*This is an Accepted Manuscript of an article published by Taylor & Francis in Transportation Planning and Technology on 23 Nov 2017, available online:*
http://www.tandfonline.com/10.1080/03081060.2018.1403745.

# Development of railway station choice models to improve the representation of station catchments in rail demand models

Marcus A. Young[a] and Simon P. Blainey[a]

[a]Transportation Research Group, University of Southampton, Boldrewood Innovation Campus, Burgess Road, Southampton, SO16 7QF, UK

**ARTICLE HISTORY**
Compiled June 30, 2017

**ABSTRACT**
This paper describes the development of railway station choice models suitable for defining probabilistic station catchments. These catchments can then be incorporated into the aggregate demand models typically used to forecast demand for new rail stations. Revealed preference passenger survey data obtained from the Welsh and Scottish Governments was used for model calibration. Techniques were developed to identify trip origins and destinations from incomplete address information and to automatically validate reported trips. A bespoke trip planner was used to derive mode-specific station access variables and train leg measures. The results from a number of multinomial logit and random parameter (mixed) logit models are presented and their predictive performance assessed. The models were found to have substantially superior predictive accuracy compared to the base model (which assumes the nearest station has a probability of one), indicating that their incorporation into passenger demand forecasting methods has the potential to significantly improve model predictive performance.

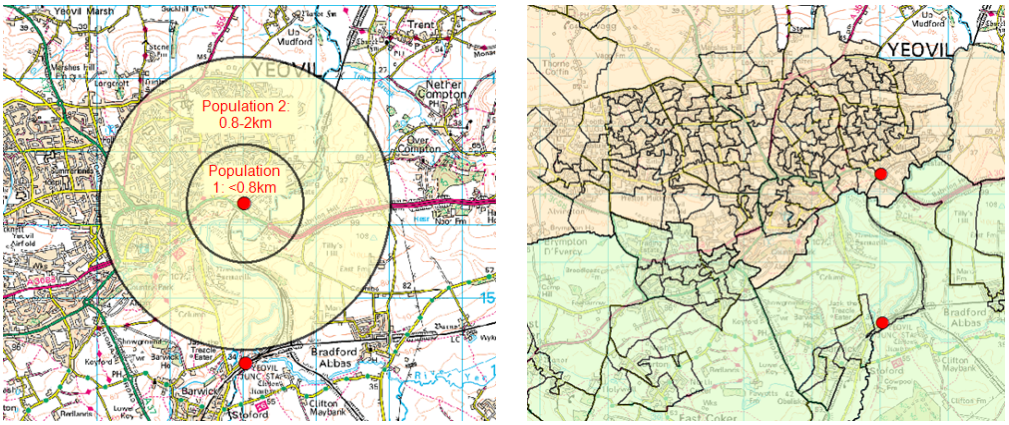
**KEYWORDS**
railway station choice; discrete choice models; passenger demand forecasting

## 1. Introduction

There has been substantial growth in rail passenger journeys in the UK and several other European countries over recent years (Amadeus 2013; Office of Rail Regulation 2017); a renaissance in rail travel that has seen more than 50 new stations open in the UK during the last decade (Railfuture 2017). This growth is projected to continue and there is considerable interest in opening new railway stations to serve local communities across the country. To assess the viability of any proposed scheme, it is important that demand for a new station can be accurately forecast. Trip end models are commonly used in the UK for this purpose, and were adopted in the appraisal stages of around two-thirds of recently opened stations/lines in the UK (where information was available) (Steer Davies Gleave 2010). Trip end models consider demand to be a function of the station's catchment population and other variables related to

CONTACT Marcus A. Young. Email: m.a.young@soton.ac.uk, Tel: 44(0)7762296430, ORCID: 0000-0003-4627-1116, Twitter: @MAYoungUK

Simon P. Blainey. Email: s.p.blainey@soton.ac.uk, Tel: 44(0)2380592834, ORCID: 0000-0003-4249-8110, Twitter: @SPBTransport




(a) Radial catchment divided into two bands (b) Zone-based catchment

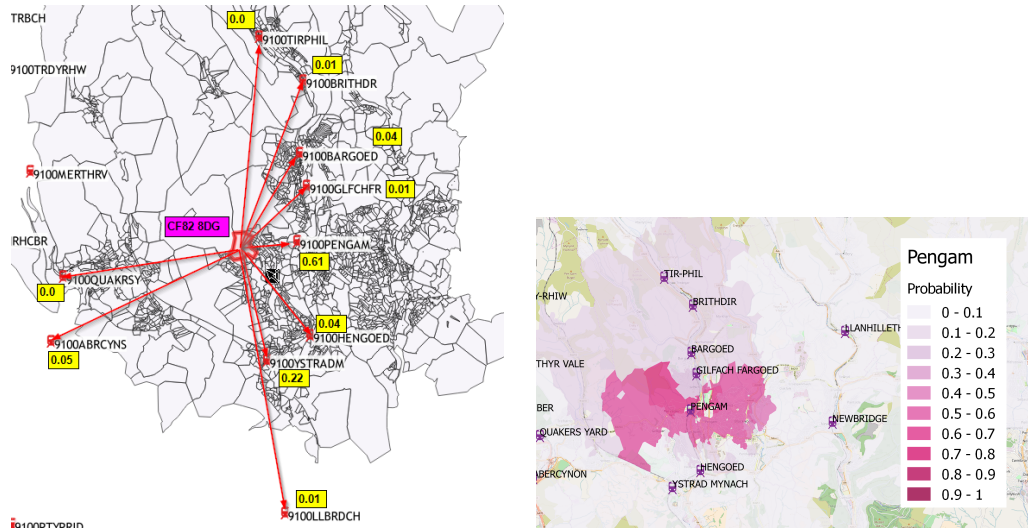
**Figure 1.** Illustration of simple catchment definitions commonly used in new station forecast models

the station's locality or the facilities and services to be provided. This type of model therefore requires a catchment to be defined, so that the population from which demand will be generated can be specified. The methods commonly used to define a catchment are simplistic, for example placing distance-based circular buffers around the station (see Figure 1a), or dividing the area of interest into zones and assigning each zone to its nearest station (see Figure 1b). However, recent research shows that in reality station catchments are far more complex entities (for example, see Blainey and Evens (2011)). Simplistic catchments account for only 50-60 percent of observed trips, station choice is not homogeneous within zones, catchments overlap, and catchments vary by access mode and station type. The inability of these simplistic catchments to capture patterns of abstraction and competition between railway stations may have contributed to the limited accuracy of many recent demand forecasts for new stations, an issue of sufficient concern that the UK Department for Transport recently commissioned a study to investigate the issue (Steer Davies Gleave 2010). This suggests that rail demand models might be more accurate if probability-based catchments were defined.

The purpose of the research described in this paper is to develop transferable station choice models which can be used to generate probabilistic station catchments that can subsequently be integrated into trip end or flow models[1]. More specifically, the aim is the ability to predict the probability of a station being chosen, from a set of alternative (or 'competing') stations, at the unit postcode level (as illustrated in Figure 2a); allowing the probabilistic catchment of a station to be generated at a high spatial resolution (as illustrated in Figure 2b). This work builds on an earlier pilot study that established suitable data processing techniques and developed some initial models using a small survey dataset (Young and Blainey 2016).

Following a brief overview of previous research in this field, the paper describes the observed station choice datasets used in this study and the procedures adopted to prepare and validate them. The modelling methodology section then considers the theoretical basis of discrete choice models; the generation of choice sets for each survey respondent; the derivation of explanatory variables; and the model calibration process.

[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. Research has shown that these models have the potential to more accurately forecast station demand (Blainey and Preston 2010)




(a) Probability of each alternative station being chosen for a postcode

(b) Probability of a station being chosen by postcode

**Figure 2.** Illustration of mechanism for defining a probability-based catchment

The results from a number of calibrated models are then presented and discussed, with consideration given to model validation and the assessment of predictive performance. Finally, conclusions are drawn and the potential for future work identified. The key steps involved in cleaning and validating the observed choice data and preparing the models for calibration are summarised in Figure 3.

## 2. Previous research

Prior station choice research, which has predominantly adopted the multinomial logit (MNL) model or, when jointly modelling access mode choice, the nested logit model, has primarily focused on explaining the factors that influence choice, rather than seeking to calibrate models that might be useful in forecasting demand for new stations. Where the aim has been to improve demand models, this has usually focused on addressing specific local needs, such as the work of Harata and Ohta (1986) and Kastrenakes (1988). Wardman and Whelan (1999) tried to incorporate probabilistic catchments into a flow model by apportioning population to five competing stations based on probabilities calculated at the postal sector level. However, the model failed to converge, which they attributed to using only a subset of their data due to time constraints and limited computer processing power. This appears to have been the only attempt to implement an approach similar to the one adopted in this paper. Lythgoe and Wardman (2002, 2004) developed a unique approach to forecasting demand for parkway stations, but its applicability is limited to long inter-urban journeys. Previous research has also lacked a rigorous assessment of model predictive performance, either against the sample used to calibrate the model or in other contexts, and despite broadly consistent reporting of the direction effects of a range of explanatory variables, no attempt has been made to develop a generalised and transferable model. Unlike many previous studies, the research described in this paper has an applied focus, seeking to develop station choice models that can be incorporated into the trip end or flow models that are used to assess proposals for new railway stations or substantial service

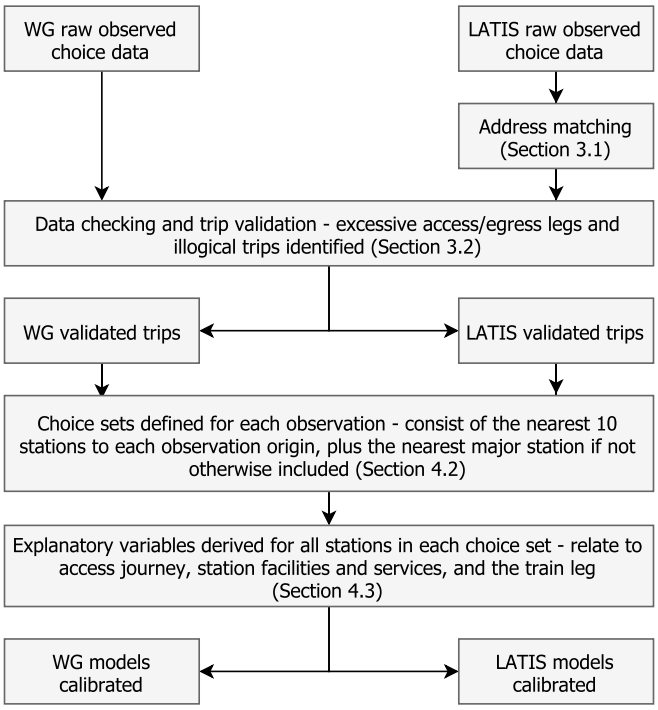



**Figure 3.** Summary of key steps involved in cleaning and validating observed choice data and preparing models for calibration

changes. For a comprehensive review of prior research in this area, see Young and Blainey (2017).

## 3. Observed station choice data

In order to calibrate a discrete choice model, data is needed on observed station choice. This study used revealed preference data from a series of on-train passenger surveys that were obtained from the Welsh Government (WG) and Transport Scotland's 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 WG surveys were carried out in 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 2014 and 2015 and, although concentrated in the Central Belt, covered stations throughout Scotland. 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 datasets contained some 7,000 and 50,000 observations respectively.

The remainder of this section describes the steps that were carried out to maximise the usefulness of the survey data and to ensure, as far as reasonably practicable, the validity of the reported trips. As several of the methods adopted are potentially novel,

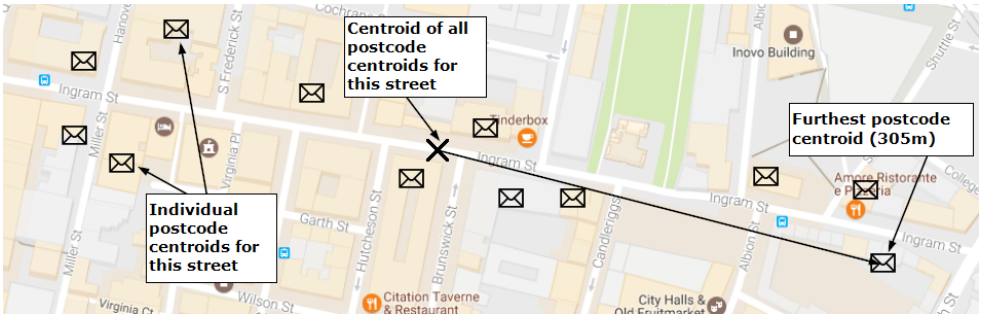



**Figure 4.** Postcode centroids for Ingram Street, Glasgow, showing calculated centroid and maximum distance from calculated centroid to any postcode centroid. Map data ©2017 Google.

and may have wider applicability to researchers using this type of data, sufficient detail is provided to enable them to be applied elsewhere.

### 3.1. *Address matching*

Many observations in the LATIS dataset had missing, incorrect or incomplete postcodes, with less than 50% of the origin addresses having a valid unit-level postcode. Survey respondents are likely to know the origin or destination postcode for particular types of trip, such as those beginning or ending at their home address, but not for others. 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's AddressBase product which contains over 28 million UK addresses. The aim was to either identify a specific postcode from the provided address information or, failing that, to approximate the geographic location of an address. The AddressBase file was imported into a PostgreSQL table and several new fields were generated. The first counted the number of distinct postcodes for each unique postal town:thoroughfare combination. The second calculated the centroid of all the individual postcode centroids belonging to each thoroughfare, and the third measured the maximum Euclidean distance from the calculated centroid to any of the individual postcode centroids. This process is illustrated in Figure 4. If the calculated centroid is used to represent the location of an origin or destination on a street, the maximum Euclidean distance indicates how far the 'real' address postcode centroid could be from that location. Next, origin and destination addresses in the LATIS survey dataset were matched to AddressBase addresses using a trigram index, with the top four matches for each observation retrieved. A manual review process was then completed, using the following 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, that postcode is accepted
    (b) if street has more than one postcode, if the maximum Euclidean distance is <= 250m, use the coordinates of the calculated street centroid as the origin or destination location.

The address matching process resulted in a 31% and 58% increase in the number of usable trip origins and destinations respectively.

### 3.2. *Data validation*

#### *3.2.1. General data checks*

A variety of data checks were carried out on both datasets, including removing observations where the access or egress mode was not provided; where the origin and/or destination station was missing or the name was incorrect or ambiguous; and where the origin station was the same as the destination station. To limit the amount of public transit schedule data that needed to be incorporated into the bespoke trip planner (see Section 4.3), only those observations where the origin was located in Wales (for WG dataset) or Scotland (for LATIS dataset) were retained. In addition, any observations with origins or destinations outside of GB or 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.

Due to the large number of observations 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.

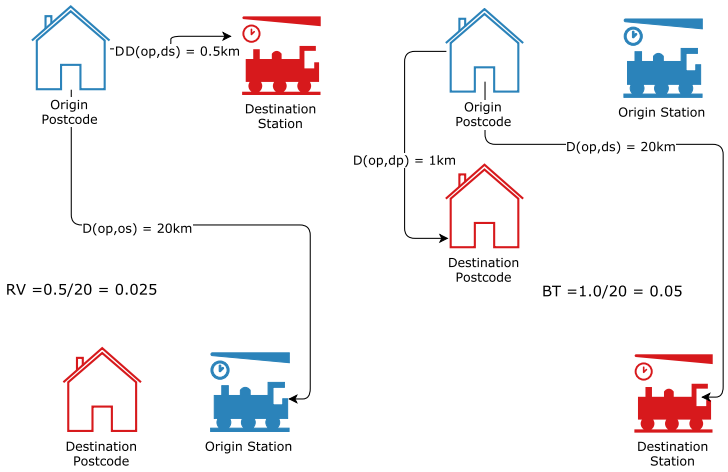
#### *3.2.2. Identifying excessive access or egress legs*

For each observation in the cleaned data, the trip planner was used to obtain the walk-time in minutes from the trip origin to the origin (boarding) station; and from the destination (alighting) station to the trip destination. A histogram and kernel density plot was then produced for access time and egress time and 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. A similar process was used to identify excessively long access and egress legs for the other modes, and those in excess of 70km were removed from the WG data and those in excess of 200km were removed from the LATIS data.[2]

#### *3.2.3. Identifying illogical trips*

There are two main types of illogical trips that are commonly observed in this type of data. The first is the 'reversed trip' where the origin station is located close to the trip destination, and the destination station is close to the trip origin. The second occurs when there is a substantial 'back-track' from the reported destination station towards the trip origin. A range of ratios were tested, 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. The closer this ratio is to zero, the more pronounced the reversal effect becomes (see Figure 5a). Observations with a ratio $< 0.5$, where the distance from the origin postcode to origin station is more than double the distance from the origin postcode to the destination station,

[2]The distribution of access and egress distance is skewed further to the right in the LATIS dataset. A review of some observations with access and egress legs of this magnitude, indicated that these could be valid trips. For example, someone travelling from the far recesses of the Highlands and Islands might choose to drive to Inverness to begin their rail journey.




(a) RV ratio. (b) BT ratio.

**Figure 5.** Illustrative example of ratios to identify illogical trips.

were removed. The second, the BT ratio, captures the 'back-track' effect and is the distance from the origin postcode to the destination postcode over the distance from the origin postcode to the destination station. The closer the ratio is to zero, the more pronounced the back-track effect becomes (see Figure 5b). Observations with a ratio $< 0.5$, where the distance from origin postcode to destination postcode is less than half the distance from origin postcode to destination station, were removed. To establish the effectiveness of the steps taken to remove illogical trips, 100 random observations were selected from the WG dataset and their reported trips were individually visualised in QGIS. All 100 of the trips were considered logical.

## 4. Modelling methodology

### 4.1. *Theoretical basis*

Given that discrete choice models (DCMs) are suitable for predicting choice between two or more alternatives, and they have been successfully applied in prior research in the station choice field, they were chosen for this study. DCMs are usually based on the theory of utility maximization. When making a choice from several alternatives (collectively called the choice set) an individual will choose the alternative that yields them maximum utility. The researcher does not know the utility of each alternative, but will attempt to measure it by identifying attributes of the alternatives and/or of the individual. The part of utility that the researcher is unable to measure is known as unobserved utility, and is treated as a random component. The utility that an individual obtains from an alternative can be expressed using the following formula:

$$U_{ni} = V_{ni} + \varepsilon_{ni} \tag{1}$$

where $U_{ni}$ is the utility for individual $n$ of alternative $i$, $V_{ni}$ is the utility measured by the researcher, and $\varepsilon_{ni}$ is the unobserved utility. In practice $V$, will be a function consisting of the selected attributes and their respective coefficients (or parameters). The function is commonly linear in parameters and the measured utility for individual

$n$ of alternative $i$ can be given by:

$$V_{ni}(\boldsymbol{X}, \beta) = \sum_{k=1}^{K} \beta_k X_{kni} \tag{2}$$

where $\boldsymbol{X}$ is a matrix of attributes and $\beta$ is a vector of parameters of those attributes. The parameters, if unknown, are obtained statistically, for example by maximum likelihood estimation. 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. Instead, the *probability* of an alternative being chosen is calculated. Assumptions made about the characteristics of the unobserved utility will determine what form of statistical model is appropriate to calculate the probabilities[3].

This section will continue by describing the procedure used to define the choice set of stations for each validated observation in this study. It will then introduce the explanatory variables that were selected as potential measures of utility and explain how these were derived. Finally, in model calibration, the process of estimating the parameter values of the variables is outlined.

### 4.2. *Defining the choice set of alternative stations*

Based on experience from an earlier pilot study (Young and Blainey 2016), a separate choice set was defined for each observation, consisting of the ten nearest stations by road distance. These choice sets accounted for 92% and 95% of observed choice in the LATIS and WG datasets respectively. It was decided to try and improve the choice sets by ensuring the nearest major station to each origin was included, and this increased the proportion of observed choice accounted for to 97% in both datasets.[4] Any observation where the chosen station was not present in the choice set was, by necessity, removed before model calibration. If an alternative origin station was the observed destination station it was removed from the choice set.[5] As it was planned to estimate mode-specific access time parameters some further adjustments to the choice sets were necessary. Observations where access mode was recorded as 'other' were removed, and where access was by bus (or Glasgow subway) alternatives were only retained if a route was available to the station using that mode (or the trip planner suggested walking).

### 4.3. *Deriving explanatory variables (measured attributes)*

To ensure that the most accurate values for the attributes used to measure utility were entered into the models, it was important to obtain realistic representations of station access journeys by different modes, and to identify the rail services available to each

[3]This introduction to discrete choice models, and the included notation, is based largely on Train (2009)

[4]The 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. For 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.

[5]In addition, if Glasgow Central or Glasgow Queen Street was the observed destination, then both these stations were removed from the choice set if present. Using either of these stations to get to the other would be illogical.

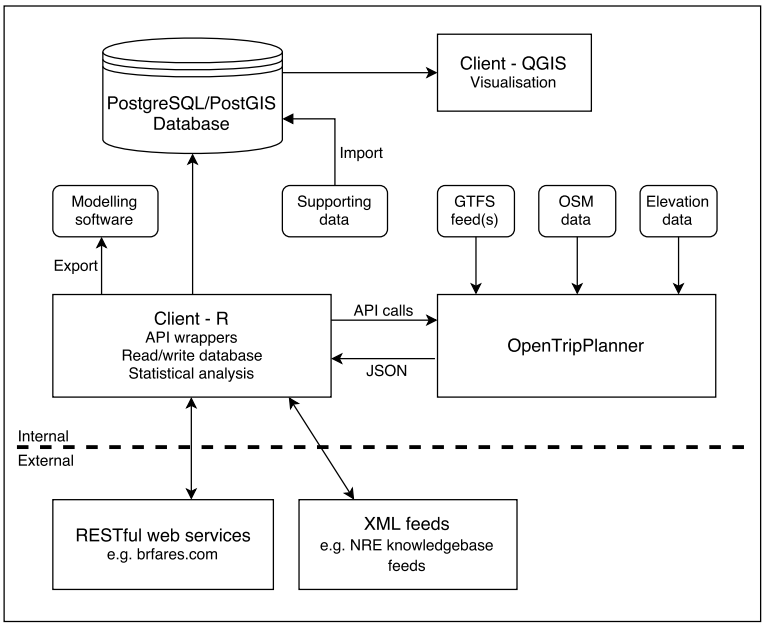



**Figure 6.** Framework to derive explanatory variables from disparate open transport data sources, from Young (2016).

respondent and the characteristics of alternative rail legs. This required development of a bespoke route planner that could generate routes for a range of motorised and non-motorised transport modes and incorporate relevant public transit schedules. To efficiently and accurately handle the collection and processing of large quantities of data from disparate open transport data sources, a data processing framework was developed that consisted of a PostgreSQL database, the R software environment, an instance of OpenTripPlanner (OTP) (an open-source route planner), and various external data sources (see Figure 6). The framework is described in more detail in Young (2016).

Socio-demographic attributes related to the respondents were not used as model variables. In DCMs attributes must vary across alternatives, and socio-demographic variables do not. For these variables to be used, either separate models would need to be calibrated for different segments (e.g. male and female), or the variables would need to be interacted (in some justifiable way) with other variables that *do* vary across alternatives (see Train (2009) for a detailed explanation). Also, as the models are to be used to define station catchments at the unit postcode level, any variables used would need to be available at this spatial resolution, which is not the case for UK census data. Furthermore, even if a variable such as level of car ownership was available at postcode level, it would be an average for the postcode and introduce the problem of ecological fallacy. In view of all these issues, it was decided to only include attributes of the alternatives, and the sections that follow explain how variables related to the station access journey; station facilities and services; and the train journey were derived.

### *4.3.1. 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. To generate journey data for access by bus (and subway

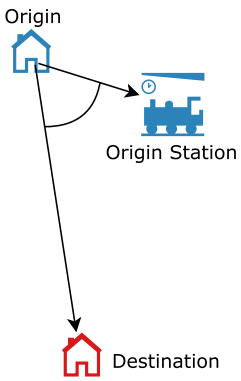



**Figure 7.** Difference in bearing (degrees) origin:origin station and origin:destination.

in Glasgow) the Scottish and Welsh components of the Traveline National Dataset (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 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 arrive by time was set to the recorded train time.[6] Two additional variables related to the access journey were generated. The 'nearest station' dummy variable indicates whether or not a station in an individual's choice set is the closest station by drive distance; and the 'bearing' variable gives the difference in bearing of origin:origin station and origin:destination in degrees (see Figure 7.).

### *4.3.2. Station facilities and service frequency*

Information on a range of potential facilities available at railway stations was obtained from the National Rail Enquiries (NRE) Stations XML feed, which forms part of the NRE Knowledgebase. This was queried for every station in the UK. The variables recorded included: free car park (y/n), car park spaces (number), station CCTV (y/n), ticket machine (y/n), waiting room (y/n), toilets (y/n), cycle spaces (number), cycle storage (y/n), cycle CCTV (y/n), and staffing level (unstaffed, part-time, full-time). To generate service frequencies the GTFS feed for GB rail services dated 25 April 2015 was downloaded from the TransitFeeds archive[7] and converted into a PostgreSQL database. A SQL query was then used to count the number of daily services and peak services (7am - 9am) scheduled to pick-up passengers at each station.

### *4.3.3. 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[8] to cover the survey period for both datasets. In addition, to allow London transfers, a GTFS feed for London Underground services was created from Transport for London

[6]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.

[7]See http://transitfeeds.com/p/association-of-train-operating-companies/284

[8]An OTP graph specifies every location in the region covered and how to travel between them. It is compiled from OpenStreetMap and GTFS data.

**Table 1.** Summary of datasets prepared for model calibration.

| | Number of choice situations | Number of cases | Average choice set size. |
|---|---|---|---|
| LATIS | 9367 | 97838 | 10.44 |
| WG | 5680 | 59833 | 10.53 |

journey planner data. 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.[9] A minimum transfer time of 6 minutes was specified, corresponding to the typical suggested connection time for a medium interchange station. The desired trip start time was set to the recorded train time. The variables used in the choice models were the journey duration and its separate components, on-train time and waiting time. Fares data was obtained using the independent BR Fares web service API (BR Fares Ltd 2016). The fare variable was populated dependent on the recorded train time, generally the cheapest anytime return fare (train times before 9am), or the cheapest off-peak fare (train times after 9am).

### 4.4. *Model calibration*

A summary of the datasets following data checking, trip validation and the preparation of choice sets is shown in Table 1. A series of models were calibrated separately for the WG and LATIS datasets using the NLOGIT 5 software package (Econometric Software Inc 2012). Separate model variants were calibrated suitable for subsequent application in either trip end models or flow models, with the latter additionally incorporating variables relating to the train leg and destination. The dependent variable in DCMs is a Boolean variable which takes the value 1 (or Y) for the single alternative in the choice set that was chosen, and 0 (or N) for all other alternatives. The explanatory variables were entered into the models using a manual forward selection procedure, and the utility function was defined as linear in parameters for all models.

## 5. Results and discussion

A summary of results for key models are shown in Tables 2, 3 and 4. In addition to measures of model fit, these tables include a measure of model predictive performance, with a lower value indicating a better model (this measure is discussed further in Section 5.3.1). The estimated parameter values ($B$) shown in the tables represent the positive or negative weighting applied to the variable in the utility function (Equation 2).

[9]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 much nearer the ultimate destination.

**Table 2.** Results of railway station choice MNL models - LATIS.

| | TE1 | | | TE7 | | | TE8 | | | TE9 | | | TE10 | | | TE17 | | | FM1 | | | FM2 | | | FM3 | | | FM4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* |
| Nearest station (yes) | 2.81 | 0.02 | *** | 0.86 | 0.03 | *** | 0.87 | 0.04 | *** | 1.02 | 0.04 | *** | 0.86 | 0.04 | *** | 0.84 | 0.04 | *** | 0.70 | 0.04 | *** | 0.70 | 0.04 | *** | 0.73 | 0.05 | *** | 0.72 | 0.05 | *** |
| Access time - walk (mins) | | | | -0.11 | 0.00 | *** | -0.11 | 0.00 | *** | -0.10 | 0.00 | *** | -0.11 | 0.00 | *** | -0.11 | 0.00 | *** | -0.13 | 0.00 | *** | -0.12 | 0.00 | *** | -0.12 | 0.00 | *** | -0.12 | 0.00 | *** |
| Access time - cycle (mins) | | | | -0.08 | 0.01 | *** | -0.09 | 0.01 | *** | -0.07 | 0.01 | *** | -0.09 | 0.01 | *** | -0.09 | 0.01 | *** | -0.13 | 0.01 | *** | -0.12 | 0.01 | *** | -0.12 | 0.01 | *** | -0.11 | 0.01 | *** |
| Access time - PT (mins) | | | | -0.05 | 0.00 | *** | -0.04 | 0.00 | *** | -0.05 | 0.00 | *** | -0.05 | 0.00 | *** | -0.05 | 0.00 | *** | -0.10 | 0.00 | *** | -0.08 | 0.00 | *** | -0.08 | 0.00 | *** | -0.07 | 0.00 | *** |
| Access time - car (mins) | | | | -0.12 | 0.00 | *** | -0.14 | 0.00 | *** | -0.13 | 0.00 | *** | -0.15 | 0.00 | *** | -0.16 | 0.00 | *** | -0.28 | 0.01 | *** | -0.24 | 0.01 | *** | -0.24 | 0.01 | *** | -0.23 | 0.01 | *** |
| Full-time staff (yes)[a] | | | | | | | 4.40 | 0.07 | *** | | | | 3.69 | 0.07 | *** | 2.73 | 0.10 | *** | 2.32 | 0.11 | *** | 2.13 | 0.11 | *** | 2.31 | 0.11 | *** | 2.33 | 0.11 | *** |
| Part-time staff (yes)[a] | | | | | | | 1.93 | 0.05 | *** | | | | 1.74 | 0.05 | *** | 1.09 | 0.07 | *** | 0.88 | 0.09 | *** | 0.87 | 0.08 | *** | 0.90 | 0.08 | *** | 0.91 | 0.08 | *** |
| Train frequency (# daily) | | | | | | | | | | 0.01 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** |
| CCTV (yes) | | | | | | | | | | | | | | | | 2.82 | 0.40 | *** | 2.84 | 0.40 | *** | 2.51 | 0.40 | *** | 2.53 | 0.41 | *** | 2.44 | 0.41 | *** |
| Car park spaces (#) | | | | | | | | | | | | | | | | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | *** |
| Free car park (yes) | | | | | | | | | | | | | | | | 0.74 | 0.36 | ** | 0.85 | 0.39 | ** | 0.67 | 0.37 | * | 0.74 | 0.38 | * | 0.84 | 0.38 | ** |
| Ticket machine (yes) | | | | | | | | | | | | | | | | 0.89 | 0.07 | *** | 0.61 | 0.08 | *** | 0.64 | 0.08 | *** | 0.63 | 0.08 | *** | 0.63 | 0.08 | *** |
| Toilets (yes) | | | | | | | | | | | | | | | | 0.56 | 0.07 | *** | 0.44 | 0.08 | *** | 0.41 | 0.08 | *** | 0.44 | 0.08 | *** | 0.47 | 0.08 | *** |
| Train leg time (mins) | | | | | | | | | | | | | | | | | | | -0.13 | 0.00 | *** | | | | | | | | | |
| On train time (mins) | | | | | | | | | | | | | | | | | | | | | | -0.09 | 0.00 | *** | -0.08 | 0.00 | *** | -0.08 | 0.00 | *** |
| Waiting time (mins) | | | | | | | | | | | | | | | | | | | | | | -0.15 | 0.00 | *** | -0.15 | 0.00 | *** | -0.15 | 0.00 | *** |
| Bearing difference (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | -0.00 | 0.00 | *** | | | |
| Bearing diff. <5km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | -0.00 | 0.00 | *** |
| Bearing diff. 5-10km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | -0.01 | 0.00 | *** |
| Bearing diff. 10-15km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | -0.01 | 0.00 | *** |
| Bearing diff. 15-20km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | -0.01 | 0.00 | *** |
| Bearing diff. 20+ km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | -0.02 | 0.00 | *** |
| Sample size (# trips) | 9367 | | | 9367 | | | 9367 | | | 9367 | | | 9367 | | | 9367 | | | 9367 | | | 9367 | | | 9367 | | | 9367 | | |
| Initial log-likelihood[b] | -21945 | | | -21945 | | | -21945 | | | -21945 | | | -21945 | | | -21945 | | | -21945 | | | -21945 | | | -21945 | | | -21945 | | |
| Final log-likelihood | -13751 | | | -10785 | | | -7348 | | | -8593 | | | -7093 | | | -6764 | | | -4803 | | | -5243 | | | -5184 | | | -5159 | | |
| McFadden's adjusted R2 | 0.37 | | | 0.51 | | | 0.66 | | | 0.61 | | | 0.68 | | | 0.69 | | | 0.78 | | | 0.76 | | | 0.76 | | | 0.76 | | |
| Predictive perf. diff. (%) | 72.0 | | | 62.0 | | | 30.0 | | | 47.3 | | | 28.1 | | | 23.5 | | | 14.4 | | | 14.8 | | | 15.1 | | | 14.6 | | |

[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 3.** Results of railway station choice MNL models - WG.

| | TE1 | | | TE7 | | | TE8 | | | TE9 | | | TE10 | | | TE17 | | | FM1 | | | FM2 | | | FM3 | | | FM4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* | *B* | *SE* | *Sig* |
| Nearest station (yes) | 3.13 | 0.03 | *** | 1.06 | 0.05 | *** | 0.95 | 0.05 | *** | 1.05 | 0.05 | *** | 0.95 | 0.05 | *** | 0.90 | 0.05 | *** | 0.88 | 0.06 | *** | 0.89 | 0.06 | *** | 0.87 | 0.06 | *** | 0.88 | 0.06 | *** |
| Access time - walk (mins) | | | | -0.11 | 0.00 | *** | -0.14 | 0.00 | *** | -0.14 | 0.00 | *** | -0.14 | 0.00 | *** | -0.13 | 0.00 | *** | -0.13 | 0.00 | *** | -0.13 | 0.00 | *** | -0.13 | 0.00 | *** | -0.13 | 0.00 | *** |
| Access time - cycle (mins) | | | | -0.14 | 0.02 | *** | -0.14 | 0.02 | *** | -0.15 | 0.02 | *** | -0.14 | 0.02 | *** | -0.14 | 0.02 | *** | -0.16 | 0.02 | *** | -0.16 | 0.02 | *** | -0.16 | 0.02 | *** | -0.16 | 0.02 | *** |
| Access time - PT (mins) | | | | -0.05 | 0.00 | *** | -0.04 | 0.00 | *** | -0.05 | 0.00 | *** | -0.04 | 0.00 | *** | -0.05 | 0.00 | *** | -0.06 | 0.00 | *** | -0.06 | 0.00 | *** | -0.06 | 0.00 | *** | -0.06 | 0.00 | *** |
| Access time - car (mins) | | | | -0.14 | 0.01 | *** | -0.15 | 0.01 | *** | -0.16 | 0.01 | *** | -0.15 | 0.01 | *** | -0.18 | 0.01 | *** | -0.24 | 0.01 | *** | -0.23 | 0.01 | *** | -0.23 | 0.01 | *** | -0.22 | 0.01 | *** |
| Full-time staff (yes)[a] | | | | | | | 3.22 | 0.07 | *** | | | | 2.36 | 0.14 | *** | 2.05 | 0.16 | *** | 1.97 | 0.18 | *** | 2.00 | 0.18 | *** | 1.96 | 0.18 | *** | 1.99 | 0.18 | *** |
| Part-time staff (yes)[a] | | | | | | | 2.08 | 0.06 | *** | | | | 1.94 | 0.06 | *** | 1.56 | 0.07 | *** | 1.36 | 0.07 | *** | 1.37 | 0.07 | *** | 1.38 | 0.07 | *** | 1.39 | 0.07 | *** |
| Train frequency (# daily) | | | | | | | | | | 0.01 | 0.00 | *** | 0.00 | 0.00 | *** | 0.00 | 0.00 | ** | -0.00 | 0.00 | *** | -0.00 | 0.00 | *** | -0.00 | 0.00 | *** | -0.00 | 0.00 | *** |
| CCTV (yes) | | | | | | | | | | | | | | | | 1.11 | 0.14 | *** | 1.16 | 0.14 | *** | 1.14 | 0.14 | *** | 1.14 | 0.14 | *** | 1.12 | 0.14 | *** |
| Car park spaces (#) | | | | | | | | | | | | | | | | 0.01 | 0.00 | *** | 0.01 | 0.00 | *** | 0.01 | 0.00 | *** | 0.01 | 0.00 | *** | 0.01 | 0.00 | *** |
| Free car park (yes) | | | | | | | | | | | | | | | | 0.36 | 0.10 | *** | 0.67 | 0.11 | *** | 0.66 | 0.11 | *** | 0.65 | 0.11 | *** | 0.64 | 0.11 | *** |
| Ticket machine (yes) | | | | | | | | | | | | | | | | 1.00 | 0.09 | *** | 0.83 | 0.10 | *** | 0.85 | 0.10 | *** | 0.88 | 0.10 | *** | 0.89 | 0.10 | *** |
| Toilets (yes) | | | | | | | | | | | | | | | | 0.28 | 0.09 | *** | 0.38 | 0.10 | *** | 0.38 | 0.10 | *** | 0.31 | 0.10 | *** | 0.32 | 0.10 | *** |
| Train leg time (mins) | | | | | | | | | | | | | | | | | | | -0.07 | 0.00 | *** | | | | | | | | | |
| On train time (mins) | | | | | | | | | | | | | | | | | | | | | | -0.07 | 0.00 | *** | -0.07 | 0.00 | *** | -0.07 | 0.00 | *** |
| Waiting time (mins) | | | | | | | | | | | | | | | | | | | | | | -0.08 | 0.00 | *** | -0.08 | 0.00 | *** | -0.08 | 0.00 | *** |
| Bearing difference (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | 0.00 | 0.00 | *** | | | |
| Bearing diff. <5km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.00 | 0.00 | *** |
| Bearing diff. 5-10km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0.00 | 0.00 | ** |
| Bearing diff. 10-15km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | -0.00 | 0.00 | |
| Bearing diff. 15-20km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | -0.01 | 0.00 | *** |
| Bearing diff. 20+ km (deg.) | | | | | | | | | | | | | | | | | | | | | | | | | | | | -0.00 | 0.00 | |
| Sample size (# trips) | 5680 | | | 5680 | | | 5680 | | | 5680 | | | 5680 | | | 5680 | | | 5680 | | | 5680 | | | 5680 | | | 5680 | | |
| Initial log-likelihood[b] | -13355 | | | -13355 | | | -13355 | | | -13355 | | | -13355 | | | -13355 | | | -13355 | | | -13355 | | | -13355 | | | -13355 | | |
| Final log-likelihood | -7215 | | | -5627 | | | -4068 | | | -4618 | | | -4043 | | | -3733 | | | -3249 | | | -3247 | | | -3236 | | | -3226 | | |
| McFadden's adjusted R2 | 0.46 | | | 0.58 | | | 0.69 | | | 0.65 | | | 0.70 | | | 0.72 | | | 0.76 | | | 0.76 | | | 0.76 | | | 0.76 | | |
| Predictive perf. diff. (%) | 64.0 | | | 56.5 | | | 34.7 | | | 42.9 | | | 33.3 | | | 27.4 | | | 22.7 | | | 22.6 | | | 22.4 | | | 22.3 | | |

[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.

### 5.1. *MNL models*

The key assumptions that underlie the MNL model are that the unobserved (stochastic) components of utility of the alternatives are independent of each other and have an identical (Gumbel) distribution (IIGD). It is a closed form model and the probability of choosing alternative $i$ from a choice set of $J$ alternatives is calculated using the following equation:

$$Pr_{(i)} = \frac{e^{V_i}}{\sum_{j=1}^{J} e^{V_j}} \tag{3}$$

where $V$ is the measured utility. As an example, the utility function for model TE7 is defined as follows:

$$V_{ik} = (\alpha * Dnearest) + \sum_{m=1}^{M} \beta_m (Dmode_m * AT_k) \tag{4}$$

where $Dnearest$ is a dummy variable with value 1 if alternative $k$ is the nearest station, and zero otherwise; $\alpha$ is the parameter for nearest station; $Dmode_m$ is a dummy variable with value 1 if individual $i$ uses access mode $m$, and zero otherwise; $AT_k$ is access time for alternative $k$; and $\beta_m$ is the access time parameter for mode $m$.

#### *5.1.1. Trip end variant models*

In the first model (TE1), the nearest station dummy variable is added. As would be expected, given that in 60-70% of the choice situations the nearest station was chosen, this model is a considerable improvement over the null model[10] for both datasets. The WG model performs rather better than the LATIS model, presumably reflecting the larger proportion of choice situations where the nearest station was chosen.

The next stage of calibration concentrated on identifying which access journey measure produced the best performing model, with both distance and time-based variables tested. In addition to estimating a single parameter, which represents only the 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 (See Equation 4). 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 adjusted $R^2$ of .58 and .51 respectively, incorporated mode-specific parameters for access time (Model TE7). The 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 4.2 units for a car driver, but by only 1.5 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.

[10]The null model assumes that there is an equal probability of each alternative in a choice set being chosen.

The station staffing level dummy variables (part-time and full-time) are added to the models next, with unstaffed excluded as the reference. The utility of a station is higher for staffed stations than unstaffed stations, and the models are substantially improved, particularly on the predictive performance measure (Model TE8). 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 is highly correlated with daily service frequency (LATIS: 0.72; WG: 0.86). In model TE9 staffing level is replaced with daily service frequency, but it is a far inferior model, indicating that staffing level is capturing additional information. Model TE10, which includes both staffing level and daily frequency, is an improvement over models TE8 and TE9, and the effect of the correlation between daily frequency and full-time staffing can be seen in lower parameter estimates for these variables.

In subsequent models several station facilities variables are introduced which result in relatively small improvements to the adjusted $R^2$, although there is a distinct improvement in model predictive performance between models TE10 and TE17. It is also noticeable, especially in the LATIS models, that the staffing level parameters are smaller once the station facilities variables are added, although they remain large suggesting that staffing level is an important factor in and of itself. Model TE17 is the best model suitable for integrating into trip end rail demand models, with an adjusted $R^2$ value of 0.69 and 0.72 for the LATIS and WG datasets respectively.

### *5.1.2. Flow variant models*

In the first of the models suitable for integration into flow-based rail demand models, the length of the train-leg (in minutes) is introduced (Model FM1). This is an improvement over model TE17, especially for the LATIS dataset where there is a substantial uplift in predictive performance. An effect of introducing the train-leg variable is an increase in the size of the mode-specific access time parameters, especially for car mode (from -.16 to -.28 and -.18 to -.24 for LATIS and WG models respectively). This may be the result of the prior models being unable to adequately explain longer access journeys to a chosen station. If decisions to travel further by car to 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 is split into on-train time and wait-time (due to transfers). In the LATIS model the wait-time parameter is 1.6 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 (ATOC 2013), although this is not replicated in the WG model where the wait-time parameter is only 1.2 times larger. There is a potential problem with the datasets that may impact the estimation of train-leg parameters. Respondents were asked 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 ultimate origin and destination stations were correctly identified it was 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. The LATIS FM2 model performs somewhat worse than the FM1 model on all the measures, whilst there is no

significant difference between the two WG models. However, it was felt that a model with separate parameters for on-train time and wait-time would be more transferable and FM2 was used as the basis for subsequent models.

The 'difference in bearing' variable is added to model FM3. In the LATIS model this has 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 for the WG model. It was hypothesised that this variable may become more important as the access journey distance increases, and might be of little consequence for short access journeys. This was investigated in model FM4 by estimating five separate parameters for the variable based on banded access time. In the LATIS model the parameters show the expected effect with a gradual increase in the size of the negative parameter as access distance increases. The effect of a 25-degree difference in bearing ranges from -0.1 for access journeys <5km to -0.5 for access journeys >20km. In the WG model the parameters for the three longer access bands have the expected negative sign, but only the parameter for the 15-20km band is significant. It is possible that the geography of the South Wales valleys is affecting 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.

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. LATIS Model FM4, with an adjusted $R^2$ value of 0.76, appears to be the most promising model to integrate into flow-based rail demand models.

### 5.2. *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 (RPL) 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:

$$P_{ni} = \int \left( \frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}} \right) f(\beta) d\beta \tag{5}$$

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

Initial RPL models were run, using the best performing MNL models as the starting

**Table 4.** Results of railway station choice RPL models (LATIS and WG).

| | WG trip end variant (RPL2) | | | | | | | | | LATIS trip end variant (RPL2) | | | | | | | | | WG flow variant (RPL4) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random parameters[a] | | | | | | | Non-random parameters | | Random parameters[a] | | | | | | | Non-random parameters | | Random parameters[a] | | | | | | | Non-random parameters | |
| *Variable* | *Mean ln(B)* | *Sig* | *Std. dev ln(B)* | *Sig* | *Med. B* | *Mean B* | *Std. Dev B* | *B* | *Sig* | *Mean ln(B)* | *Sig* | *Std. dev ln(B)* | *Sig* | *Med. B* | *Mean B* | *Std. Dev B* | *B* | *Sig* | *Mean ln(B)* | *Sig* | *Std. dev ln(B)* | *Sig* | *Med. B* | *Mean B* | *Std. Dev B* | *B* | *Sig* |
| Nearest station (yes) | | | | | | | | 0.54 | *** | | | | | | | | 0.46 | *** | | | | | | | | 0.65 | *** |
| Time - walk (mins) | -1.56 | *** | 0.87 | *** | 0.21 | 0.31 | 0.32 | | | -1.93 | *** | 0.44 | *** | 0.14 | 0.16 | 0.07 | | | -1.60 | *** | 0.62 | *** | 0.20 | 0.25 | 0.17 | | |
| Time - cycle (mins) | | | | | | | | -0.17 | *** | -2.26 | *** | 0.38 | *** | 0.10 | 0.11 | 0.04 | | | | | | | | | | -0.22 | *** |
| Time bus/PT (mins) | -2.80 | *** | 0.67 | *** | 0.06 | 0.08 | 0.06 | | | -2.66 | *** | 0.85 | *** | 0.07 | 0.10 | 0.10 | | | -2.48 | *** | 0.57 | *** | 0.08 | 0.10 | 0.06 | | |
| Time (car) mins | -1.41 | *** | 0.56 | *** | 0.24 | 0.29 | 0.17 | | | -1.39 | *** | 0.83 | *** | 0.25 | 0.35 | 0.35 | | | -1.20 | *** | 0.33 | *** | 0.30 | 0.32 | 0.11 | | |
| On train time (mins) | | | | | | | | | | | | | | | | | | | -2.61 | *** | 0.10 | ns | 0.07 | 0.07 | 0.01 | | |
| Wait time (mins) | | | | | | | | | | | | | | | | | | | -2.45 | *** | 1.02 | *** | 0.09 | 0.14 | 0.20 | | |
| Full-time staff (yes)[b] | | | | | | | | 2.30 | *** | | | | | | | | 2.92 | *** | | | | | | | | 1.78 | *** |
| Part-time staff (yes)[b] | | | | | | | | 1.80 | *** | | | | | | | | 1.15 | *** | | | | | | | | 1.46 | *** |
| Train frequency (# daily) | | | | | | | | 0.00 | *** | | | | | | | | 0.00 | *** | | | | | | | | | |
| CCTV (yes) | | | | | | | | 0.92 | *** | | | | | | | | 2.85 | *** | | | | | | | | 1.28 | *** |
| Car park spaces (#) | | | | | | | | 0.01 | *** | | | | | | | | 0.00 | *** | | | | | | | | 0.01 | *** |
| Free car park (yes) | | | | | | | | 0.52 | *** | | | | | | | | 0.73 | ns | | | | | | | | 0.77 | *** |
| Ticket machine (yes) | | | | | | | | 1.19 | *** | | | | | | | | 1.04 | *** | | | | | | | | 1.01 | *** |
| Toilets (yes) | | | | | | | | 0.08 | ns | | | | | | | | 0.73 | *** | | | | | | | | 0.24 | ** |
| Sample size (# trips) | 5680.00 | | | | | | | | | 9366.00 | | | | | | | | | 5680.00 | | | | | | | | |
| Initial log-likelihood[c] | -13355 | | | | | | | | | -21945 | | | | | | | | | -13355 | | | | | | | | |
| Final log-likelihood | -3649 | | | | | | | | | -6553 | | | | | | | | | -3149 | | | | | | | | |
| McFadden's adjusted $R^2$ | 0.73 | | | | | | | | | 0.70 | | | | | | | | | 0.76 | | | | | | | | |
| Predictive perf. diff. (%) | 25.90 | | | | | | | | | 23.60 | | | | | | | | | 21.10 | | | | | | | | |

[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.

point, with all parameters specified as random to test whether the standard deviation (SD) of each 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 each of the model variables is 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.

#### *5.2.1. Trip end variant models*

Using Model TE17 as the initial model, with all parameters specified as random, the SD of the mode-specific access time parameters was significant for both WG and LATIS (excluding cycle mode in the WG model). In addition, the SD of the nearest station and car park spaces parameters was significant in the LATIS model. The z-values for the other parameters were very low, and not close to critical values. Based on these findings an RPL model with the access time parameters specified as random was run for both datasets, and the results are shown in Table 4 (Model RPL2). Both the LATIS and WG models have higher log-likelihood and adjusted $R^2$ values than the MNL equivalent model, and although predictive performance is slightly better for the WG model, it is marginally worse for the LATIS model. The SD of all the random parameters is significant, indicating that the parameter estimates are individual-specific and for any individual the parameter may be different from the mean parameter estimate (Hensher, Rose, and Greene 2016). Interestingly, the variability in the parameter for walk access time is much greater in the WG model (SD 0.32) than it is in the LATIS model (SD 0.07), while there is greater variability in the parameter for car access time in the LATIS model (SD 0.35) compared with the WG model (SD 0.17). The RPL model also has an effect on the non-random parameters, compared with the MNL model, most noticeably a substantially smaller parameter for the nearest station variable.

#### *5.2.2. Flow variant models*

Using Model FM2 as the initial model, with all parameters specified as random, the SD of the access time parameters was again significant, apart from cycle mode for both the LATIS and WG models. The SD of on-train time and waiting time was significant in the LATIS model, while the SD of waiting time, nearest station and CCTV was significant in the WG model. Based on these findings an RPL model with the access time and train leg parameters specified as random was run for both datasets. The LATIS model failed to fit as initial iterations were unable to improve the log-likelihood function of the MNL model that was used for starting values. In the WG model (RPL4) the SD is significant for access time and wait time, but not significant for on-train time, as found in the initial model. Model RPL4 performs somewhat better than the equivalent MNL model (FM2), with a higher log-likelihood and slightly improved predictive performance.

### 5.3. *Model appraisal*

#### *5.3.1. Predictive performance*

Rather than use the fundamentally flawed '% correctly predicted' measure (see Train (2009) p.69 for a discussion), which assesses a model by assuming each individual

**Table 5.** Summary of model predictive performance.

| | Predictive performance (absolute difference as % of total choice situations) | | |
|---|---|---|---|
| Model | LATIS | WG | Comments |
| Base model (probability of nearest station = 1) | 50.9 | 41.0 | |
| TE17 | 23.5 | 27.4 | |
| RPL2 | 23.6 | 25.9 | Trip end variant |
| FM2 | 14.8 | 22.6 | |
| FM4 | 14.6 | na[a] | |
| RPL4 | na[b] | 21.1 | Flow variant |
| FM2 (using parameter values from alternate dataset) | 20.2 | 34.8 | Transferability test |

[a]Invalid model for the WG dataset.
[b]Model failed to fit.

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, Rose, and Greene (2016) p.502). To assess the overall performance of the models, the absolute difference between the two figures has been summed for all stations and expressed as a percentage of the total number of choice situations in the model. A predictive performance of zero percent would indicate no deviation between observed and predicted choice. The predictive performance of each model is included in Tables 2, 3 and 4. Table 5 summarises the performance of the best models and, given that the aim of this work is to improve on the simplistic models that assume the nearest station is chosen, compares them with a base model where the probability of choosing the nearest station equals 1. The graphs in Figures 8 and 9 show the number of times each station was actually chosen and by how much the model under or over-predicted this choice, for the base model and LATIS FM4, illustrating the substantially better predictive performance of the latter.

### *5.3.2. Transferability*

One of the 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 10 plots the parameter estimates along with confidence intervals for model FM2[11] for both case study areas. It suggests reasonable correspondence of many of the parameters, but also identifies potentially problematic variables, such as 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 (88% of unique alternatives in the LATIS dataset), and it may only be a factor that actually influences choice for a

[11]Model FM2 was selected for this exercise as WG Model FM4 is invalid due to issues with the 'difference in bearing' variable discussed in Section 5.1.2.

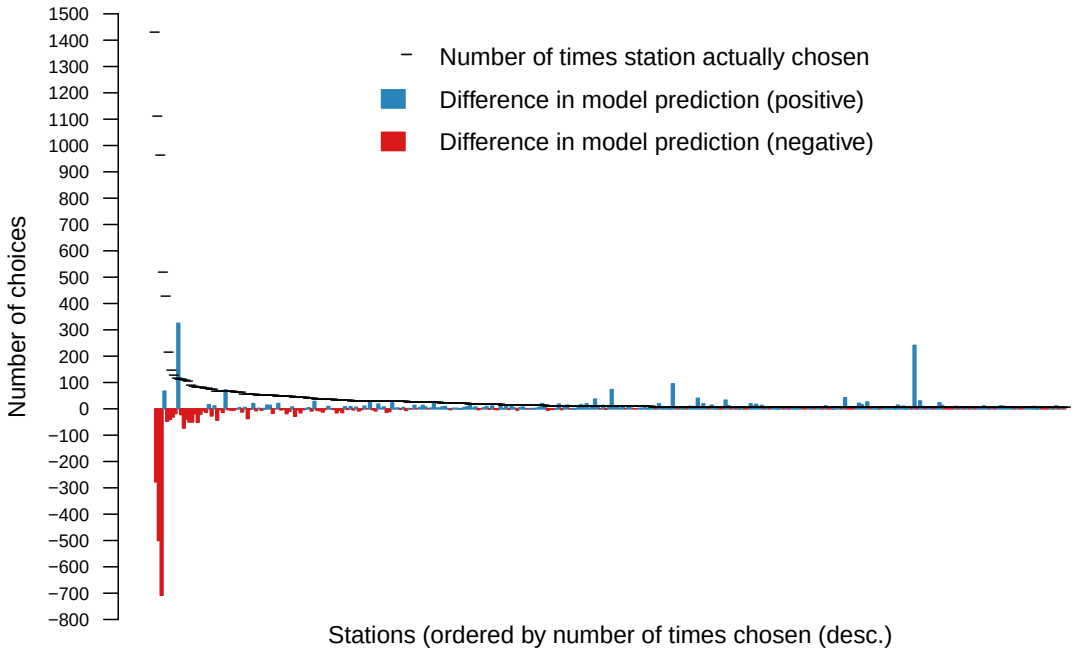



**Figure 8.** Model predictive performance - LATIS base model (nearest station probability = 1).

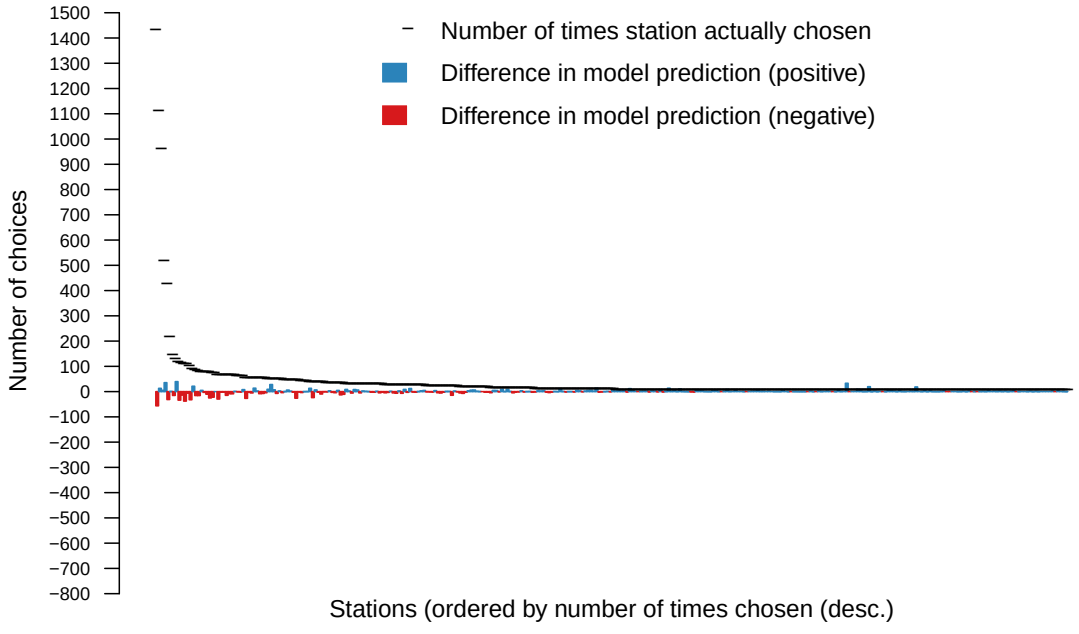



**Figure 9.** Model predictive performance - LATIS model FM4.

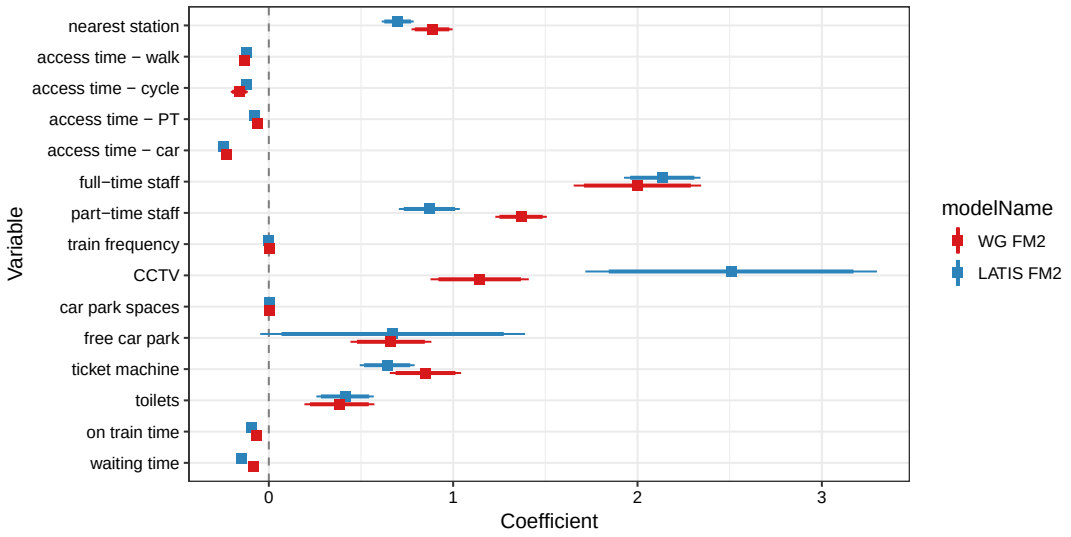



**Figure 10.** Parameter estimates for WG and LATIS model FM2 showing 95% and 99% confidence intervals.

small number of observations. In the next step to assess model transferability, the parameters from the LATIS FM2 model were used to predict choice in the WG dataset, and vice versa. The predictive performance of these models when applied to the alternative dataset are reported in Table 5. The WG model performs reasonably well against the LATIS dataset, slightly better than TE17 but not as good as FM4. However, the LATIS model does not perform particular well against the WG dataset, it is an improvement over the base model but not as good as TE17.

## 6. Conclusions and future work

This paper has shown that it is possible to calibrate station choice models, using two independent datasets, that are suitable for integration into both trip-end and flow rail demand models. The models have a very good fit as measured by adjusted $R^2$ and predict station choice substantially better than the base model that assumes the nearest station has a probability of one. There is good coincidence in parameter estimates for many of the explanatory variables across the two datasets, suggesting that calibration of a transferable model may be possible. Transferability has been tested by applying a WG calibrated model to the LATIS dataset and vice versa, with somewhat mixed results. Further work is needed to identify if problematic variables are having an adverse effect on model transferability, and to review poor predictive performance at the level of individual or neighbouring stations with a view to identifying shortcomings of the models that can be addressed. There is also scope to introduce additional explanatory variables, for example related to land-use characteristics; or to calibrate separate models based on specific socio-demographic characteristics, such as trip purpose or car ownership. While the latter might identify different attribute weightings dependent upon the demographic profile of an individual, the challenge will be in obtaining suitable data at sufficient spatial resolution to use these models for forecasting purposes.

The RPL models do not offer sufficient improvement over the MNL models to justify the extra complexity that would be involved in simulating station probabilities at

the postcode level. However, an important issue that remains to be addressed is the proportional substitution behaviour that is a characteristic of the MNL model. This means that the addition of a new station to a choice set will reduce the probability of each existing station in the choice set by the same percentage. However, it would be expected that a new station would, all else being equal, abstract more passengers from nearer stations than more distant ones. Although largely ignored in prior station choice research, there are several potential solutions to account for this spatial correlation, such as the inclusion of an accessibility term (for example, see Ho and Hensher (2016)); a bespoke generalised extreme value (GEV) model, such as the Generalised Spatially Correlated Logit (GSCL) model developed by Sener, Pendyala, and Bhat (2011); or an alternative (error components) formulation of the mixed logit model to create correlations between alternatives that result in an appropriate substitution pattern. Further research is needed to assess the feasibility and suitability of these approaches.

These limitations aside, the superior predictive performance of the station choice models compared to the base model, suggests that this research has the potential to improve the models that are used to assess proposals for new railway stations and to enable better forecasting of the effects of changing service patterns. Future work will focus on developing a methodology for incorporating probabilistic catchments derived from the station choice models into the rail demand models. The accuracy of rail demand models using either deterministic or probabilistic catchments will then be compared, ideally under a real-world scenario. The results should provide a methodology suitable for incorporating into national passenger demand forecasting frameworks, such as the UK's Passenger Demand Forecasting Handbook (ATOC 2013). The methodology should also have broader transferability to other national contexts where modifications to rail networks and services are planned.

**Acknowledgements**

The authors wish to thank 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. This work was supported by the EPSRC under DTG Grant EP/M50662X/1. Map data, Code.Point and Code.Point Polygons ©Crown Copyright and Database Right 2017. Ordnance Survey (Digimap Licence). This work uses public sector information licensed under the Open Government Licence v3.0. An earlier version of this paper was presented by the corresponding author to the 49th Universities Transport Study Group Annual Conference hosted by Trinity College Dublin (Young 2017).

**Data access statement**

The passenger survey data were obtained from the Welsh Government and the Transport Scotland LATIS Service. These datasets are subject to licence restrictions and cannot be made openly available. The multiple datasets used to derive explanatory variables for model calibration, including data used in the trip planner, are available from a number of different sources. Full details are available in the documentation at: TBC. All underlying data for the figures, graphs and tables that appear in this paper are openly available from the University of Southampton data archive at: TBC

## References


Amadeus. 2013. *The Rail Journey to 2020: Facts, figures and trends that will define the future of European passenger rail traffic.*

Association of Train Operating Companies (ATOC). 2013. *Passenger Demand Forecasting Handbook v5.1.*

Blainey, Simon, and Preston, John. 2010. "Modelling local rail demand in South Wales." *Transportation Planning and Technology* 33: 55–73.

Blainey, Simon, and Samantha Evens. 2011. "Local station catchments: reconciling theory with reality." Paper presented at AET European Transport Conference, October.

BR Fares Ltd. 2016. *BR Fares.* [online] Available at: `http://www.brfares.com`

Econometric Software Inc. 2012. *Nlogit 5.*

Harata, N., and K. Ohta. 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, 1729–1740. Centre for Transportation Studies.

Hensher, David A., John M. Rose, and William H. Greene. 2016. *Applied Choice Analysis.* 2nd ed. Cambridge University Press.

Ho, Chinh Q., and David A. Hensher. 2016. "A workplace choice model accounting for spatial competition and agglomeration effects." *Journal of Transport Geography* 51: 193–203.

Kastrenakes, Cheryl Rosen. 1988. "Development of a Rail Station Choice Model for NJ Transit." *Transportation Research Record* 1162: 16–21.

Lythgoe, WF, and M Wardman. 2002. "Estimating Passenger Demand for Parkway Stations." Paper presented at AET European Transport Conference, September.

Lythgoe, WF, and M Wardman. 2004. "Modelling passenger demand for parkway rail stations." *Transportation* 31 (2): 125–151.

Office of Rail Regulation. 2017. *Passenger Rail Usage 2016-17 Q4 Statistical Release.* [online] Available at: `http://www.orr.gov.uk/__data/assets/pdf_file/0019/24832/passenger-rail-usage-2016-17-q4.pdf`

Railfuture. 2017. *Britain's growing railway.*

Sener, Ipek N, Ram M Pendyala, and Chandra R Bhat. 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.

Steer Davies Gleave. 2010. *Station Usage and Demand Forecasts for Newly Opened Railway Lines and Stations.* Final report prepared for Department for Transport.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation.* Cambridge University Press.

Wardman, M, and GA Whelan. 1999. "Using Geographical Information Systems to improve rail demand models." Final Report to Engineering and Physical Sciences Research Council.

Young, Marcus. 2016. "An automated framework to derive model variables from open transport data using R, PostgreSQL and OpenTripPlanner." Paper presented at 24th GIS Research UK Conference, March.

Young, Marcus. 2017. "Developing railway station choice models to improve rail industry demand models." Paper presented at 49th Annual UTSG Conference, Dublin, Ireland, January.

Young, Marcus, and Simon Blainey. 2016. "Defining probability-based rail station catchments for demand modelling." Paper presented at 48th Annual UTSG Conference, Bristol, GB, January.

Young, Marcus, and Simon Blainey. 2017. "Railway Station Choice Modelling: A Review of Methods and Evidence." *Transport Reviews.* Advance online publication. doi:10.1080/01441647.2017.1326537