Developing railway station choice models to improve rail industry demand models

Mr Marcus Young
PhD Student
Transportation Research Group, University of Southampton

Abstract
This paper describes the development of railway station choice models suitable for defining probabilistic station catchments for use in 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. Results of 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 performance.

1 Introduction

Rail travel in GB has grown considerably in recent years. This has been accompanied by expansion of the network, and many new stations have been built and closed stations reopened. 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. The models predominantly used in the UK to forecast demand for new railway stations require a catchment for the station to be defined first, so that model inputs, such as the population from which demand will be generated, can be specified. The common methods used to define catchments are simplistic, for example placing a distance-based circular buffer around a station, or dividing the area of interest into zones and assigning each zone to its nearest station. However, recent research shows that in reality station catchments are far more complex than this (for example, see Blainey & 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 a probability-based catchment was defined.

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 focussed 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 focussed on addressing specific local needs, such as the work of Harata & Ohta (1986) and Kastrenakes (1988). Wardman & Whelan (1999) attempted to incorporate probabilistically defined catchments into a flow model as part of a larger piece of work to improve rail demand models using Geographical Information Systems (GIS), but this was largely unsuccessful; and Lythgoe & 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 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 changes. 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 & Blainey, 2016a). For a comprehensive review of prior research in this area, see Young & Blainey (2016b).

2 Data considerations

2.1 Survey data

Data from a series of ultimate origin-destination (OD) surveys carried out in Wales and Scotland were obtained from the Welsh Government (WG) and Transport Scotland's LATIS service. The WG surveys were carried out in the spring of 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. Prior to subsequent processing and validation the WG and LATIS datasets contained some 7,000 and 50,000 observations respectively.

2.2 Data cleaning

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 particular types of trip, such as those beginning or ending at their home address, and 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's 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. The AddressBase file was imported into a PostgreSQL database 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 1. 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 appended to the dataset. A manual review process was then completed, 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, that postcode is accepted
   b) if street has more than one postcode, if the maximum Euclidean distance is <= 250m use the calculated street centroid as the origin or destination location

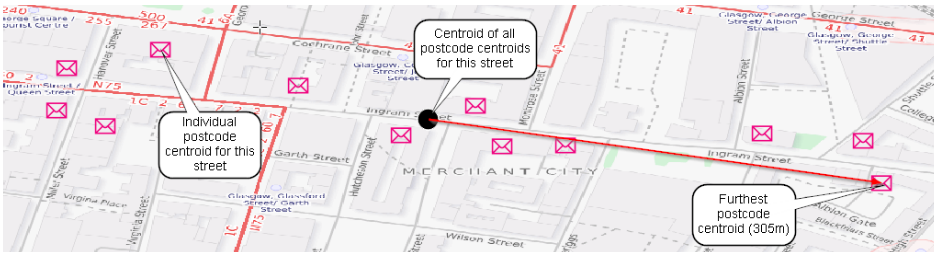

Centroid of all postcode centroids for this street
Individual postcode centroid for this street
Furthest postcode centroid (305m)
Cochrane Street
Ingram Street
Garth Street
Hutcheson Street
Montrose Street
Albion Street
George Street
Hanover Street
Miller Street
Virginia Place
Virginia Street
Wilson Street
MERCHANT CITY
Albion Gate
Blackfriars Street
Glasgow, Glassford Street/ Garth Street
Glasgow, George Street/ Albion Street
Glasgow, George Street/ Shuttle Street
Ingram Street / Queen Street

Figure 1: Postcode centroids for Ingram Street, Glasgow, showing calculated centroid and maximum distance from calculated centroid to any postcode centroid

A variety of other data checks were carried out on both the WG and LATIS 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 trip planner (see Section 2.4), 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.

2.3 Automated trip validation

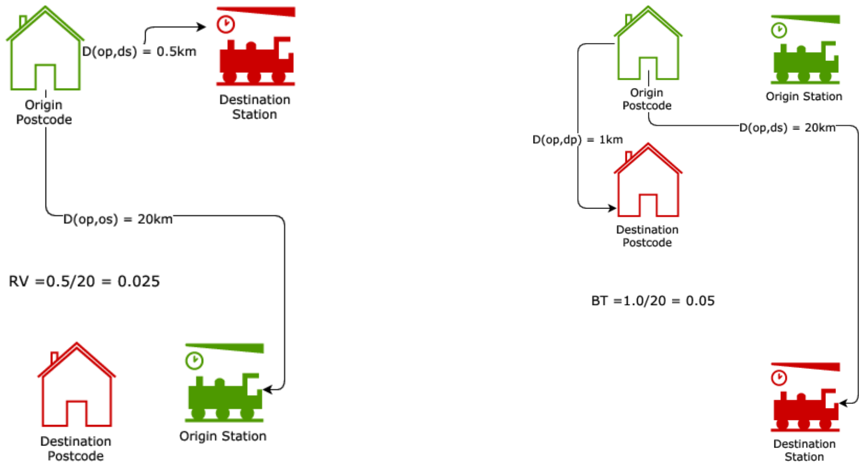
Due to the large number of survey observations in the datasets it was not practical to manually check each one to ensure the reported trip was sensible. An alternative strategy was adopted that generated information inherent in the reported trip and used that to automatically validate the trip. This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

2.3.1 Excessive access or egress legs

For each observation in the cleaned data, a trip planner 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 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[1].

2.3.2 Illogical trips

There are two main types of illogical trips that are observed in this type of data. The first is the ‘reversed trip’ where the origin station is located close to the ultimate destination, and the destination station is close to the ultimate 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 the ratio is to zero, the more pronounced the reversal effect becomes (see Figure 2a). 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, were removed.



(a) RV Ratio (b) BT Ratio

Figure 2: Illustrative example of ratios to identify illogical trips

[1] The distribution of access and egress distance is skewed further to the right in the LATIS dataset.

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 2b). 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. 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 excessive access and/or egress legs) and their reported trips were individually visualised in QGIS. All 100 of the trips were considered logical.

2.4 Deriving explanatory variables

A key objective of this research was to obtain realistic representations of station access journeys by different transport modes and to identify, as accurately as possible, the rail services available to each respondent when they made their trip and the characteristics of alternative rail legs, thus ensuring that appropriate explanatory variables for the station choice models could be generated. 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. It was also recognised that deriving the explanatory variables would require the collection and processing 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. A data processing framework was therefore 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. The framework is described in more detail in Young (2016).

2.4.1 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. To generate journey data for access by bus (and also subway 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.[2] 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 difference in bearing of origin:origin station and origin:destination in degrees (see Figure 3).

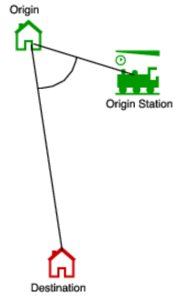


Figure 3: Difference in bearing (degrees) origin:origin station and origin:destination

[2] 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.

2.4.2 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[3] and converted into a PostgreSQL database. A SQL query was then used to count the number of daily services and peak services (7am - 9am) that pick-up passengers at each station.

2.4.3 Train journey

Two GTFS feeds for GB rail services dated 17 March 2014 and 4 April 2015 were downloaded from the TransitFeeds archive[4] and incorporated into separate OTP graphs[5] to cover the survey period for both the WG and LATIS datasets. In addition, to allow London transfers, a GTFS feed for London Underground services was created from Transport for London 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.[6] 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[7]. 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).

2.5 Defining choice sets

Based on experience from an earlier pilot study (Young & Blainey, 2016a), a separate choice set was defined for each observation, consisting of the ten nearest stations by road distance. For the WG and LATIS datasets a separate 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. For each origin:station pair the drive distance was obtained from an API call to OTP and the stations then ranked by drive 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.[8] 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. Including the nearest major station increased the proportion of observed choice accounted for to 97% in both datasets. Any observation where the chosen station was not present in the choice set was, by necessity, removed before model calibration. If an

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

[4] See Footnote 3

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

[6] 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.

[7] See Footnote 2

[8] 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.

alternative origin station was the observed destination station it was removed from the choice set.[9]

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 bus route was available to the station or OTP suggested walking to the station. For walk access mode, choices were not restricted to stations within 60 minutes of the origin as this would have resulted in some choice sets with only the chosen station and they could not have been included during model calibration. Using the same choice sets for all model calibration enabled direct comparison of model fit using log likelihood and R^2 measures.

3 Model calibration

A summary of the datasets following data cleaning, trip validation and preparation of choice sets is shown in Table 1, and Figure 4 disaggregates the choice situations by the distance-rank of the chosen station.

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

A series of models were calibrated separately for the WG and LATIS datasets using the NLOGIT 5 software package (Econometric Software, Inc, 2012). As this research was seeking to develop station choice models that could be incorporated into the rail demand models that are typically used to forecast demand for new stations, choice models suitable for use in trip end models were distinguished from those suitable for flow modes, with the latter additionally incorporating variables relating to the train leg and destination. Explanatory variables were entered into the models using a manual forward selection procedure, and a summary of results for key models are shown in Tables 2 and 3. In addition to reporting the model log-likelihood and McFadden's adjusted R-squared, these tables include a measure of the predictive performance of each model, with a lower value indicating a better model (this measure is discussed further in Section 4.1).

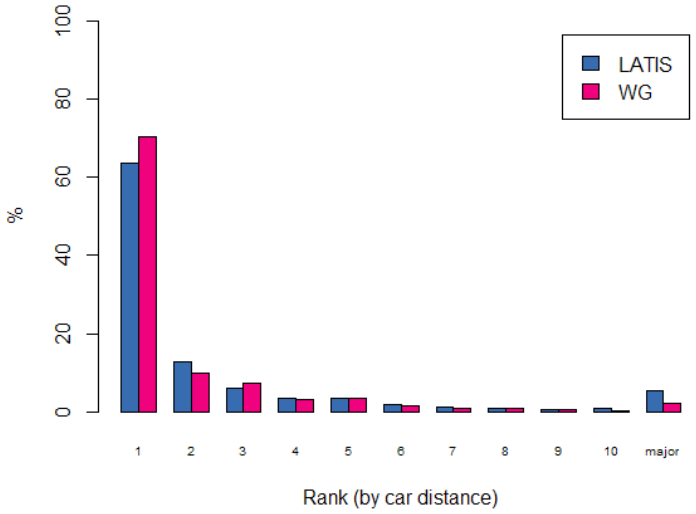


Figure 4: Percentage of choice situations where chosen station was of specified rank (based on car distance) or a major station not otherwise ranked 1:10

[9] 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. This is not the case for Cardiff or Edinburgh where travel between the two main stations by rail is a logical trip.

Table 2 – LATIS multinomial model results

	Nearest station	Time walk mins	Time cycle mins	Time pt mins	Time car mins	Full-time staff	Part-time staff	Train freq. daily	Cctv	Car park sp.	Free car park	Ticket mach.	Toilets	Train leg time mins	On train time	Wait time	Bear. diff	Bear. <5 km	Bear. 5-10 km	Bear. 10-15 km	Bear. 15-20 km	Bear. 20+ km	logLik	AdjR^2	Pred. perf. diff (%)
TE1	2.81***																						-13751	.37	72.0
TE7	.86***	-.11***	-.08***	-.05***	-.12***																		-10785	.51	62.0
TE8	.87***	-.11***	-.09***	-.04***	-.14***	4.40***	1.93***																-7348	.66	30.0
TE9	1.02***	-.10***	-.07***	-.05***	-.13***			.006***															-8593	.61	47.3
TE10	.86***	-.11***	-.09***	-.05***	-.15***	3.69***	1.74***	.003***															-7093	.68	28.1
TE13	.87***	-.11***	-.09***	-.05***	-.16***	3.57***	1.61***	.003***	3.27***														-7015	.68	27.4
TE14	.85***	-.11***	-.09***	-.05***	-.16***	3.45***	1.55***	.002***	3.27***	.001***													-6965	.68	27.5
TE17	.84***	-.11***	-.09***	-.05***	-.16***	2.73***	1.09***	.002***	2.82***	.001***	.74**	.89***	.56***										-6764	.69	23.5
FM1	.70***	-.13***	-.13***	-.10***	-.28***	2.32***	.88***	.001***	2.84***	.002***	.85**	.61***	.44***	-.13***									-4803	.78	14.4
FM2	.70***	-.12***	-.12***	-.08***	-.24***	2.13***	.87***	.001***	2.51***	.002***	.67*	.64***	.41***		-.09***	-.15***							-5243	.76	14.8
FM3	.73***	-.12***	-.12***	-.08***	-.24***	2.31***	.90***	.001***	2.53***	.002***	.74*	.63***	.44***		-.08***	-.15***	-.005***						-5184	.76	15.1
FM4	.72***	-.12***	-.11***	-.07***	-.23***	2.33***	.91***	.001***	2.44***	.002***	.84**	.63***	.47***		-.08***	-.15***		-.004***	-.007***	-.01***	-.01***	-.02***	-5159	.76	14.6

Table 3 – WG multinomial model results

	Nearest station.	Time walk mins	Time cycle mins	Time bus mins	Time car mins	Full-time staff	Part-time staff	Train freq. daily	Cctv	Car park sp.	Free car park	Ticket mach.	Toilets	Train leg time mins	On train time	Wait time	Bear. diff	Bear. <5 km	Bear. 5-10 km	Bear. 10-15 km	Bear. 15-20 km	Bear. 20+ km	logLik	AdjR^2	Pred. perf. diff (%)
TE1	3.13***																						-7215	.46	64.0
TE7	1.06***	-.11***	-.14***	-.05***	-.14***																		-5627	.58	56.5
TE8	.95***	-.14***	-.14***	-.04***	-.15***	3.22***	2.08***																-4068	.69	34.7
TE9	1.05***	-.14***	-.15***	-.05***	-.16***			.008***															-4618	.65	42.9
TE10	.95***	-.14***	-.14***	-.04***	-.15***	2.36***	1.94***	.003***															-4043	.70	33.3
TE13	.94***	-.14***	-.15***	-.04***	-.15***	2.19***	1.74***	.003***	1.38***														-3976	.70	32.1
TE14	.88***	-.13***	-.15***	-.04***	-.18***	2.38***	1.76***	.001***	1.29***	.006***													-3816	.71	27.8
TE17	.90***	-.13***	-.14***	-.05***	-.18***	2.05***	1.56***	.001**	1.11***	.006***	.36***	1.0***	.28***										-3733	.72	27.4
FM1	.88***	-.13***	-.16***	-.06***	-.24***	1.97***	1.36***	-.002***	1.16***	.006***	.67***	.83***	.38***	-.07***									-3249	.76	22.7
FM2	.89***	-.13***	-.16***	-.06***	-.23***	2.0***	1.37***	-.002***	1.14***	.006***	.66***	.85***	.38***		-.07***	-.08***							-3247	.76	22.6
FM3	.87***	-.13***	-.16***	-.06***	-.23***	1.96***	1.38***	-.002***	1.14***	.006***	.65***	.88***	.31***		-.07***	-.08***	.003***						-3236	.76	22.4
FM4	.88***	-.13***	-.16***	-.06***	-.22***	1.99***	1.39***	-.002***	1.12***	.006***	.64***	.89***	.32***		-.07***	-.08***		.003***	.002**	-.001^{ns}	-.008***	-.005^{ns}	-3226	.76	22.3

3.1 Multinomial logit (MNL) models

3.1.1 Trip end-related 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 for a variable, which represents only average effects 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. 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-squared 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. The station staffing level dummy variables (part-time and full-time) are added to the models next, and these have to be interpreted with reference to the excluded unstaffed level. The results show that 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 level 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 the subsequent models several station facilities variables are introduced[11] which result in relatively small improvements to the adjusted R-squared, although there is a distinct improvement in model predictive performance. It is also noticeable, especially in the LATIS models, that the parameters for the staffing level variables become smaller as 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-squared of 0.69 and 0.72 for the LATIS and WG datasets respectively.

3.1.2 Flow-related models

In the first of the models suitable for integration into rail demand flow 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

[10] As choice sets are defined on an individual basis, the null model is derived on the assumption that within each choice set the probability of choosing any of the alternative stations is equal

[11] The car park spaces and free car park variables were interacted with a dummy variable representing car as access mode, so these parameters were only estimated against choice situations where access mode was car.

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. 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. 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, described in Section 2.4.1, 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 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 model FM4 by estimating five separate parameters for the variable based on banded access journey 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. 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. Model FM4 is the best model suitable for integrating into rail demand flow models, with an adjusted R-squared of 0.76 for both the LATIS and WG datasets.

3.2 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. Initial RPL models were run, using the best performing MNL models as the starting point, with all parameters specified as random to test whether the standard deviation of each parameter was significantly different from zero. If the standard deviation is not significant, it indicates that there is no individual taste variation for that parameter. As the parameter for all the model variables is expected to have the same sign for all individuals, the log normal distribution was specified, 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. Using Model TE17 as the starting point, the mode-specific access time parameters had a significant standard deviation for both WG and LATIS (excluding cycle mode in the WG model). Additionally, the standard deviation of the nearest station and car park spaces parameters were significant in the LATIS model. The z-values of the standard deviations for the other parameters were very low, and not close to critical values. Based on these findings an RPL model with mode-specific access times parameters specified as random was run for both datasets, and the results are shown in Table 4. Both the LATIS and WG models have higher log-likelihood and adjusted R-squared values than the MNL equivalent model, and while predictive performance was slightly better for the WG model, it was marginally worse for the LATIS model. The standard deviations are significant for all the random parameters, 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., 2015). 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), whilst 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.

Table 4: Trip end station choice models (RPL)

	WG – Trip end station choice model (RPL2) AdjR^2 = .73 LogLik = -3649 Pred. Perf. Diff (%) = 25.9						LATIS – Trip end station choice model (RPL2) AdjR^2 = .70 LogLik = -6553 Pred. Perf. Diff (%) = 23.6					
	Random Parameters[1]					Non-random parameters	Random parameters[1]					Non-random parameters
	Mean of ln(coef.)	Std. dev of ln(coef.)	Median coef.	Mean coef.	Std. dev. of coef.	Coefficient	Mean of ln(coef.)	Std. dev of ln(coef.)	Median coef.	Mean coef.	Std. dev. of coef.	Coefficient
Nearest station						.54***						.46***
Time (walk) mins	-1.56***	.87***	.21	.31	.32		-1.93****	.44****	.14	.16	.07	
Time (cycle) mins						-.17***	-2.26****	.38****	.10	.11	.04	
Time (bus\pt) mins	-2.80***	.67***	.06	.08	.06		-2.66****	.85****	.07	.10	.10	
Time (car) mins	-1.41***	.56***	.24	.29	.17		-1.39****	.83****	.25	.35	.35	
Full-time staff						2.30***						2.92***
Part-time staff						1.80***						1.15***
Daily frequency						.003***						.002***
CCTV						.92***						2.85***
Car park spaces						.006***						.002***
Free car park						.52***						.73^{ns}
Ticket machine						1.19***						1.04***
Toilets						.08^{ns}						.73***

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

4 Model appraisal

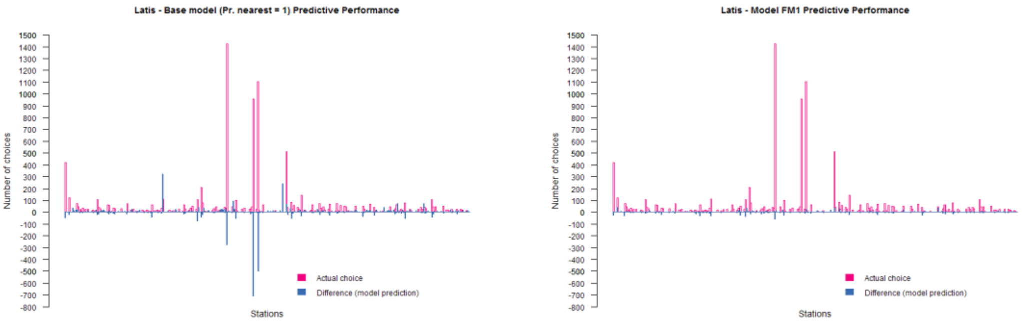
4.1 Predictive performance

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 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., 2015 p.502). To assess the overall performance of the models reported in this paper, 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 Figure 5 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 (a) and LATIS FM1 (b), illustrating the substantially better predictive performance of the latter.

Table 5: Summary of model predictive performance

Model	Predictive performance (abs. diff as % of total choice situations)		
	LATIS	WG	Comments
Base model (prob. Nearest = 1)	50.91	40.99	
TE17	23.53	27.35	
RPL2	23.58	25.85	
FM4	14.61	22.35	
RPL4	n/a[12]	21.13	RPL for flow-related model
FM2 (model calibrated on other dataset)	20.16	34.80	

[12] Model failed to fit. Initial iterations unable to improve log-likelihood fn. of MNL model (used for starting values).



(a) Base model (nearest prob = 1)

(b) Model FM1

Figure 5: Model predictive performance (LATIS)

4.2 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 plots the parameter estimates along with confidence intervals for model FM2 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 small number of observations.

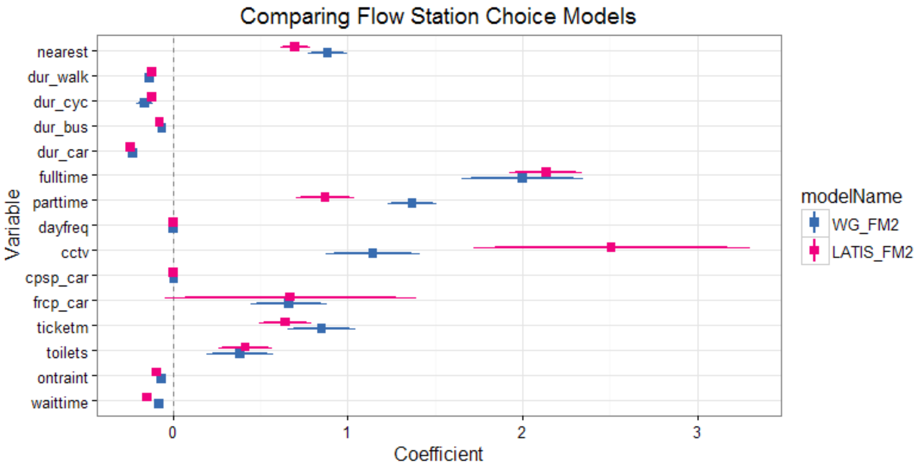


Figure 6: Parameter estimates for WG and LATIS model FM2 showing 95% and 99% confidence intervals

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.

5 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-squared 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. The superior predictive performance of the models compared to the base model, suggests that using them to define probabilistic station catchments could significantly improve the accuracy of the aggregate demand models that are commonly used in the UK to assess the viability of proposed schemes for new stations. Future work will now 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.

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; and Ordnance Survey Ltd for providing AddressBase data. This work forms part of a PhD funded by EPSRC DTG Grant EP/M50662X/1.

References

Association of Train Operating Companies (ATOC). (2013). *Passenger demand forecasting handbook v5.1*.

Blainey, S., & Evens, S. (2011, October). *Local station catchments: reconciling theory with reality*. Paper presented at AET European Transport Conference.

BR Fares Ltd. (2016). *BR Fares*. webpage. URL: http://www.brfares.com

Econometric Software, Inc. (2012). Nlogit 5 [Computer software].

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: Proc. of world conference on transport research (Vol. 2, pp. 1729-1740).

Hensher, D. A., Rose, J. M., & Greene, W. H. (2015). *Applied choice analysis: a primer* (2nd ed.). Cambridge University Press.

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.

Kastrenakes, C. R. (1988). *Development of a rail station choice model for NJ Transit. Transportation Research Record, 1162, 16-21.*

Ramsey, P. (2011). *Indexed nearest neighbour search in PostGIS*. webpage. Retrieved from http://boundlessgeo.com/2011/09/indexed-nearest-neighbour-search-in-postgis/

Steer Davies Gleave. (2010). *Station usage and demand forecasts for newly opened railway lines and stations (Final Report prepared for Department for Transport)*.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Wardman, M., & Whelan, G. (1999). *Using geographical information systems to improve rail demand models*. (Report to Engineering and Physical Sciences Research Council)

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., & Blainey, S. (2016a, January). *Defining probability-based rail station catchments for demand modelling*. Paper presented at 48th Annual UTSG Conference, Bristol, GB.

Young, M., & Blainey, S. (2016b). *Railway Station Choice Modelling: A Review of Methods and Evidence*. Manuscript submitted for publication.