <− "NA"
168   }
169
170   # get cycle storage availability − manadatory True/False (InterChange is not manadatory)
171
172   xpath <− "/x:Station/x:Interchange/x:CycleStorageAvailability"
173   # Check if tag exists
174   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
175    # get tag value
176    services$cycleStorage <−
177     xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
178   } else {
179    services$cycleStorage <− "NA"
180   }
181
182   # get cycle storage spaces − optional Integer (number)
183
184   xpath <− "/x:Station/x:Interchange/x:CycleStorageSpaces"
185   # Check if tag exists
186   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
187    # get tag value
188    services$cycleSpaces <−
189     xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
190   } else {
191    services$cycleSpaces <− 0
192   }
193
194   # get cycle storage sheltered − optional yes/partial/no/unknown
195
196   xpath <− "/x:Station/x:Interchange/x:CycleStorageSheltered"
197   # Check if tag exists
198   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
199    # get tag value
200    services$cycleShelter <−
201     xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
202   } else {
203    services$cycleShelter <− "NA"
204   }
205
206   # get cycle CCTV − optional True/False
207
```

```
208   xpath <- "/x:Station/x:Interchange/x:CycleStorageCctv"
209   # Check if tag exists
210   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
211    # get tag value
212    services$cycleCctv <-
213     xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
214   } else {
215    services$cycleCctv <- "NA"
216   }
217
218   # Is a free car park available - Self-closing tag (no content). Existence indicates that there is no charge for using this car
        ↪ park at any time.
219   xpath <- "/x:Station/x:Interchange/x:CarPark/x:Charges/x:Free"
220   # Check at least one Spaces tag exists
221   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
222    # there is a free car park
223    services$freeCarPark <- "true"
224   } else {
225    services$freeCarPark <- "false"
226   }
227
228   # get total car parking spaces - optional
229   xpath <- "/x:Station/x:Interchange/x:CarPark/x:Spaces"
230   # Check at least one Spaces tag exists
231   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
232    #Use Reduce to sum the spaces for all carparks tags
233    services$carSpaces <-
234     Reduce(sum, (as.numeric(
235      xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
236     )))
237   } else {
238    services$carSpaces <- 0
239   }
240
241
242   # get Taxi Rank - optional true/false/unknown
243   xpath <- "/x:Station/x:Interchange/x:TaxiRank/y:Available"
244   # Check if tag exists
245   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
246    # get tag value
247    services$taxiRank <-
248     xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
249   } else {
250    services$taxiRank <- "NA"
251   }
252
253   # get Bus Services - optional true/false/unknown
254   xpath <- "/x:Station/x:Interchange/x:BusServices/y:Available"
255   # Check if tag exists
256   if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
257    # get tag value
258    services$busServices <-
259     xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
260   } else {
261    services$busServices <- "NA"
262   }
263
264   # get Metro Services - optional true/false/unknown
265   xpath <- "/x:Station/x:Interchange/x:MetroServices/y:Available"
```

```
    # Check if tag exists
    if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
      # get tag value
      services$metroServices <-
        xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
    } else {
      services$metroServices <- "NA"
    }

    # get Car Hire - optional true/false/unknown
    xpath <- "/x:Station/x:Interchange/x:CarHire/y:Available"
    # Check if tag exists
    if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
      # get tag value
      services$carHire <-
        xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
    } else {
      services$carHire <- "NA"
    }

    # get Cycle Hire - optional true/false/unknown
    xpath <- "/x:Station/x:Interchange/x:CycleHire/y:Available"
    # Check if tag exists
    if (length(xpathSApply(xml, xpath, namespaces = ns)) > 0) {
      # get tag value
      services$cycleHire <-
        xpathSApply(xml, xpath, namespaces = ns, fun = xmlValue)
    } else {
      services$cycleHire <- "NA"
    }

    # Write results to database table
    dbWriteTable(
      conn = con,
      name = c('data', 'stations'),
      data.frame(services),
      append = TRUE,
      row.names = FALSE
    )

  } else {
    # response code was not 200, write details to errors dataframe
    newRow <-
      data.frame(
        crsCode = crsCode,
        statusCode = req$status_code,
        bodyEmpty = isTRUE(paste(req[6]) == "raw(0)")
      )
    errors <- rbind(errors, newRow)
  }
  # end the loop
}
```

A.3 Querying brfares.com API to obtain fares

Example for LATIS dataset.

```
# load libraries
library(httr)
library(jsonlite)
library(stringr)

# Create dataframe of unique origin:destination stations identified by CRS code
fares_lookup <-
  unique(latis_alternatives_main[c("alternative", "destcrs")])

# initialize errors dataframe
errors <- data.frame(
  id = integer(),
  correctOrig = character(),
  correctDest = character(),
  correctRlc = character(),
  noFares = character()
)

total <- nrow(fares_lookup)

for (i in 1:total) {
  # set values
  originCrs <- fares_lookup$alternative[i] # is the alternative
  destCrs <- fares_lookup$destcrs[i]

  # set up the feed URL
  feedUrl <-
    paste("http://api.brfares.com/queryextra?orig=",
          originCrs,
          "&dest=",
          destCrs,
          sep =
            "")

  # make API call, using gzip encoding
  req <- GET(feedUrl, config(accept_encoding = "gzip"))

  # convert response content into text
  text <- content(req, as = "text")

  # convert from JSON to list of R objects
  asRlist <- fromJSON(text)

  # Check that the api call response is valid
  if (asRlist$correct$orig == TRUE &&
      asRlist$correct$dest == TRUE &&
      asRlist$correct$rlc == TRUE &&
      !is.null(asRlist$fares$adult$fare)) {
    # Extract the values of interest
    fareCategory <- asRlist$fares$category$desc
    routeCode <- asRlist$fares$route$code
    routeName <- asRlist$fares$route$name
    ticketCode <- asRlist$fares$ticket$code
```

```
54  ticketName <- asRlist$fares$ticket$name
55  restrictionCode <- asRlist$fares$restriction$code
56  adultFare <- asRlist$fares$adult$fare
57
58  # Create ticket dataframe
59  dfTickets <-
60    data.frame(
61      fareCategory,
62      routeCode,
63      routeName,
64      ticketCode,
65      ticketName,
66      restrictionCode,
67      "adultFare" = adultFare /
68        100,
69      stringsAsFactors = FALSE
70    )
71
72  # Subset the ticket dataframe to WALKUP fares only
73  dfTickets <- dfTickets[dfTickets$fareCategory == "WALKUP",]
74
75  # Subset to off-peak returns
76  dfOffPeak <-
77    dfTickets[dfTickets$ticketCode == "CDR" |
78                dfTickets$ticketCode == "SVR" |
79                dfTickets$ticketCode == "BFR" |
80                dfTickets$ticketCode == "G2R" |
81                dfTickets$ticketCode == "SMG",]
82
83  # Subset to anytime returns
84  dfAnytime <-
85    dfTickets[dfTickets$ticketCode == "SDR" |
86                dfTickets$ticketCode == "SOR" |
87                dfTickets$ticketCode == "GOR" |
88                dfTickets$ticketCode == "GTR",]
89
90  # Extract fares - need to take minimum as alternative routes may be possible
91
92  # get off-peak fare - use CDR if available, otherwise ...
93  if (nrow(dfOffPeak) > 0) {
94    if ("CDR" %in% dfOffPeak$ticketCode) {
95      idx <- which.min(dfOffPeak$adultFare[dfOffPeak$ticketCode == "CDR"])
96      offpeakReturn <- dfOffPeak$adultFare[idx]
97      offpeakRestriction <- dfOffPeak$restrictionCode[idx]
98    } else {
99      idx <- which.min(dfOffPeak$adultFare)
100     offpeakReturn <- dfOffPeak$adultFare[idx]
101     offpeakRestriction <- dfOffPeak$restrictionCode[idx]
102   }
103 } else {
104   offpeakReturn <- "NULL"
105   offpeakRestriction <- "NULL"
106 }
107
108 # get anytime fare - use SDR if available, otherwise ...
109 if (nrow(dfAnytime) > 0) {
110   if ("SDR" %in% dfAnytime$ticketCode) {
111     idx <-
112       which.min(dfAnytime$adultFare[dfAnytime$ticketCode == "SDR"])
```

```
      anytimeReturn <- dfAnytime$adultFare[idx]
      anytimeRestriction <- dfAnytime$restrictionCode[idx]
    } else {
      idx <- which.min(dfAnytime$adultFare)
      anytimeReturn <- dfAnytime$adultFare[idx]
      anytimeRestriction <- dfAnytime$restrictionCode[idx]
    }
  } else {
    anytimeReturn <- "NULL"
    anytimeRestriction <- "NULL"

  }

  # now update the fares_lookup dataframe
  fares_lookup[i, "offpeakfare"] <- offpeakReturn
  fares_lookup[i, "offpeakrestriction"] <-
    str_trim(offpeakRestriction)
  fares_lookup[i, "anytimefare"] <- anytimeReturn
  fares_lookup[i, "anytimerestriction"] <-
    str_trim(anytimeRestriction)

 } else {
  # there is a problem with the api call response - record errors in errors dataframe for later review
  newRow <-
    data.frame(
      id = i,
      correctOrig = asRlist$correct$orig,
      correctDest = asRlist$correct$dest,
      correctRlc = asRlist$correct$rlc,
      nofares = paste(is.null(asRlist$fares$adult$fare))
    )
  errors <- rbind(errors, newRow)
 }
}
```

A.4 Retrieving address matches from AddressBase

```
# Do addressbase search for each unique posttown, excluding where posttown is "unknown"

for (posttown in sort(unique(add2015chk$Origin.posttown[add2015chk$Origin.posttown != "unknown"])))

{
 # set up temp table and index
 query <-
   paste(
     "create temp table tmp as (select \"POSTCODE\", postcode_count, max_d_2ct, ST_X
       ↪ (ST_Transform (stpc_cent_geom, 4326))as stpc_cent_X, ST_Y (ST_Transform
       ↪ (stpc_cent_geom, 4326))as stpc_cent_Y, full_text_address, address_short from
       ↪ addressbase where \"POST_TOWN\" = '",
     posttown,
     "')"
     ,
     sep = ""
```

```
14    )
15   dbGetQuery(con, query)
16
17   # create gin index on full_text_address field
18   #
19   query <-
20    paste(
21     "CREATE INDEX idx_tmp_trgm ON tmp USING gin (full_text_address COLLATE
       ↪ pg_catalog.\"default\" gin_trgm_ops)"
22     ,
23     sep = ""
24    )
25   dbGetQuery(con, query)
26
27   # create gin index on short address field
28   query <-
29    paste(
30     "CREATE INDEX idx_tmp_trgm2 ON tmp USING gin (address_short COLLATE
       ↪ pg_catalog.\"default\" gin_trgm_ops)"
31     ,
32     sep = ""
33    )
34   dbGetQuery(con, query)
35
36   # run search
37   for (id in add2015chk$ID[add2015chk$Origin.posttown == posttown]) {
38    query <-
39     paste(
40      "SELECT \"POSTCODE\", postcode_count, max_d_2ct, stpc_cent_X, stpc_cent_Y,
       ↪ full_text_address as address, similarity(full_text_address, '"
41      ,
42      add2015chk$Origin_full_address[add2015chk$ID == id]
43      ,
44      "') FROM tmp WHERE full_text_address % '"
45      ,
46      add2015chk$Origin_full_address[add2015chk$ID == id]
47      ,
48      "' UNION SELECT \"POSTCODE\", postcode_count, max_d_2ct, stpc_cent_X, stpc_cent_Y,
       ↪ address_short as address, similarity(address_short, '"
49      ,
50      add2015chk$Origin_full_address[add2015chk$ID == id]
51      ,
52      "') FROM tmp WHERE address_short % '"
53      ,
54      add2015chk$Origin_full_address[add2015chk$ID == id]
55      ,
56      "' ORDER BY similarity DESC LIMIT 4"
57      ,
58      sep = ""
59     )
60    result <- dbGetQuery(con, query)
61
62    # Save results
63
64    # check first that we have some results - check nrows not null
65    if (nrow(result) > 0) {
66     # loop through the results
67     for (r in 1:nrow(result)) {
68      # set variables
```

```
      sim <- paste('M', r, '.s', sep = "")
      addr <- paste('M', r, '.add', sep = "")
      pc <- paste('M', r, '.pc', sep = "")
      pcnt <- paste('M', r, '.pcnt', sep = "")
      maxd2ct <- paste('M', r, '.maxd2ct', sep = "")
      stpcc <- paste('M', r, '.stpcc', sep = "")
      add2015chk[[sim]][add2015chk$ID == id] <-
       round(result[r, "similarity"], 2)
      add2015chk[[addr]][add2015chk$ID == id] <-
       result[r, "address"]
      add2015chk[[pc]][add2015chk$ID == id] <-
       result[r, "POSTCODE"]
      add2015chk[[pcnt]][add2015chk$ID == id] <-
       result[r, "postcode_count"]
      add2015chk[[maxd2ct]][add2015chk$ID == id] <-
       result[r, "max_d_2ct"]
      add2015chk[[stpcc]][add2015chk$ID == id] <-
       paste(round(result[r, "stpc_cent_y"], 5), ",", round(result[r, "stpc_cent_x"], 5), sep =
            "")
    }
   }
  }

  # drop temp table (and index?)

  query <-
   paste ("drop table tmp")
  dbGetQuery(con, query)

}
```

A.5 Creating observed station catchments

This example is for the LATIS dataset.

```
# get distinct list of origin stations in the dataset
query1 <-
 paste("SELECT DISTINCT origincrs FROM latis.survey_val", sep = "")
query1 <- gsub(pattern = '\\s',
           replacement = "  ",
           x = query1)
df <- dbGetQuery(con, query1)

# loop through each origin station and create a temporary table to hold distinct origin postcodes for that station and then
    ↪ create a polygon linking the postcode centroids and write to database. Polygon created using ST_ConcaveHull function
    ↪ set at 0.99 target percent

# Note: need at least 3 records for each station to build a catchment

for (i in 1:nrow(df)) {
 query2 <-
  paste(
```

```
    "CREATE TEMP TABLE catchment AS
       SELECT DISTINCT ON (originlatlong) originlatlong, origincrs, id, origin_geom
       FROM latis.survey_val
       where origincrs = '",
    df[i, 1],
    "' ORDER BY originlatlong",
    sep =
     ""
   )
  query2 <- gsub(pattern = '\\s',
           replacement = " ",
           x = query2)
  dbGetQuery(con, query2)
  # check there are at least 3 points in the temp table - pgr_pointsASPolygon needs at least 3
  count_rows <- dbGetQuery(con, "select count() from catchment")
  if (count_rows > 2) {
   query3 <-
    paste(
     "INSERT INTO latis.catchment_allorigins_polygons (origin_geom, origincrs)
          VALUES (
          (select ST_ConcaveHull(ST_Collect(origin_geom), 0.99)
          from catchment),
          (SELECT origincrs from catchment LIMIT 1)
          )"
)
   query3 <- gsub(pattern = '\\s',
            replacement = " ",
            x = query3)
   dbGetQuery(con, query3)
  }
  # drop the temp table
  dbGetQuery(con, "DROP TABLE catchment")
}
```

A.6 Station catchments that each unit-level postcode intersects

```
# Step 1: get the set of postcode polygons that intersect the station catchments produced above and create a table. Use distinct
    ↪ otherwise will get multiple postcodes because of intesection with different catchments

query <-
 paste(
  "create table latis.pc_in_obs_catchments as (
    select distinct(a.postcode), a.geom
    from data.postcode_polygons as a, latis.catchment_allorigins_polygons as b
    where ST_Intersects(a.geom, b.origin_geom_gb) and a.geom is not null
  )",
 sep = ""
 )
query <- gsub(pattern = '\\s',
         replacement = " ",
         x = query)
dbGetQuery(con, query)

```

```
# Step 2:  Create a station catchment count field for the postcodes in latis.pc_in_obs_catchments

query <-
 paste("ALTER TABLE latis.pc_in_obs_catchments ADD COLUMN in_catchments integer",
     sep = "")
query <- gsub(pattern = '\\s',
          replacement = " ",
          x = query)
dbGetQuery(con, query)

# Step 3: generate the catchment count

query <-
 paste(
  "with tmp2 as (
    with tmp as(
    SELECT a.postcode, count(b.origincrs)
    FROM latis.pc_in_obs_catchments as a
    LEFT JOIN latis.catchment_allorigins_polygons as b
    ON ST_Intersects(a.geom,b.origin_geom_gb)
    group by b.origincrs, a.postcode
    )
    select postcode, count() from tmp
    group by postcode)
    Update latis.pc_in_obs_catchments as c
    set in_catchments = tmp2.count
    from tmp2
    where c.postcode = tmp2.postcode",
  sep = ""
  )
query <- gsub(pattern = '\\s',
          replacement = " ",
          x = query)
dbGetQuery(con, query)
```

Appendix B

PostgreSQL code segments

B.1 AddressBase

B.1.1 Generate `postcode_count` field

```
1  ALTER TABLE addressbase
2   ADD COLUMN postcode_count SMALLINT;
3
4  WITH tmp2 AS (
5   WITH tmp AS (
6      SELECT
7       DISTINCT ON ("POST_TOWN", "DEPENDENT_THOROUGHFARE", "THOROUGHFARE",
         ↪ "DOUBLE_DEPENDENT_LOCALITY", "DEPENDENT_LOCALITY", "POSTCODE")
8       "POST_TOWN",
9       "DEPENDENT_THOROUGHFARE",
10      "THOROUGHFARE",
11      "DOUBLE_DEPENDENT_LOCALITY",
12      "DEPENDENT_LOCALITY",
13      "POSTCODE"
14     FROM data.addressbase
15     WHERE "THOROUGHFARE" <> '')
16   SELECT
17    "POST_TOWN",
18    "DEPENDENT_THOROUGHFARE",
19    "THOROUGHFARE",
20    "DOUBLE_DEPENDENT_LOCALITY",
21    "DEPENDENT_LOCALITY",
22    count()
23   FROM tmp
24   GROUP BY "POST_TOWN", "DEPENDENT_THOROUGHFARE", "THOROUGHFARE", "DOUBLE_DEPENDENT_LOCALITY",
       ↪ "DEPENDENT_LOCALITY")
25 UPDATE data.addressbase
26 SET postcode_count = tmp2.count
27 FROM tmp2
28 WHERE data.addressbase."POST_TOWN" = tmp2."POST_TOWN"
29     AND data.addressbase."DEPENDENT_THOROUGHFARE" = tmp2."DEPENDENT_THOROUGHFARE"
```

```
    AND data.addressbase."THOROUGHFARE" = tmp2."THOROUGHFARE"
    AND data.addressbase."DOUBLE_DEPENDENT_LOCALITY" = tmp2."DOUBLE_DEPENDENT_LOCALITY"
    AND data.addressbase."DEPENDENT_LOCALITY" = tmp2."DEPENDENT_LOCALITY";
```

B.1.2 Generate `stpc_cent_geom` field

```
ALTER TABLE addressbase
 ADD COLUMN stpc_cent_geom GEOMETRY(Point, 27700);

WITH tmp2 AS (
 WITH tmp AS (
    SELECT
     DISTINCT ON ("POST_TOWN", "DEPENDENT_THOROUGHFARE", "THOROUGHFARE",
      ↪ "DOUBLE_DEPENDENT_LOCALITY", "DEPENDENT_LOCALITY", "POSTCODE")
     "POST_TOWN",
     "DEPENDENT_THOROUGHFARE",
     "THOROUGHFARE",
     "DOUBLE_DEPENDENT_LOCALITY",
     "DEPENDENT_LOCALITY",
     "POSTCODE",
     b.the_geom
    FROM data.addressbase AS a
     LEFT JOIN data.onspd_nov_2015 AS b
      ON a."POSTCODE" = b.pcds
    WHERE "THOROUGHFARE" <> '')
 SELECT
  "POST_TOWN",
  "DEPENDENT_THOROUGHFARE",
  "THOROUGHFARE",
  "DOUBLE_DEPENDENT_LOCALITY",
  "DEPENDENT_LOCALITY",
  st_centroid(st_collect(the_geom)) AS geom
 FROM tmp
 GROUP BY "POST_TOWN", "DEPENDENT_THOROUGHFARE", "THOROUGHFARE", "DOUBLE_DEPENDENT_LOCALITY",
     ↪ "DEPENDENT_LOCALITY")
UPDATE data.addressbase
SET stpc_cent_geom = tmp2.geom
FROM tmp2
WHERE data.addressbase."POST_TOWN" = tmp2."POST_TOWN"
    AND data.addressbase."DEPENDENT_THOROUGHFARE" = tmp2."DEPENDENT_THOROUGHFARE"
    AND data.addressbase."THOROUGHFARE" = tmp2."THOROUGHFARE"
    AND data.addressbase."DOUBLE_DEPENDENT_LOCALITY" = tmp2."DOUBLE_DEPENDENT_LOCALITY"
    AND data.addressbase."DEPENDENT_LOCALITY" = tmp2."DEPENDENT_LOCALITY";
```

B.1.3 Generate `max_d_2ct` field

```
ALTER TABLE data.addressbase
 ADD COLUMN max_d_2ct INTEGER;

WITH tmp2 AS (
```

```
 WITH tmp AS (
   SELECT
    DISTINCT ON ("POST_TOWN", "DEPENDENT_THOROUGHFARE", "THOROUGHFARE",
     ↪ "DOUBLE_DEPENDENT_LOCALITY", "DEPENDENT_LOCALITY", "POSTCODE")
    "POST_TOWN",
    "DEPENDENT_THOROUGHFARE",
    "THOROUGHFARE",
    "DOUBLE_DEPENDENT_LOCALITY",
    "DEPENDENT_LOCALITY",
    "POSTCODE",
    stpc_cent_geom,
    b.the_geom
   FROM data.addressbase AS a
    LEFT JOIN data.onspd_nov_2015 AS b
     ON a."POSTCODE" = b.pcds
   WHERE "THOROUGHFARE" <> '')
 SELECT
  "POST_TOWN",
  "DEPENDENT_THOROUGHFARE",
  "THOROUGHFARE",
  "DOUBLE_DEPENDENT_LOCALITY",
  "DEPENDENT_LOCALITY",
  round(ST_MaxDistance(st_collect(the_geom), st_collect(stpc_cent_geom))) AS dist
 FROM tmp
 GROUP BY "POST_TOWN", "DEPENDENT_THOROUGHFARE", "THOROUGHFARE", "DOUBLE_DEPENDENT_LOCALITY",
     ↪ "DEPENDENT_LOCALITY")
UPDATE data.addressbase
SET max_d_2ct = tmp2.dist
FROM tmp2
WHERE data.addressbase."POST_TOWN" = tmp2."POST_TOWN"
   AND data.addressbase."DEPENDENT_THOROUGHFARE" = tmp2."DEPENDENT_THOROUGHFARE"
   AND data.addressbase."THOROUGHFARE" = tmp2."THOROUGHFARE"
   AND data.addressbase."DOUBLE_DEPENDENT_LOCALITY" = tmp2."DOUBLE_DEPENDENT_LOCALITY"
   AND data.addressbase."DEPENDENT_LOCALITY" = tmp2."DEPENDENT_LOCALITY";
```

B.2 Station daily train frequency

This query calculates train frequency for a particular day — in this example, 25 November 2013. Based on information provided in Zervaas (2014).

```

WITH tmp AS (
  SELECT
   t.,
   st.
  FROM gtfs2013.stop_times AS st, gtfs2013.trips AS t
  WHERE st.stop_id IN (SELECT stop_id
                FROM gtfs2013.stops
                WHERE parent_station = '",
i,
"') AND st.trip_id = t.trip_id
      AND t.service_id IN (SELECT service_id
```

```
                    FROM gtfs2013.calendar
                    WHERE start_date <= '2013-11-25' AND end_date >= '2013-11-25'
                        AND monday = 1)
)
SELECT count()
FROM tmp
```

B.3 Procedural code block to identify station pairs

```
DO
$do$
DECLARE pc CHARACTER VARYING;
BEGIN
 -- loop through each postcode in the probability table
 FOR pc IN SELECT DISTINCT postcode
        FROM demandmodels.pc_probs_n10_cmb
 LOOP
  -- we will insert the possible station pairs for this postcode into a table called station_pairs
  INSERT INTO demandmodels.station_pairs (j, i)
   WITH a AS (
     SELECT i
     -- unnest the array (tmp). This creates a CTE table (i) of one column containing the crs codes for this postcode
     FROM unnest(array(
            -- Use CTE to create table tmp which is an array of the station CRS codes for this postcode
              WITH tmp AS (
                SELECT array_agg(crscode)
                OVER (
                 PARTITION BY postcode )
                FROM demandmodels.pc_probs_n10_cmb
                WHERE postcode = pc)
              SELECT DISTINCT array_agg
              FROM tmp)) AS s(i)
   )
   -- select unique stations pairs for this postcode from CTE table (i) by using a cross join
   SELECT
    a.i AS j,
    b.i AS i
   FROM
    a
    CROSS JOIN a AS b
   WHERE
    a < b
   ORDER BY a, b;
 END LOOP;
END
$do$;

-- create a new table just containing the distinct station pairs
CREATE TABLE demandmodels.unique_stn_pairs AS
 SELECT DISTINCT
  j,
  i
 FROM demandmodels.station_pairs
```

B.4 Procedural code block to calculate accessibility term

```
DO
$do$
DECLARE
 pc      CHARACTER VARYING;
 r       INTEGER := 1;
 this_alt CHARACTER VARYING;
BEGIN
 -- loop through each record in probability table
 FOR r IN SELECT id
        FROM demandmodels.pc_probs_n10_cmb
        ORDER BY id
 LOOP
  -- populate variables related to this record
  SELECT INTO pc, this_alt
   postcode,
   crscode
  FROM demandmodels.pc_probs_n10_cmb
  WHERE id = r;
  -- use Common Table Expression to select the other stations for this pc along with relevant category weightings and station
     ↪ pair distances
  WITH cdm AS (
    SELECT
     a.id,
     a.crscode,
     a.category,
     b.fxd_entsexits,
     c.distance
    FROM demandmodels.pc_probs_n10_cmb AS a
     LEFT JOIN demandmodels.cat_weights AS b
      ON a.category = b.category
     LEFT JOIN demandmodels.station_pair_distance AS c
      ON (a.crscode = c.i AND this_alt = c.j) OR (a.crscode = c.j AND this_alt = c.i)
    WHERE postcode = pc AND crscode <> this_alt)
  -- calculate the accessibility term using select query on the CTE table and update the table
  UPDATE demandmodels.pc_probs_n10_cmb
  SET fxdwact = (SELECT round(cast(avg(fxd_entsexits / distance) AS NUMERIC), 4)
            FROM cdm)
  WHERE id = r;
 END LOOP;
END
$do$;
```

B.5 Station catchment queries

B.5.1 Simple catchment

In this example, the simple unweighted catchment population for Honiton station (CRS code is 'HON') is retrieved from the probability table.

```
SELECT sum(b.population)
FROM demandmodels.pc_nearest_15_stations AS a LEFT JOIN data.pc_pop_2011_clean AS b ON a.postcode = b.postcode
WHERE crscode = 'HON' AND distance_rank = 1
"
```

B.5.2 Simple weighted catchment

In this example, the simple catchment population weighted by the decay function for Honiton station (CRS code is 'HON') is retrieved from the probability table.

```
WITH nw_pop AS (
  SELECT
    -- first part of query does not apply decay function for postcodes within 750m of the station
    sum(population)
  FROM demandmodels.pc_nearest_15_stations AS A
    LEFT JOIN data.pc_pop_2011_clean AS b ON a.postcode = b.postcode
  WHERE crscode = 'HON'
      -- we only include those postcodes where 'HON' is the nearest station
      AND distance_rank = 1 AND total_dist / 1000 <= 0.75
), w_pop AS (
  SELECT
    -- second part of query applies the decay function for postcodes > 750m from the station
    sum(population  power(((total_dist / 1000) + 1), -1.5212))
  FROM demandmodels.pc_nearest_15_stations AS a
    LEFT JOIN data.pc_pop_2011_clean AS b ON a.postcode = b.postcode
  WHERE crscode = 'HON'
      -- we only include those postcodes where 'HON' is the nearest station
      AND distance_rank = 1 AND total_dist / 1000 > 0.75
)
-- use COALESCE function to set population sum to zero if query result was null
SELECT round(COALESCE(nw_pop.sum, 0) + COALESCE(w_pop.sum, 0)) AS w_pop
FROM nw_pop, w_pop
```

B.5.3 Probabilistic catchment

In this example, the probabilistic catchment population (with two-stage decay function applied) for Honiton station (CRS code is 'HON') is retrieved from the probability table.

```
WITH nw_pop AS (
  SELECT
    -- first part of query weights population only by probability where distance from postcode to station is within 750m
    sum(c.te19_prob  b.population)
  FROM demandmodels.pc_nearest_15_stations AS a
```

```
6       LEFT JOIN data.pc_pop_2011_clean AS b ON a.postcode = b.postcode
7       LEFT JOIN demandmodels.pc_probs_n10_cmb AS c ON a.postcode = c.postcode AND a.crscode = c.crscode
8     WHERE distance_rank < 11 AND a.crscode = 'HON' AND total_dist / 1000 <= 0.75
9   ), w_pop AS (
10    SELECT
11      -- second part of query weights population by probability and the decay function where distance from postcode to station
          ↪ is > 750m
12      sum(c.te19_prob  b.population  power(((total_dist / 1000) + 1), -1.5212))
13    FROM demandmodels.pc_nearest_15_stations AS a
14      LEFT JOIN data.pc_pop_2011_clean AS b ON a.postcode = b.postcode
15      LEFT JOIN demandmodels.pc_probs_n10_cmb AS c ON a.postcode = c.postcode AND a.crscode = c.crscode
16    WHERE distance_rank < 11 AND a.crscode = 'HON' AND total_dist / 1000 > 0.75
17  )
18  SELECT round(COALESCE(nw_pop.sum, 0) + COALESCE(w_pop.sum, 0)) AS w_pop
19  FROM nw_pop, w_pop
```

Appendix C

Station demand forecasts for Wales: report to the Welsh Government

This report (pp. 292–314) has been removed from the public version of this thesis due to client confidentiality.

Appendix D

Miscellaneous

D.1 Trip-end models

D.1.1 Travelcard boundary stations

Category E and F stations identified as travelcard boundary stations are shown in Table D.1.

D.1.2 Assigned categories

Categories that were assigned to stations (opened prior to January 2012) with no official category designation are shown in Table D.2.

D.1.3 Station ticketing groups

The fares feed dated 10 January 2017 was downloaded from `http://data.atoc.org/data-download`. Information about station groups is contained in the file `RJFAF359.LOC` (or similarly named file). The station groups are located at the top of this file, prior to other groups, for example bus groups. Group entries begin RG. The first 7 digits after RG is the group code. The rest of file can then be searched to find the stations that are part of this group. These are the RM entries. So, for example, the Bicester NTH/VIL group entries is: `RG7079340311229992807201528072015BICESTER NTH/VIL`; which has the group ID: `7079340`. The group members are: `RM7079340311229997030480BCS` and `RM7079340311229997031040BIT` (CRS codes BCS and BIT). More information can be found in the 'RJIS Datafeeds Interface Specification for Fares and Associated Data' PDF document provided with the feed download. The station groups for non-London stations are shown in Table D.3, and for London stations in Table D.4.

CRS code	Station name	Category	Travelcard Region
APB	Appley Bridge	F1	Greater Manchester
BLK	Blackrod	F1	Greater Manchester
BMC	Bromley Cross	E	Greater Manchester
BML	Bramhall	E	Greater Manchester
BYN	Bryn	F1	Greater Manchester
GLZ	Glazebrook	E	Greater Manchester
GNF	Greenfield	E	Greater Manchester
HAL	Hale	E	Greater Manchester
HDG	Heald Green	E	Greater Manchester
LTL	Littleborough	F2	Greater Manchester
MDL	Middlewood	F2	Greater Manchester
ORR	Orrell	F1	Greater Manchester
PAT	Patricroft	F2	Greater Manchester
SRN	Strines	F2	Greater Manchester
CWH	Crews Hill	F2	London
ELS	Elstree & Borehamwood	E	London
ENL	Enfield Lock	E	London
EWE	Ewell East	E	London
HDW	Hadley Wood	E	London
HTE	Hatch End	E	London
KCK	Knockholt	E	London
SGR	Slade Green	E	London
TUR	Turkey Street	E	London
WDT	West Drayton	E	London
WRU	West Ruislip	F1	London
ELP	Ellesmere Port	E	Merseyrail
GSW	Garswood	E	Merseyrail
HGN	Hough Green	E	Merseyrail
HSW	Heswall	F2	Merseyrail
MEC	Meols Cop	F2	Merseyrail
NLW	Newton-Le-Willows	E	Merseyrail
RNF	Rainford	F2	Merseyrail
DRT	Darton	F1	South & West Yorkshire
MRP	Moorthorpe	F1	South & West Yorkshire
SES	South Elmsall	F1	South & West Yorkshire
DBD	Denby Dale	F1	South Yorkshire
DOR	Dore & Totley	F2	South Yorkshire
KVP	Kiveton Park	F2	South Yorkshire
TNN	Thorne North	E	South Yorkshire
TNS	Thorne South	F2	South Yorkshire
BPT	Bishopton	E	Strathclyde
BRR	Barrhead	E	Strathclyde
CAC	Caldercruix	F2	Strathclyde
CRF	Carfin	F	Strathclyde
CRO	Croy	E	Strathclyde
CUB	Cumbernauld	E	Strathclyde
DLR	Dalreoch	E	Strathclyde
HLY	Holytown	F	Strathclyde
MIN	Milliken Park	F	Strathclyde
BLO	Blaydon	F2	Tyne & Wear
BKT	Blake Street	E	West Midlands
BWN	Bloxwich North	F2	West Midlands
DDG	Dorridge	E	West Midlands
EWD	Earlswood	F2	West Midlands
LOB	Longbridge	E	West Midlands
HBD	Hebden Bridge	E	West Yorkshire
HRS	Horsforth	F1	West Yorkshire
KNO	Knottingley	F1	West Yorkshire
MIK	Micklefield	F1	West Yorkshire
MSN	Marsden	F1	West Yorkshire
SON	Steeton & Silsden	F1	West Yorkshire
WDN	Walsden	F2	West Yorkshire

TABLE D.1: Category E and F stations identified as travelcard boundary stations, by travelcard region.

D.1.4 Stations excluded from unit postcode choice sets

The stations that were excluded when the choice sets were defined for every unit postcode in mainland GB are shown in Table D.5.

CRS code	Station name	Staffing level	Entries/exits 2015/16 (million)	Assumed category	Comment
ALO	Alloa	unstaffed	0.4	F	
AMR	Amersham	fullTime	2.3	C	
BSV	Buckshaw Parkway	partTime	0.3	D	
ZCW	Canada Water	unstaffed	23.7	B	Assume staffed (LO)
CFO	Chalfont & Latimer	fullTime	0.8	C	
CLW	Chorleywood	fullTime	0.5	D	
DLJ	Dalston Junction	unstaffed	5.1	C	Assume staffed (LO)
DUN	Dunbar	partTime	0.5	D	
GFD	Greenford	fullTime	0.3	D	
HGG	Haggerston	unstaffed	3.2	C	Assume staffed (LO)
HOH	Harrow-On-The-Hill	fullTime	2.4	C	
HAF	Heathrow Terminal 4	fullTime	n/a	n/a	
HWV	Heathrow Terminal 5	fullTime	n/a	n/a	
HXX	Heathrow Terminals 1-3	fullTime	n/a	n/a	
HOX	Hoxton	unstaffed	3	C	Assume staffed (LO)
MCE	Metrocentre	unstaffed	0.4	F	
OKE	Okehampton	unstaffed	0.003	F	
RIL	Rice Lane	fullTime	0.3	D	
RIC	Rickmansworth	fullTime	1.1	C	
ROE	Rotherhithe	unstaffed	1.7	C	Assume staffed (LO)
SMC	Sampford Courtenay	unstaffed	0	F	
SDE	Shadwell	unstaffed	5	C	Assume staffed (LO)
SDC	Shoreditch High Street	unstaffed	8	C	Assume staffed (LO)
SIA	Southend Airport	fullTime	0.4	D	
SFA	Stratford International	fullTime	1.6	C	
SQE	Surrey Quays	unstaffed	4.2	C	Assume staffed (LO)
TNA	Thornton Abbey	unstaffed	0.001	F	
WPE	Wapping	unstaffed	2.5	C	Assume staffed (LO)
ZLW	Whitechapel	unstaffed	14	B	Assume staffed (LO)

Notes: London Overground (LO)

TABLE D.2: Categories that were assigned to stations (opened prior to 1 January 2012) with no official category designation, based on staffing level and stations entries/exits.

StationGroupName	GroupMembers	CRScode
RG700254031122999100520120405201 2COLCHESTERSTNS	RM700254031122999706853 0CET	CET
	RM700254031122999706861 0COL	COL
RG700258031122999301020143010201 4CATFORDSTATIONS	RM700258031122999705047 0CFB	CFB
	RM700258031122999705077 0CTF	CTF
RG700259031122999301020143010201 4EDENBRIDGESTNS	RM700259031122999705359 0EBT	EBT
	RM700259031122999705473 0EBR	EBR
RG700260031122999131120141111201 4FARNBOROUGHSTNS	RM700260031122999705521 0FNB	FNB
	RM700260031122999705688 0FNN	FNN
RG700262031122999301020143010201 4PENGESTATIONS	RM700262031122999705072 0PNE	PNE
	RM700262031122999705378 0PNW	PNW
RG700263031122999120720121905201 2ENFIELDCHSE/TWN	RM700263031122999706010 0ENC	ENC
	RM700263031122999706959 0ENF	ENF
RG700265031122999270220132702201 3WHAMPSTEADSTNS	RM700265031122999701421 0WHD	WHD
	RM700265031122999701525 0WHP	WHP
RG700268031122999270220132702201 3PONTEFRACTSTNS	RM700268031122999708540 0PFR	PFR
	RM700268031122999708548 0PFM	PFM
RG700271031122999270220132702201 3THORNESTATIONS	RM700271031122999706530 0TNN	TNN
	RM700271031122999706531 0TNS	TNS
RG700403031122999270220132702201 3READINGSTATIONS	RM700403031122999703149 0RDG	RDG
	RM700403031122999703160 0RDW	RDW
RG700404031122999270220132702201 3HELENSBURGHSTNS	RM700404031122999709981 0HLC	HLC
	RM700404031122999709982 0HLU	HLU
RG700410031122999100520120305201 2BEDFORDSTATIONS	RM700410031122999701510 0BSJ	BSJ
	RM700410031122999701512 0BDM	BDM
RG700411031122999100520120405201 2SOUTHENDSTNS	RM700411031122999707420 0SOV	SOV
	RM700411031122999707456 0SOC	SOC
RG700413031122999270220132702201 3HERTFORDSTNS	RM700413031122999706085 0HFN	HFN
	RM700413031122999706818 0HFE	HFE
RG700415031122999190320131903201 3GAINSBOROUGH	RM700415031122999706424 0GBL	GBL
	RM700415031122999706465 0GNB	GNB
RG700416031122999131120141111201 4DORKINGSTATIONS	RM700416031122999705297 0DKT	DKT
	RM700416031122999705357 0DKG	DKG
	RM700416031122999705412 0DPD	DPD
RG700418031122999190320131903201 3BIRMINGHAMSTNS	RM700418031122999701006 0BSW	BSW
	RM700418031122999701127 0BHM	BHM
	RM700418031122999704515 0BMO	BMO
RG700424031122999270220132702201 3BRADFORDYKSTNS	RM700424031122999708345 0BDI	BDI
	RM700424031122999708346 0BDQ	BDQ
RG700428031122999120920131109201 3CANTERBURYSTNS	RM700428031122999705007 0CBW	CBW
	RM700428031122999705164 0CBE	CBE
RG700429031122999131120141111201 4DORCHESTERSTNS	RM700429031122999705961 0DCH	DCH
	RM700429031122999705962 0DCW	DCW
RG700431031122999131119971311199 7FALKIRKSTATIONS	RM700431031122999709930 0FKG	FKG
	RM700431031122999709931 0FKK	FKK
RG700432031122999120920131109201 3FOLKESTONESTNS	RM700432031122999705027 0FKW	FKW
	RM700432031122999705035 0FKC	FKC
RG700433031122999060819980608199 8GLASGOWCEN/QST	RM700433031122999709813 0GLC	GLC
	RM700433031122999709950 0GLQ	GLQ
RG700435031122999270220132702201 3LIVERPOOLSTNS	RM700435031122999702226 0MRF	MRF
	RM700435031122999702242 0LVC	LVC
	RM700435031122999702244 0LVJ	LVJ
	RM700435031122999702246 0LIV	LIV
RG700437031122999120920131109201 3MAIDSTONESTNS	RM700437031122999705115 0MDE	MDE
	RM700437031122999705222 0MDW	MDW
	RM700437031122999705237 0MDB	MDB
RG700438031122999270220132702201 3MANCHESTERSTNS	RM700438031122999702963 0DGT	DGT
	RM700438031122999702966 0MCO	MCO
	RM700438031122999702968 0MAN	MAN
	RM700438031122999702970 0MCV	MCV
RG700440031122999131120141111201 4PORTSMOUTHSTNS	RM700440031122999705537 0PMS	PMS
	RM700440031122999705540 0PMH	PMH
RG700441031122999190320131903201 3NEWARKSTATIONS	RM700441031122999706498 0NCT	NCT
	RM700441031122999706499 0NNG	NNG
RG700443031122999131119971311199 7TYNDRUMSTATIONS	RM700443031122999708728 0TYL	TYL
	RM700443031122999708838 0UTY	UTY
RG700444031122999190320131903201 3WAKEFIELDSTNS	RM700444031122999708584 0WKK	WKK
	RM700444031122999708591 0WKF	WKF
RG700445031122999270220132702201 3WARRINGTONSTNS	RM700445031122999702384 0WBQ	WBQ
	RM700445031122999702390 0WAC	WAC
RG700446031122999270220132702201 3WIGANSTATIONS	RM700446031122999702363 0WGN	WGN
	RM700446031122999702406 0WGW	WGW
RG700447031122999190320131903201 3WORCESTERSTNS	RM700447031122999704891 0WOS	WOS
	RM700447031122999704893 0WOF	WOF
RG700449031122999301020143010201 4CROYDONSTATIONS	RM700449031122999705355 0ECR	ECR
	RM700449031122999705411 0WCY	WCY
RG701780031122999290919990909199 9BOOTLESTATIONS	RM701780031122999702195 0BNW	BNW
	RM701780031122999702239 0BOT	BOT
RG707468031122999030220002701200 0TILBURYSTATIONS	RM707468031122999707461 0TBR	TBR
	RM707468031122999707462 0TIL	TIL
RG707934031122999280720152807201 5BICESTERNTH/VIL[1]	RM707934031122999703048 0BCS	BCS
	RM707934031122999703104 0BIT	BIT

Notes: [1]BICESTERNTH/VIL is the most recently created group, dating from 28 July 2015.

TABLE D.3: Station groups and group stations (not London).

StationGroupName	GroupMembers	CRScode
RG70107203112299901042000010420000LONDONTERMINALS	RM70107203112299970144400EUS	EUS
	RM70107203112299970147500MYB	MYB
	RM70107203112299970155500STP	STP
	RM70107203112299970308700PAD	PAD
	RM70107203112299970511200BFR	BFR
	RM70107203112299970514300CHX	CHX
	RM70107203112299970514800LBG	LBG
	RM70107203112299970542600VIC	VIC
	RM70107203112299970559800WAT	WAT
	RM70107203112299970612100KGX	KGX
	RM70107203112299970696500LST	LST
	RM70107203112299970749000FST	FST
RG70445203112299901011991181120160LONDONTHAMESLNK	RM70445203112299970057700ZFD	ZFD
	RM70445203112299970155500STP	STP
	RM70445203112299970511200BFR	BFR
	RM70445203112299970512100CTK	CTK
	RM70445203112299970514800LBG	LBG
	RM70445203112299970524600EPH	EPH

TABLE D.4: Station groups and group stations (London).

Station		
Name	Crscode	Reason for exclusion
Altnabreac	ABC	No access by road (forestry tracks only)
Bordesley	BBS	No weekday service
Brading	BDN	Isle of Wight
Brigg	BGG	No weekday service
Barlaston	BRT	No weekday service
Buckenham (Norfolk)	BUC	No weekday service
Berney Arms	BYA	Access via long countryside walk
Corrour	CRR	No access by road (forestry tracks only)
Dunrobin Castle	DNO	No weekday service
Denton	DTN	No weekday service
Falls Of Cruachan	FOC	No weekday service
Gainsborough Central	GNB	No weekday service
Heathrow Express	HAF	Serves airport only[1]
Heysham Port	HHB	No weekday service
Heathrow Express	HWV	Serves airport only[1]
Heathrow Express	HXX	Serves airport only[1]
Kirton Lindsey	KTL	No weekday service
Lakenheath	LAK	No weekday service
Lake	LKE	Isle of Wight
Lympstone Commando	LYC	No public access
Manchester United Football Ground	MUF	No weekday service
Norton Bridge	NTB	No weekday service
Okehampton	OKE	No weekday service
Pilning	PIL	No weekday service
Redcar British Steel	RBS	No public access
Reddish South	RDS	No weekday service
Ryde Esplanade	RYD	Isle of Wight
Ryde Pier Head	RYP	Isle of Wight
Ryde St Johns Road	RYR	Isle of Wight
Smallbrook Junction	SAB	Isle of Wight
Sandown	SAN	Isle of Wight
Shanklin	SHN	Isle of Wight
Sampford Courtenay	SMC	No weekday service
Stanlow & Thornton	SNT	No public access
Tees-Side Airport	TEA	No weekday service
Wedgwood	WED	No weekday service

Note: [1]These stations have an atypical 'catchment', and are not a viable origin station choice in most circumstances.

TABLE D.5: Stations excluded from unit postcode choice sets.

References

Adcock, S. J. (1997, September). *A passenger station choice model for the British rail network.* Paper presented at AET European Transport Conference.

Alderson, J., & McDonald, I. (Eds.). (2017). *Britain's growing railway* (6th ed.). Railway Development Society.

Association of Train Operating Companies. (2013). *Passenger demand forecasting handbook v5.1*.

Association of Train Operating Companies. (2015). *Passenger Demand Forecasting Council.* Webpage. Retrieved 7 April 2018, from https://www.raildeliverygroup.com/pdfc.html

Atkins Limited. (2011). *PLANET long distance framework: Approach to station choice modelling - a report for HS2.*

Basar, G., & Bhat, C. (2004). A parameterized consideration set model for airport choice: an application to the San Francisco Bay area. *Transportation Research Part B: Methodological*, *38*(10), 889–904.

Basemap Ltd. (2014). *Visography TRACC help pages.* PDF Document.

BBC. (2012, September 19). *Conon Bridge railway station to reopen in 2013.* Webpage. Retrieved 01/11/2017, from http://www.bbc.co.uk/news/uk-scotland-highlands-islands-19654403

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand (Transportation Studies Book 9)*. The MIT Press.

Ben-Akiva, M. E. (1973). *Structure of passenger travel demand models.* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Bernardin, V. L., Jr., Koppelman, F., & Boyce, D. (2009). Enhanced destination choice models incorporating agglomeration related to trip chaining while controlling for spatial competition. *Transportation Research Record*, *2132*(1), 143-151.

Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2010). Analysis of implicit choice set generation using a constrained multinomial logit model. *Transportation Research Record: Journal of the Transportation Research Board*, *2175*(1), 92–97.

Blainey, S. (2009). *Forecasting the use of new local railway stations and services using GIS* (Unpublished doctoral dissertation). University of Southampton.

Blainey, S. (2010). Trip end models of local rail demand in England and Wales. *Journal of Transport Geography*, *18*(1), 153–165.

Blainey, S. (2017). *A new station demand forecasting model for Wales.* (Unpublished working paper)

Blainey, S., & Evens, S. (2011, October). *Local station catchments: reconciling theory with reality.* Paper presented at AET European Transport Conference.

Blainey, S., & Preston, J. (2010). Modelling local rail demand in South Wales. *Transportation Planning and Technology*, *33*(1), 55–73.

Blainey, S., & Preston, J. (2013a). Extending geographically weighted regression from points to flows: a rail-based case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, *227*(6), 724–734.

Blainey, S., & Preston, J. (2013b). A GIS-based appraisal framework for new local railway stations and services. *Transport Policy*, *25*, 41–51.

Boyce, D. E., & Williams, H. (2016). *Forecasting Urban Travel: Past, Present and Future*. Edward Elgar Publishing Ltd.

BR Fares Ltd. (2016). *BR Fares.* Webpage. Retrieved 27-04-2017, from http://www.brfares.com

Brons, M., Givoni, M., & Rietveld, P. (2009). Access to railway stations and its potential in increasing rail use. *Transportation Research Part A: Policy and Practice*, *43*(2), 136–149.

Brons, M., & Rietveld, P. (2009). Improving the quality of the door-to-door rail journey: a customer-oriented approach. *Built Environment*, *35*(1), 122–135.

Campaign for Better Transport. (2017). *Re-opening rail lines.* Webpage. Retrieved 5 September 2017, from http://www.bettertransport.org.uk/re-opening-rail-lines

Campaign for Borders Rail. (2016, December 3). *Rail monitoring group attacks Borders Railway 'forecasting failure'.* Webpage. Retrieved 5 September 2017, from https://campaignforbordersrail.wordpress.com/2016/12/03/rail-monitoring-group-attacks-borders-railway-forecasting-failure

Cantillo, V., & Ortúzar, J. d. D. (2005). A semi-compensatory discrete choice model with explicit attribute thresholds of perception. *Transportation Research Part B: Methodological*, *39*(7), 641–657.

Castro, M., Martinez, F., & Munizaga, M. (2009, March). *Calibration of a logit discrete choice model with elimination of alternatives.* Paper presented at the International Choice Modelling Conference.

Cervero, R., Round, A., Goldman, T., & Wu, K.-L. (1995). *Rail access modes and catchment areas for the BART system* (Working Paper No. UCTC No. 307). The University of California Transportation Center.

Chakour, V., & Eluru, N. (2014). Analyzing commuter train user behavior: a decision framework for access mode and station choice. *Transportation*, *41*(1), 211–228.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2017). Shiny: Web application framework for R [Computer software]. Retrieved from https://CRAN.R-project.org/package=shiny (R package version 1.0.5)

Chen, C., Xia, J. C., Smith, B., & Han, R. (2014). Development of a conceptual framework for modeling train station choice under uncertainty for park-and-ride users. In

Transportation Research Board 93rd Annual Meeting.

Chen, C., Xia, J. C., Smith, B., Olaru, D., Taplin, J., & Han, R. (2015). Influence of parking on train station choice under uncertainty for park-and-ride users. *Procedia Manufacturing*, *3*, 5126–5133.

Chorus, C. (2012). Random regret minimization: An overview of model properties and empirical evidence. *Transport Reviews*, *32*(1), 75–92.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S. K., & Tiffin, N. (2016). RPostgreSQL: R interface to the PostgreSQL database system [Computer software]. Retrieved from https://CRAN.R-project.org/package=RPostgreSQL (R package version 0.4-1)

Davidson, B., & Yang, L. (1999). *Modeling of commuter rail station choice and access mode combinations.* Paper presented at the Transportation Research Board Annual Meeting.

Debrezion, G., Pels, E., & Rietveld, P. (2007a). Choice of departure station by railway users. *European Transport*, *37*, 78–92.

Debrezion, G., Pels, E., & Rietveld, P. (2007b). *Modelling the joint access mode and railway station choice* (Discussion Paper No. TI 2007-012/3). Tinbergen Institute.

Debrezion, G., Pels, E., & Rietveld, P. (2009). Modelling the joint access mode and railway station choice. *Transportation Research Part E: logistics and transportation review*, *45*(1), 270–283.

Department for Transport. (2011). *Guidance note on passenger demand forecasting for third party funded local rail schemes.*

Department for Transport. (2015). *Free flow vehicle speed statistics: Great Britain 2014*.

Department for Transport. (2017a). *Connecting people: a strategic vision for rail.* Retrieved 24/03/2018, from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/663124/rail-vision-web.pdf

Department for Transport. (2017b). *Rail factsheet.* Retrieved 12 January 2018, from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/663116/rail-factsheet-2017.pdf

Desfor, G. (1975). Binary station choice models for a rail rapid transit line. *Transportation Research*, *9*(1), 31–41.

Econometric Software Inc. (2012). NLOGIT 5 [Computer software].

Edinburgh Evening News. (2015, December 2). *Borders Railway service 'miserable', campaigners say.* Webpage. Retrieved 9 November 2011, from http://www.edinburghnews.scotsman.com/news/borders-railway-service-miserable-campaigners-say-1-3964218

Fan, K.-S., Miller, E. J., & Badoe, D. (1993). Modeling rail access mode and station choice. *Transportation Research Record*, *1413*, 49–59.

Fortmann-Roe, S. (2018). *Accurately Measuring Model Prediction Error.* Webpage. Retrieved 28 August 2018, from http://scott.fortmann-roe.com/docs/MeasuringError.html

Fotheringham, A. S. (1986). Modelling hierarchical destination choice. *Environment and Planning A*, *18*(3), 401–418.

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, *23*(7), 1025–1044.

Fox, J. (2005, October). *Modelling park-and-ride in the PRISM Model for the West Midlands Region.* Paper presented at the AET European Transport Conference.

Fox, J., Daly, A., Patruni, B., & Milthorpe, F. (2011, September). *Extending the Sydney Strategic Model to represent toll road and park-and-ride choices.* Paper presented at the 34th Australasian Transport Research Forum, Adelaide, Australia.

GB Rail. (2015). *GB Rail GTFS.* Webpage. Retrieved 27 April 2017, from http://www.gbrail.info/

Geofabrik. (2015). *Downloads.* Webpage. Retrieved from http://www.geofabrik.de/data/download.html

Giannopoulos, G., & Boulougaris, G. (1989). Definition of accessibility for railway stations and its impact on railway passenger demand. *Transportation Planning and Technology*, *13*(2), 111–120.

Givoni, M., & Rietveld, P. (2007). The access journey to the railway station and its role in passengers' satisfaction with rail travel. *Transport Policy*, *14*(5), 357–365.

Givoni, M., & Rietveld, P. (2014). Do cities deserve more railway stations? The choice of a departure railway station in a multiple-station region. *Journal of Transport Geography*, *36*, 89–97.

Glasgow, G. (2001, July). Mixed logit models in political science. In *Eighteenth Annual Political Methodology Summer Conference.*

Google. (2015). *The Google distance matrix API.* Webpage. Retrieved 10 April 2018, from https://developers.google.com/maps/documentation/distance-matrix/usage-limits

GoogleTransitDataFeed. (2016). *GoogleTransitDataFeed wiki.* Webpage. Retrieved 10 March 2016, from https://code.google.com/archive/p/googletransitdatafeed/wikis/GoogleTransitDataFeed.wiki

Green, C., & Hall, P. (2009). *Better rail stations* (An independent review presented to Lord Adonis, Secretary of State for Transport).

Greene, W. H. (2012). NLOGIT version 5 reference guide [Computer software manual].

Guan, H., Yin, Y., Yan, H., Han, Y., & Qin, H. (2007). Urban railway accessibility. *Tsinghua Science & Technology*, *12*(2), 192–197.

Guerra, E., Cervero, R., & Tischler, D. (2012). Half-mile circle. *Transportation Research Record: Journal of the Transportation Research Board*, *2276*(1), 101–109.

Hagen, M. v., & Heiligers, M. (2011, October). *Effect of station improvement measures on customer satisfaction.* Paper presented at the AET European Transport Conference.

Hanson, S. (1977). Measuring the cognitive levels of urban residents. *Geografiska Annaler: Series B, Human Geography*, *59*(2), 67–81.

Harata, N., & Ohta, K. (1986). Some findings on the application of disaggregate nested logit model to railway station and access mode choice. In *Research for tomorrows transport requirements: Proceedings of the world conference on transport research* (Vol. 2, pp. 1729–1740).

Hensher, D. A., Rose, J. M., & Greene, W. H. (2005). *Applied choice analysis: a primer*. Cambridge University Press.

Hensher, D. A., Rose, J. M., & Greene, W. H. (2016). *Applied choice analysis* (2nd ed.). Cambridge University Press.

Hess, S., Beck, M., & Crastes dit Sourd, R. (2017, January). *Can a better model specification avoid the need to move away from random utility maximisation?* Paper presented at 96th Annual Meeting of the Transportation Research Board.

Hess, S., Daly, A., & Batley, R. (2018, Mar). Revisiting consistency with random utility maximisation: theory and implications for practical work. *Theory and Decision*, *84*(2), 181–204. doi: 10.1007/s11238-017-9651-7

Ho, C. Q., & Hensher, D. A. (2016, February). A workplace choice model accounting for spatial competition and agglomeration effects. *Journal of Transport Geography*, *51*, 193–203.

Horni, A., Charypar, D., & Axhausen, K. W. (2010, September). *Empirically approaching destination choice set formation.* Paper presented at 10th Swiss Transport Research Conference, Ascona, Switzerland.

Hunt, L. M., Boots, B., & Kanaroglou, P. S. (2004). Spatial choice modelling: new opportunities to incorporate space into substitution patterns. *Progress in Human Geography*, *28*(6), 746–766.

Ivanescu, A. E., Li, P., George, B., Brown, A. W., Keith, S. W., Raju, D., & Allison, D. B. (2015). The importance of prediction model validation and assessment in obesity and nutrition research. *International Journal of Obesity*, *40*(6), 887.

Jiang, Y., Zegras, C., & Mehndiratta, S. (2012). Walk the line: station context, corridor type and bus rapid transit walk access in Jinan, China. *Journal of Transport Geography*, *20*(1), 1–14.

Jones, S., & Hensher, D. A. (2008). An evaluation of open-and closed-form distress prediction models: The nested logit and latent class models. In S. Jones & D. A. Hensher (Eds.), *Advances in credit risk modelling and corporate bankruptcy prediction* (pp. 80–113). Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Kastrenakes, C. R. (1988). Development of a rail station choice model for NJ Transit. *Transportation Research Record*, *1162*, 16–21.

Keijer, M., & Rietveld, P. (2000). How do people get to the railway station? the Dutch experience. *Transportation Planning and Technology*, *23*(3), 215–235.

Koppelman, F. S., & Bhat, C. (2006). A self instructing course in mode choice modeling: multinomial and nested logit models. *US Department of Transportation, Federal Transit Administration*, *31*.

Koppelman, F. S., & Sethi, V. (2000). Closed-form discrete-choice models. In *Handbook of Transport Modelling* (pp. 211–227). Elsevier Science Ltd.

Lang, D. T., & the CRAN Team. (2017). XML: Tools for parsing and generating XML within R and S-Plus [Computer software]. Retrieved from https://CRAN.R-project.org/package=XML (R package version 3.98-1.9)

Lin, T. G., Xia, J. C., Robinson, T. P., Goulias, K. G., Church, R. L., Olaru, D., ... Han, R. (2014). Spatial analysis of access to and accessibility surrounding train stations: a case study of accessibility for the elderly in Perth, Western Australia. *Journal of Transport Geography*, *39*, 111–120.

Liou, P. S., & Talvitie, A. P. (1974). Disaggregate access mode and station choice models for rail trips. *Transportation Research Record*, *526*, 42–65.

live-departures.info. (2017). *Stations served by electric trains* [List derived from current timetable feed where power type specified as E (Electric) or EMU (Electric Multiple Unit).]. Webpage. Retrieved 27 March 2017, from https://live-departures.info/rail/trivia/StationsServedByElectricTrains

Local Transport Today. (2013, March 11). *Full speed ahead for Borders Railway despite BCR of just 0.5.* Webpage. Retrieved 16/01/2018, from https://www.transportxtra.com/publications/local-transport-today/news/33878/full-speed-ahead-for-borders-railway-despite-bcr-of-just-0-5

Lythgoe, W. (2004). *Enhancing cross-sectional rail passenger demand models* (Unpublished doctoral dissertation). University of Leeds (Institute for Transport Studies).

Lythgoe, W., & Wardman, M. (2002, September). *Estimating passenger demand for parkway stations.* Paper presented at AET European Transport Conference.

Lythgoe, W., & Wardman, M. (2004). Modelling passenger demand for parkway rail stations. *Transportation*, *31*(2), 125–151.

Lythgoe, W., Wardman, M., & Toner, J. (2004, October). *Enhancing rail passenger demand models to examine station choice and access to the rail network.* Paper presented at AET European Transport Conference.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2017). Cluster: Cluster analysis basics and extensions [Computer software]. (R package version 2.0.6)

Mahmoud, M. S., Eng, P., & Shalaby, A. (2014, January). *Park-and-ride access station choice model for cross-regional commuter trips in the Greater Toronto and Hamilton Area (GTHA).* Paper presented at Transportation Research Board 93rd Annual Meeting.

Maindonald, J. H., & Braun, W. J. (2015). DAAG: Data Analysis and Graphics Data and Functions [Computer software]. Retrieved from https://CRAN.R-project.org/package=DAAG (R package version 1.22)

McFadden, D., & Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, *15*(5), 447-470.

Midlothian Council. (2017, January 30). *Borders Railway boosts tourism.* Webpage. Retrieved 27/11/2017, from https://www.midlothian.gov.uk/news/article/2114/borders_railway_boosts_tourism

Mirkes, E. (2011). *K-means and K-medoids applet.* Webpage. Retrieved 30/01/2018, from http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html

Munizaga, M. A., & Alvarez-Daziano, R. (2001, June). *Mixed logit vs. nested logit and probit models.* Paper presented at 5th tri-annual Invitational Choice Symposium, Asilomar.

MVA Consultancy. (2011). *Making better decisions. assessment of aspirations for track access on the West Coast Main Line.*

National Rail Enquiries. (2016). *Knowledgebase XMLs.* Webpage. Retrieved 10 March 2016, from http://www.nationalrail.co.uk/100298.aspx

Network Rail. (2018, February). *A better railway for a better Britain - strategic business plan 2019 - 2024.* Retrieved 23/03/2018, from https://cdn.networkrail.co.uk/wp-content/uploads/2018/02/Strategic-business-plan-high-level-summary.pdf

Nomis. (2013). *Postcode headcounts and household estimates - 2011 census* [CSV format]. Retrieved 14 August 2017, from https://www.nomisweb.co.uk/census/2011/postcode_headcounts_and_household_estimates

Nomis. (2014, May 23). *WP102EW - population density (workplace population)* [CSV format]. Retrieved 14 August 2017, from https://www.nomisweb.co.uk/census/2011/wp102ew

North Star. (2014, September 28). *Anger as trains skip newly-reopened Ross station.* Webpage. Retrieved from http://www.north-star-news.co.uk/News/Anger-as-trains-skip-newly-reopened-Ross-station-27092014.htm

Office of Rail and Road. (2013, May). *Estimates of station usage - 2011-12 report and data* [Excel spreadsheet]. Webpage. Retrieved from http://orr.gov.uk/statistics/published-stats/station-usage-estimates

Office of Rail and Road. (2017). *Passenger journeys by year - Table 12.5.* Webpage. Retrieved 5 September 2017, from http://dataportal.orr.gov.uk/browsereports/12

Office of Rail Regulation. (2017). *Passenger rail usage 2016-17 Q4 statistical release.* Retrieved 22-06-2017, from http://www.orr.gov.uk/__data/assets/pdf_file/0019/24832/passenger-rail-usage-2016-17-q4.pdf

Openshaw, S. (1984). *The modifiable areal unit problem* (Vol. 38). Geo Books, Norwich.

OpenStreetMap. (2015). *United Kingdom tagging guidelines.* Webpage. Retrieved 1 June 2015, from http://wiki.openstreetmap.org/wiki/United_Kingdom_Tagging_Guidelines

OpenTripPlanner. (2018). An open source multi-modal trip planner [Computer software]. Retrieved from https://github.com/opentripplanner/OpenTripPlanner

Ordnance Survey. (2015, June). *AddressBase technical specification v2.1.* Retrieved 12 July 2017, from https://www.ordnancesurvey.co.uk/docs/technical-specifications/addressbase-technical-specification.pdf

Ordnance Survey. (2016). *OS Open Roads - version:11/16* [ESRI Shapefile]. Retrieved from https://www.ordnancesurvey.co.uk/opendatadownload/products.html#OPROAD

Ordnance Survey. (2017). *OS Open Roads: User guide and technical specification.* Retrieved from https://www.ordnancesurvey.co.uk/docs/user-guides/os-open-roads-user-guide.pdf

Ortuzar, J. d. D. (1980). *Multimodal choice modelling - some relevant issues*. Institute of Transport Studies, University of Leeds.

Ortúzar, J. d. D., & Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons.

Pagliara, F., & Timmermans, H. (2009). Choice set generation in spatial contexts: a review.

Transportation Letters, *1*(3), 181–196.

Pang, H., & Khani, A. (2018). Modeling park-and-ride location choice of heterogeneous commuters. *Transportation*, *45*(1), 71–87.

Park, S., Kang, J., & Choi, K. (2014). Finding determinants of transit users' walking and biking access trips to the station: A pilot case study. *KSCE Journal of Civil Engineering*, *18*(2), 651–658.

Passenger Focus. (2007). *Getting to the station. findings of research conducted at Witham*.

Passenger Focus. (2011). *The challenge of getting to the station passenger experiences*.

Passenger Focus. (2012). *Future priorities for the West Coast Main Line: Released capacity from a potential high speed line*.

Pellegrini, P. A., & Fotheringham, A. S. (2002). Modelling spatial choice: a review and synthesis in a migration context. *Progress in Human Geography*, *26*(4), 487–510.

Potter, R. B. (1979). Perception of urban retailing facilities: An analysis of consumer information fields. *Geografiska Annaler: Series B, Human Geography*, *61*(1), 19–29.

Preston, J. (1991a). Demand forecasting for new local rail stations and services. *Journal of Transport Economics and Policy*, 183–202.

Preston, J. (1991b). *Passenger demand forecasting for new rail services - manual of advice* (Working Paper No. 352). Institute of Transport Studies, University of Leeds.

Preston, J., & Aldridge, D. (1991). *Greater Manchester PTE new railway station demand prediction model*. Institute of Transport Studies, University of Leeds.

Preston, J., Blainey, S., Wall, G., Chintakayala, P., & Wardman, M. (2008, October). *The effects of station enhancements on rail demand.* Paper presented at AET European Transport Conference.

Preston, J., & Dargay, J. (2005, October). *The dynamics of rail demand.* Paper presented at Third Conference on Railroad Industry Structure, Competition and Investment.

Prior, M., Vickers, J., Segal, J., & Quill, J. (2011, October). *Modelling open access train services.* Paper presented at AET European Transport Conference.

Puello, L. L. P., & Geurs, K. (2015). Modelling observed and unobserved factors in cycling to railway stations: application to transit-oriented-developments in the Netherlands. *EJTIR*, *15*(1), 27–50.

RailEngineer. (2016, April 22). *After Borders, what next?* Webpage. Retrieved 12 January 2018, from https://www.railengineer.uk/2016/04/22/after-borders-what-next

Railfuture. (2018). *New stations.* Webpage. Retrieved 12 January 2018, from http://www.railfuture.org.uk/New+stations

Ramsey, P. (2011). *Indexed nearest neighbour search in PostGIS.* Webpage. Retrieved 24 May 2015, from https://boundlessgeo.com/2011/09/indexed-nearest-neighbour-search-in-postgis/

Rodrigue, J. P., Comtois, C., & Slack, B. (2013). *The geography of transport systems*. Routledge.

Scotland's Census. (2013). *Statistical bulletin - release 1C (part two) - Table A1: Census day estimates of usually resident population and households by postcode, 2011.* Webpage. Retrieved 14 August 2017, from http://www.scotlandscensus.gov.uk/bulletin-figures

-and-tables

Scotland's Census. (2016, November 24). *WP101SC - population count - workplace population* [Excel spreadsheet]. Webpage. Retrieved 16 August 2017, from http://www.scotlandscensus.gov.uk/ods-web/data-warehouse.html

Sener, I. N., Pendyala, R. M., & Bhat, C. R. (2011). Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. *Journal of Transport Geography*, *19*(2), 294–303.

Sharma, B., Hickman, M., & Nassir, N. (2017). Park-and-ride lot choice model using random utility maximization and random regret minimization. *Transportation*. (Advanced online publication) doi: 10.1007/s11116-017-9804-0

Smith, G. C. (1976). The spatial information fields of urban consumers. *Transactions of the Institute of British Geographers*, *1*(2), 175–189.

Sober, E. (2002). Instrumentalism, parsimony, and the Akaike framework. *Philosophy of Science*, *69*, S112–S123.

Song, Y., & Rodríguez, D. A. (n.d.). *The measurement of the level of mixed land uses: A synthetic approach* (White Paper Series). Carolina Transportation Program. Retrieved 29/01/2018, from http://planningandactivity.unc.edu/Mixed%20land%20uses%20White%20Paper.pdf

Spaven, D. (2017). *Waverley Route: The Battle for the Borders Railway* (Third ed.). Stenlake Publishing.

Steer Davies Gleave. (2010). *Station usage and demand forecasts for newly opened railway lines and stations* (Final Report prepared for Department for Transport). Retrieved from https://www.gov.uk/government/publications/new-stations-study

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, *54*(8), 774–781.

The Inverness Courier. (2015, September 12). *Concern poor rail service forcing people back to roads amid reports passengers left standing at Beauly and Conon Bridge.* Webpage. Retrieved 1/11/2017, from https://www.inverness-courier.co.uk/News/Concern-poor-rail-service-forcing-people-back-to-roads-amid-reports-passengers-left-standing-at-Beauly-and-Conon-Bridge-11092015.htm

The PostGIS Development Group. (2017). *PostGIS Manual: Geometry Processing: ST_ConcaveHull.* Webpage. Retrieved 14 July 2017, from https://postgis.net/docs/ST_ConcaveHull.html

The PostGIS Development Group. (2018). *PostGIS Manual: Geometry Processing: ST_Azimuth.* Webpage. Retrieved 25 Jan 2018, from https://postgis.net/docs/ST_Azimuth.html

The PostgreSQL Global Development Group. (2017). *PostgreSQL 9.4.12 Documentation, Appendix F. Additional Supplied Modules, F.31. pg_trgm.* Webpage. Retrieved 12 July 2017, from https://www.postgresql.org/docs/9.4/static/pgtrgm.html

The Scotsman. (2016, September 9). *Minister orders improvements to fix Borders Railway's 'unacceptable' performance.* Webpage. Retrieved 9 November 2011,

from http://www.scotsman.com/news/transport/minister-orders-improvements-to-fix-borders-railway-s-unacceptable-performance-1-4225643

Thill, J. C. (1992). Choice set formation for destination choice modelling. *Progress in Human Geography*, *16*(3), 361–382.

Tobler, W. (1970). Computer movie simulating urban growth in Detroit region. *Economic Geography*, *46*(2), 234–240.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

TransitFeeds. (2017). *ATOC GTFS*. Webpage. Retrieved 11 September 2017, from http://transitfeeds.com/p/association-of-train-operating-companies/284

Transport Focus. (2015a). *National Rail Passenger Survey: Spring 2015: Data* [CSV]. Retrieved 13 June 2017, from https://data.gov.uk/dataset/national-rail-passenger-survey

Transport Focus. (2015b). *National Rail Passenger Survey: Spring 2015 main report*.

Transport Scotland. (2012, November). *Borders Railway final business case - final version*. Retrieved 16/01/2018, from https://www.transport.gov.scot/media/10321/ts_borders_fbc_final_version_issued.pdf

Transport Scotland. (2016, August 19). *Galashiels transport interchange celebrates first year*. Webpage. Retrieved 24/11/2017, from http://www.bordersrailway.co.uk/news/galashiels-transport-interchange-celebrates-first-year

Transport Scotland. (2017, June). *Borders Railway year 1 evaluation*. Retrieved 25 August 2017, from https://www.transport.gov.scot/media/39335/sct04173824741.pdf

UK Data Service. (2011). *Census geography* [CSV format]. Webpage. Retrieved from https://borders.ukdataservice.ac.uk

Vanwinckelen, G., & Blockeel, H. (2012). On estimating model accuracy with repeated cross-validation. In *Benelearn 2012: Proceedings of the 21st belgian-dutch conference on machine learning* (pp. 39–44).

Waddell, P., Bhat, C., Eluru, N., Wang, L., & Pendyala, R. M. (2007). Modeling interdependence in household residence and workplace choices. *Transportation Research Record: Journal of the Transportation Research Board*, *2003*(1), 84–92.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

Wardman, M., & Tyler, J. (2000). Rail network accessibility and the demand for inter-urban rail travel. *Transport Reviews*, *20*(1), 3–24.

Wardman, M., & Whelan, G. (1999). *Using geographical information systems to improve rail demand models*. (Final Report to Engineering and Physical Sciences Research Council)

Wei, T., & Simko, V. (2017). R package "corrplot": Visualization of a correlation matrix [Computer software]. Retrieved from https://github.com/taiyun/corrplot ((Version 0.84))

Weiss, A., & Habib, K. N. (2016). Examining the difference between park and ride and kiss and ride station choice using spatially weighted error correlation choice model. In *Transportation Research Board 95th Annual Meeting*.

Weiss, A., & Habib, K. N. (2017). Examining the difference between park and ride and kiss

and ride station choices using a spatially weighted error correlation (SWEC) discrete choice model. *Journal of Transport Geography*, *59*, 111–119.

Wen, C. H., & Koppelman, F. S. (2001). The generalized nested logit model. *Transportation Research Part B: Methodological*, *35*(7), 627–641.

Whelan, G., Toner, J., Mackie, P., & Preston, J. (2001, July). *Modelling quality bus partnerships.* Paper presented at the 9th World Conference on Transport Research.

Why do some stations have just one train a week? (2015, March 28). Webpage. Retrieved 16 August 2017, from http://www.rail.co.uk/rail-news/2015/closures

Wieland, T. (2017). Market Area Analysis for Retail and Service Locations with MCI. *The R Journal*, *9*(1), 298–323. Retrieved from https://journal.r-project.org/archive/2017/RJ-2017-020/RJ-2017-020.pdf

Wikipedia contributors. (2017). *Galashiels — wikipedia, the free encyclopedia.* Retrieved 12 January 2018, from https://en.wikipedia.org/w/index.php?title=Galashiels&oldid=812721848

Wikipedia contributors. (2018). *Borders railway — wikipedia, the free encyclopedia.* Retrieved 12 January 2018, from https://en.wikipedia.org/w/index.php?title=Borders_Railway&oldid=819971833

Worsley, T. (2012). *Rail demand forecasting - using the passenger demand forecasting handbook* (On the Move - Supporting Paper 2). RAC Foundation.

Young, M. (2016, March). *An automated framework to derive model variables from open transport data using R, PostgreSQL and OpenTripPlanner.* Paper presented at 24th GIS Research UK Conference.

Young, M. (2017a, January). *Developing railway station choice models to improve rail industry demand models.* Paper presented at 49th Annual UTSG Conference, Dublin, Ireland.

Young, M. (2017b, October). *Development of integrated demand and station choice models for local railway stations and services.* Paper presented at AET European Transport Conference, Barcelona, Spain.

Young, M., & Blainey, S. (2016, January). *Defining probability-based rail station catchments for demand modelling.* Paper presented at 48th Annual UTSG Conference, Bristol, GB.

Young, M., & Blainey, S. (2018a). Development of railway station choice models to improve the representation of station catchments in rail demand models. *Transportation Planning and Technology*, *41*(1), 80–103.

Young, M., & Blainey, S. (2018b). Railway station choice modelling: a review of methods and evidence. *Transport Reviews*, *38*(2), 232–251.

Zervaas, Q. (2014). *The definitive guide to GTFS*. Author.

Zhao, F., Chow, L.-F., Li, M.-T., Ubaka, I., & Gan, A. (2003). Forecasting transit walk accessibility: regression model alternative to buffer method. *Transportation Research Record: Journal of the Transportation Research Board*, *1835*(1), 34–41.

Zolfaghari, A., Sivakumar, A., & Polak, J. (2013). Simplified probabilistic choice set formation models in a residential location choice context. *Journal of Choice Modelling*, *9*, 3–13.