



## **DEVELOPMENT OF INTEGRATED DEMAND AND STATION CHOICE MODELS FOR LOCAL RAILWAY STATIONS AND SERVICES**

Marcus Young  
Transportation Research Group, University of Southampton

### **1. INTRODUCTION**

#### **1.1. Background**

The railway in Britain has experienced considerable growth in recent years. Over the last ten years alone, total passenger journeys have increased by 51% (an additional 584 million journeys) (ORR, 2017), 61 new stations have opened, and several new lines have been built (Railfuture, 2017). This growth looks set to continue, with new lines and stations currently under construction or planned, and campaigns being run nationwide by communities eager to be connected to the rail network (Campaign for Better Transport, 2017). However, there are concerns about the accuracy of the station demand forecasts that are used to determine the viability of proposed new schemes. A report commissioned by the UK Government to investigate the issue, compared forecast and observed demand at 23 newly opened stations. It found that forecast demand was above or below observed demand by a margin of more than 20% in 14 cases, including an under-prediction in excess of 100% at three stations (Steer Davies Gleave, 2010). More recently, the demand forecast for the new Borders Railway line in Scotland was described as a 'shocking failure', after usage figures revealed that passenger trips in the first year of operation were up to eight times higher than forecast for three of the new stations, and lower than predicted for the other four (Campaign for Borders Rail, 2016). Inaccurate forecasts can have potentially serious consequences. Under-prediction might lead to the unnecessary rejection of a proposal on the grounds of the perceived benefit-cost ratio, or to the inadequate provision of station and network infrastructure. Conversely, over-prediction, or not adequately accounting for abstraction from existing stations, could result in a new station that fails to deliver the expected economic and societal benefits.

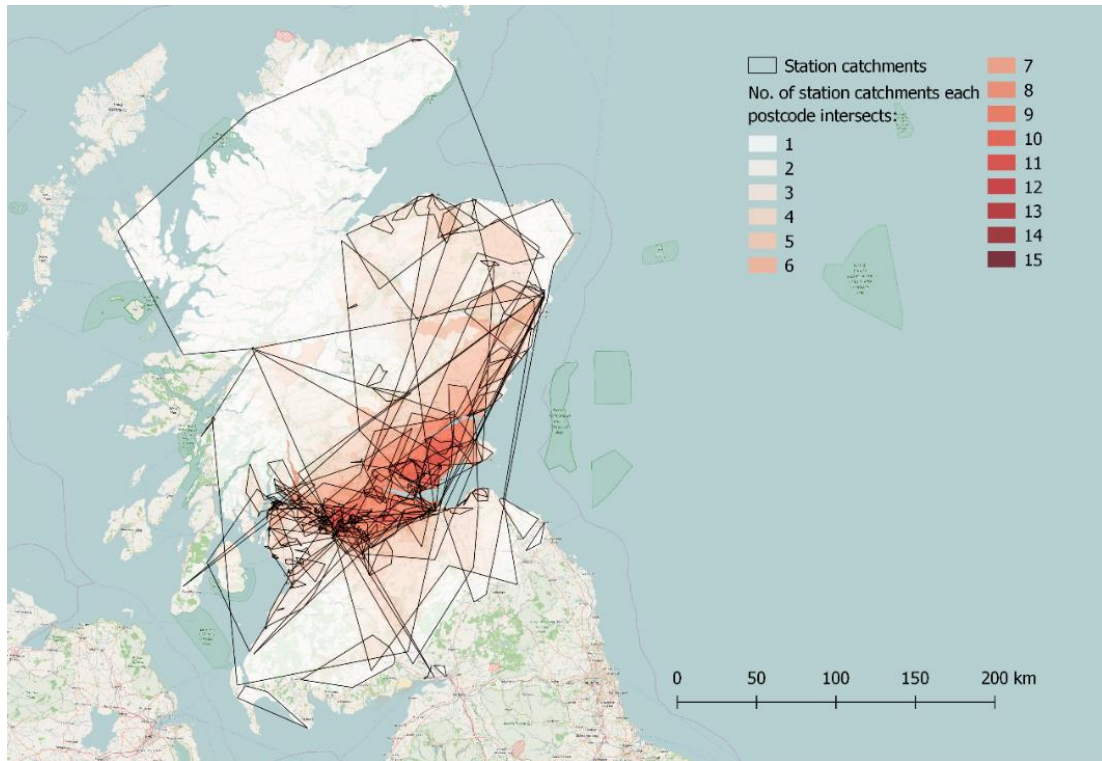
#### **1.2. The station catchment problem**

Although the UK Department for Transport has published some general guidance for those carrying out or commissioning demand forecasts for new local railway stations (DfT, 2011), the models used are usually developed for, and applied to, a specific local context. In most cases trip-rate or trip-end

models are adopted, as was the case in two-thirds of the stations/lines considered in the Steer Davies Gleave report. Trip-rate models assume the number of trips to be some function of the population in the area surrounding a station (its catchment), while trip-end models include additional variables relating to station services, facilities or the locality. The research summarised in this paper builds upon previous work by Blainey (2010) and Blainey and Preston (2013) to develop national trip-end models suitable for general application in forecasting demand for new local rail stations in England and Wales. A weakness of this work, in common with trip-end models generally, lies in how the station catchments are defined. Two methods are typically used; either a distance- or time-based buffer is placed around the station, or the study area is divided into zones and each zone is assigned to its nearest station. The latter was the method adopted by Blainey (2010), with census output areas used as the zonal units. Both approaches produce discrete non-overlapping catchments which imply that station choice is a deterministic process (anyone within a zone will always use the same station) and that stations do not compete with one another. However, analysis of passenger survey data reveals that station catchments are far more complex entities. Figure 1 shows approximate catchments for stations in Scotland, created by computing the polygon that encompasses the observed origin postcode<sup>2</sup> of passengers using each station; with the choropleth indicating the number of station catchments that each postcode intersects<sup>1</sup>. This reveals substantial overlapping of catchments and confirms that station choice is not deterministic, even when small-scale origin zonal units are used.

This research seeks to address the problem by allocating a set of alternative stations to each unit-level postcode and then using station choice models to calculate the probability of each of these stations being chosen. Zonal population can then be apportioned to each station based on these probabilities, thereby creating probabilistic catchments. In addition, by using full postcodes as the zonal unit, the station catchments are defined at a much higher spatial resolution than in previous work. If station catchments are not correctly defined, then inappropriate weight will be given to other model variables, such as service quality measures, as drivers of trip generation, rather than the catchment population. By defining more realistic catchments, the parameter estimates will be more robust, and the models will be more transferable (Wardman and Whelan, 1999).

Figure 1: Approximate observed station catchments generated from Scottish passenger survey data, with each postcode classified to show the number of station catchments that it intersects. Basemap is © openstreetmap.org contributors



### 1.3. Previous research

While there is a substantive body of prior station choice research (see Young and Blainey (2017) for a comprehensive review), relatively little attention has been given to how station choice models can be used to improve rail demand models. There are two notable exceptions. The most refined methodology is probably that proposed by Lythgoe and Wardman (2002, 2004), where station choice is an intrinsic component of a flow model<sup>3</sup>, with a station's generation potential represented by the population within 40km allocated to a grid of zones. However, this approach was intended to forecast demand for parkway stations and is limited to modelling inter-urban journeys greater than 80km (subsequently reduced to 40km by Lythgoe et al. (2004)). The only research to adopt a similar approach to that taken in this paper, is the work of Wardman and Whelan (1999). They attempted to incorporate probability-based station catchments into a flow model by apportioning the population of postal sectors<sup>4</sup> to one of five competing stations. However, due to time and computer resource constraints they had to use a subset of the flow data, which resulted in the model failing to converge. They recommended further work, noting that they had 'seriously underestimated the complexity of [the] task and the computing and time resources required'. However, this approach has not been revisited

since, despite the considerable advances in computing power over the past two decades.

In the next section the paper will outline the models that have been developed to generate station choice probabilities. The methodology used to calibrate trip-end models for local stations in GB, using both deterministic and probabilistic catchments, is then described and the calibration results presented. Both model variants are then used to forecast demand for several recently opened individual stations, and for a set of stations on a newly constructed line. Forecast demand is then compared with observed demand and, where information is available, with the forecast made as part of the scheme appraisal process.

## 2. STATION CHOICE MODEL CALIBRATION

As this paper is primarily concerned with the integration of station choice models into trip-end models, only a brief overview of the station choice models will be given here, concentrating on modifications made to the models since earlier work. Full details about data preparation and validation, derivation of variables and the modelling approach adopted can be found in Young and Blainey (in press). Multinomial logit (MNL) models of station choice were developed using revealed preference data obtained from on-train passenger surveys carried out in Wales and Scotland during 2014 and 2015. Models were calibrated using the combined dataset of 14,422 observations (choice situations) from both the Welsh and Scottish surveys. The full unit postcode was taken as the trip origin, and the choice set was defined as the ten nearest stations by drive distance to each postcode. In the previous work, mode-specific parameters for access time were found to yield the best performing models. However, as choice of access mode is not being modelled, the square root of access distance has been included in the new models. This was found to perform better than an untransformed or log-transformed variable, and was superior to time-based access variables. A range of explanatory variables were tested using a manual forward selection procedure, and the best performing model was SC1, where the probability of individual  $n$  at origin  $i$  choosing station  $k$  from a choice set of  $K$  alternative stations, is given by the following formula:

$$Pr_{nik} = \frac{\exp(\beta N_k + \gamma \sqrt{D_{ik}} + \delta U_k + \epsilon \ln F_k + \zeta C_k + \eta P_k + \theta T_k + \iota B_k)}{\sum_{k=1}^K \exp(\beta N_k + \gamma \sqrt{D_{ik}} + \delta U_k + \epsilon \ln F_k + \zeta C_k + \eta P_k + \theta T_k + \iota B_k)} \quad (1)$$

where  $D$  is the access distance by road from origin  $i$  to station  $k$ ;  $F$  is the daily service frequency at station  $k$ ;  $P$  is the number of car parking spaces at station  $k$ ;  $N$ ,  $U$ ,  $C$ ,  $T$  and  $B$  are dummy variables that take the value of 1 if station  $i$  is the nearest station, unstaffed, has CCTV, has a ticket machine, or has a bus interchange respectively, and zero otherwise; and  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ,  $\eta$ ,  $\theta$  and  $\iota$  are parameters to be estimated. The results from calibrating this model are summarised in Table 1.

## 2.1. The spatial choice problem

Due to the underlying assumptions of the MNL model, it exhibits proportional substitution behaviour (a consequence of the independence of irrelevant alternatives (IIA) property). This means that when a new alternative is added to a choice set the probability of each of the existing alternatives will be reduced by the same proportion. However, it is a reasonable assumption that stations that are closer to each other in space will be better substitutes for one another than stations that are further apart. A new station would be expected to abstract proportionately more passengers from a station closer to it than one further away. A potential mechanism to address this issue is to incorporate a measure of the accessibility of each alternative to all other alternatives within a choice set. This ‘accessibility term’ is often a Hansen-type measure, where the distance between alternatives is weighted by a size-based attraction variable. As the term includes information from other alternatives the IIA property no longer holds and the model can capture competition (or agglomeration) effects. Examples include Fotheringham’s competing destinations model (CDM) (see Pellegrini and Fotheringham (2002) for a review); and a recent application to account for spatial competition in workplace choice models (Ho and Hensher, 2016). To assess the potential of this approach, the following form of the accessibility term, as proposed by Fotheringham, was tested:

$$A_{ni} = \left( \frac{1}{M-1} \sum_{\substack{k \\ k \neq j}} \frac{W_k}{d_{jk}} \right)^\theta \quad (2)$$

where  $M$  is the total number of alternatives for individual  $n$  at origin  $i$ ,  $W$  is a weight,  $d$  is the distance from alternative  $j$  to alternative  $k$ , and  $\theta$  is a parameter to be estimated. As  $A$  increases an alternative is closer to more ‘attractive’ alternatives. Fotheringham states that the CDM can be derived, and under certain circumstances be consistent with random utility theory, simply by included the accessibility term in the utility function (Fotheringham, 1986). If  $\theta$

$< 0$  then alternatives that are more isolated will have a higher probability of being chosen (competition effect), but if  $\theta > 0$  then alternatives closer together will have a higher probability of being chosen (agglomeration effect). Two variants of the accessibility term were tested. In the first (model SC2), the weight was defined as the total number of entries and exits at the station in 2014/15. As this figure will be unknown for a proposed new station, a second model (SC3) was estimated using a fixed weight for each station category, based on thresholds specified in the category definitions (see Annex C in Green and Hall (2009)), as shown in Table 2. The logarithmic transformation of the accessibility term was added to the models, as suggested by Fotheringham, with the model form becoming:

$$Pr_{nik} = \frac{\exp(\beta N_k + \gamma \sqrt{D_{ik}} + \delta U_k + \epsilon \ln F_k + \zeta C_k + \eta P_k + \theta T_k + \iota B_k + \kappa \ln A_k)}{\sum_{k=1}^K \exp(\beta N_k + \gamma \sqrt{D_{ik}} + \delta U_k + \epsilon \ln F_k + \zeta C_k + \eta P_k + \theta T_k + \iota B_k + \kappa \ln A_k)} \quad (3)$$

where  $A$  is the accessibility term, and  $\kappa$  the associated parameter to be estimated.

The results, shown in Table 1, indicate that models SC2 and SC3, incorporating the accessibility term, perform better than SC1. They have lower AIC values, and the AIC weights indicate a high probability that they are the better models. The parameter for the accessibility term is negative and significant in both models, indicating that there is a competition effect at play. The estimated parameter is very similar in the two models, suggesting that the fixed weight is a suitable proxy for the actual number of entries and exits. As the trip-end models are only intended to predict Category E or F stations, the appropriate weight will be known for any proposed new station (given that category F stations are unstaffed).

### 3. INTEGRATING TRIP-END AND STATION CHOICE MODELS

Previous research carried out at the University of Southampton Transportation Research Group has successfully developed linear regression models to forecast the number of trips made to/from local railway stations in England and Wales (Blainey, 2010). Station catchments were defined by allocating 2001 census output areas in England and Wales to their nearest station by road distance and applying a distance decay function to the population associated with each output area, reflecting the expectation that the number of trips generated by the population of an output area will fall as the distance from the

station increases. The best models were found to explain over 75% of variation in the observed data, and to better predict actual demand on the Ebbw Vale

Table 1: Station choice model calibration results

Variable	Model SC1			Model SC2			Model SC3		
	B	z	Sig	B	z	Sig	B	z	Sig
	No accessibility term			Accessibility term with trip entries/exits as weighting			Accessibility term using fixed weighting for each station category		
Nearest station (y/n)	0.6907	18.44	***	0.6855	18.25	***	0.6908	18.41	***
$\sqrt{V}$ (access distance)	-2.2618	-56.31	***	-2.2675	-56.35	***	-2.2652	-56.39	***
Unstaffed (y/n)	-0.6767	-16.01	***	-0.6522	-15.26	***	-0.6383	-14.53	***
$\ln$ (daily frequency)	1.1986	34.57	***	1.2100	34.74	***	1.2145	34.63	***
CCTV (y/n)	1.0708	8.59	***	1.0539	8.44	***	1.0760	8.63	***
Car park spaces (no.)	0.0013	16.48	***	0.0011	9.59	***	0.0012	13.67	***
Ticket machine (y/n)	0.9839	19.08	***	0.9758	18.91	***	0.9633	18.57	***
Bus interchange (y/n)	0.7585	13.61	***	0.7346	13.08	***	0.7308	12.95	***
$\ln$ (accessibility term)				-0.1314	-3.66	***	-0.1413	-3.22	***
McFadden's adjusted R <sup>2</sup>			0.71			0.71			0.71
AIC			19317.70			19306.20			19309.30
Delta AIC			11.50			0.00			3.10
Akaike weight			0.00			0.82			0.17

Table 2: Fixed weights for each main station category, used in the accessibility term

Station category	Weight (entries/exits)
A	2,000,000
B	2,000,000
C	1,000,000
D	500,000
E	250,000
F	125,000

branch line (which opened in 2008) than the methods used in the feasibility study carried out prior to scheme approval. As part of consultancy work carried out for the Welsh Government, the models were later re-calibrated using more recent data, including output area population from the 2011 census and station entries and exits (the basis of the dependent variable) from 2011/12 (Blainey, 2017). These models have been taken as the starting point for developing new trip-end models that incorporate probability-based catchments derived using the station choice models described in Section 2. The new models extend the earlier work in several key respects. Firstly, they are calibrated for stations in

the whole of GB, and not restricted to England and Wales. Secondly, unit postcodes are used to define catchment zones rather than census output areas; providing a much higher spatial resolution to the population data (there are some 1.5 million postcodes covering GB, compared to less than 0.25 million output areas). Thirdly, rather than assigning the population of each zone to its nearest station, the population is allocated to each station in a zone's choice set based on the probability that each station will be chosen, thus defining a probabilistic catchment.

### 3.1. Model formulation

The model form using simple (deterministic) catchments, as proposed by Blainey and Preston (2013), is as follows:

$$\ln \hat{V}_i = \alpha + \beta \left( \ln \sum_z P_z w_z \right) + \gamma \ln F_i + \delta \ln T_i + \epsilon \ln J_{it} + \zeta \ln Pk_i + \eta T e_i + \theta E l_i + \iota B_i \quad (4)$$

where  $\hat{V}_i$  is the estimated annual passenger entries and exits for station  $i$ ;  $P_z$  is the resident population of zone  $z$ ,  $Z$  is all zones where the closest station by car travel time is station  $i$ ;  $w_z$  is a decay function;  $F_i$  is weekday train frequency at station  $i$ ;  $T_i$  is distance in km from station  $i$  to the nearest Category A-D station;  $J_{it}$  is the number of jobs within a  $t$  minute drive of station  $i$ ,  $Pk_i$  is the number of parking spaces at station  $i$ ; and  $T e_i$ ,  $E l_i$  and  $B_i$  are dummy variables that take the value of 1 if station  $i$  is a terminus station, served by electric trains or a travelcard boundary station respectively, and zero otherwise; and  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ,  $\eta$ ,  $\theta$ , and  $\iota$  are parameters to be estimated.

To incorporate probabilistic station catchments, the model shown in Equation 4 can be amended to the following form:

$$\ln \hat{V}_i = \alpha + \beta \left( \ln \sum_z P r_{zi} P_z w_{zi} \right) + \gamma \ln F_i + \delta \ln J_{it} + \epsilon \ln Pk_i + \zeta T e_i + \eta E l_i + \theta B_i \quad (5)$$

where  $P r_{zi}$  is the probability of an individual located in zone  $z$  choosing station  $i$ ;  $Z$  now consists of all zones which have station  $i$  within their choice set; and  $\ln T_i$  has been removed.  $\ln T_i$  was added to try and capture potential competition effects of nearby large stations, something that should now be captured by the station choice component. This is the proposed general form of the model, with the nature of the zone being defined by the researcher. In the case of the models reported here, the zone is defined as the unit postcode.



An analysis of access trips from the passenger survey data indicated that a two-stage distance- or time-based decay function would be appropriate, with no decay for access journeys  $\leq 0.75\text{km}$  or  $\leq$  two-minutes. A power function (slope -1.5212) and an exponential function (slope -0.2432) gave the best fit for access distance and access time data respectively. The distance-based decay function  $w_{zi}$ , is therefore given by:

$$w_{zi} = \begin{cases} (d + 1)^{-1.5212} & \text{if } d > 0.75 \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

where  $d$  is the road distance in km from zone  $z$  to station  $i$ ; and the time-based function is given by:

$$w_{zi} = \begin{cases} e^{(-.2432 \times t)} & \text{if } t > 2 \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

where  $t$  is road travel time in minutes from zone  $z$  to station  $i$ .

### 3.2. Generating station choice probabilities

In order to generate the station choice probabilities ( $Pr_{zi}$  in equation 5), it is necessary to first define a station choice set for every unit postcode in GB (mainland only). Then, for each choice set, the probability of each station being chosen needs to be calculated. The unit postcode represents the spatial level at which resident population will be weighted, both by the decay function and the calculated choice probabilities, before being allocated to each station in the trip-end model.

An OD cost matrix analysis was run in ArcGIS to identify the 10 nearest stations (the destinations) to each postcode (the origins). Only stations that were in operation during 2011 were included, to correspond with the date of the census and the 2011/12 station usage data from the Office of Rail and Road (ORR). Any stations considered inaccessible to typical passengers (either on private property or some distance from the public road network), with no weekday service<sup>5</sup>, or located on the Isle of Wight were excluded. As ArcGIS failed to successfully complete an OD cost matrix analysis with all the origins loaded (some 1.45 million), the analysis was run in seven batches of approximately 200,000 origins. The data was exported into DBF format, and processed in R. A probability table was then created in a PostgreSQL database, consisting of ten rows for each postcode, one for each of the ten stations, along with predictor variables pulled in from other database tables. For each table row, the exponentiated measured utility was then calculated using the estimated



parameters from each station choice model and written to new columns. Using window functions (with records partitioned by postcode), the sum of measured utility for each choice set, and the probability of each station within a choice set being chosen, were calculated and written to new columns.

### **3.3. Model calibration**

In line with the earlier work, the calibration dataset was defined as those railway stations assigned to Network Rail categories E and F. For stations in England and Wales, the categories were obtained from a recent review of stations commissioned by the Department for Transport (Green and Hall, 2009). As no official source of the categories could be identified for stations in Scotland, an internal source was used. Any station opened after these lists were compiled was manually reviewed and allocated to a category based on the descriptions contained in Green and Hall (2009).

Only stations that had been open for the entirety of financial year 2011/12 were included in the dataset. This was the reporting period used by the ORR when compiling the total number of station entries and exits from ticketing data, the dependent variable used in the models (ORR, 2013). Stations were removed from the dataset if they had no weekday service, restricted public access, or were located on the Isle of Wight. For ticketing purposes some stations are grouped under a single common location, allowing passengers to travel to or from any station in a group using the same ticket. As the usage data for individual stations in each group is estimated and likely to be unreliable, they were also removed. The final calibration dataset consisted of 1792 stations.

The same predictor variables as those found to perform best in the previous work by Blainey and Preston (2013) were used to calibrate the trip-end models. These are summarised, with brief details about their source and derivation, in Table 3.

### **3.4. Model Results**

Initial calibration runs established that assigning each postcode to its nearest station by road distance, rather than by drive-time, produced the best deterministic catchment models; while the distance-based decay function was preferred over the time-based function for both the deterministic and probabilistic catchments. In addition, workplace population within a one minute drive of the station was found to perform better than the other thresholds.

Table 4 shows the results of four models: TE1 uses a simple catchment with the population of each postcode assigned to its nearest station; TE2 is the same as TE1 but applies the distance decay function; TE3 uses a probabilistic catchment with postcode population weighted by station choice probabilities (using SC1) and the distance-decay function; TE4 is the same as TE3 but uses SC3 (with the accessibility term) to derive the probabilities.

All the models fit the data very well, with TE4 the best fitting model (adjusted  $R^2 = 0.8506$ ). TE4 has the lowest AIC, and the AIC weights indicate a 98% probability that TE4 is the best of the four models. Turning to the parameter estimates, it is apparent that the population parameter is larger in models TE3 and TE4, while the daily frequency and terminus dummy parameters are smaller. This suggests that too much weight is being given to station service quality and characteristic in models TE1 and TE2, due to inadequacies in the catchment definition. It appears that models TE3 and TE4 can better account for differences in station usage that are explained by station catchments and their generation potential, and as a consequence should be more robust and transferable.

## **4. MODEL APPLICATION AND APPRAISAL**

### **4.1. Methodology**

In order to investigate the predictive performance of the integrated trip-end and station choice models, and assess whether probabilistic catchments produce more accurate estimates of station demand, it was first necessary to develop a methodology for generating the station choice and trip-end model inputs under the changed circumstances that result from a new station or new line being introduced. A crucial component is the procedure for redefining the set of alternative stations available at each unit postcode, so that any new stations appear as available choices when appropriate. It would not be practical to regenerate the nearest 10 stations for every postcode in GB, each time a new station needed to be modelled. Analysis of passenger survey data revealed that only a very small number (0.6%) of station access journeys exceed 60 minutes, irrespective of the chosen access mode. It was therefore decided that for any new station the 'area of interest' could be limited to those unit postcodes within 60 minutes' drive time. The nearest 10 stations, from the universal set that now includes the proposed new station(s), to each of these postcodes can then be generated. Any postcodes which do not include the new station(s) amongst the nearest 10 can then be discarded, as they will have will no influence on the

catchment definition. This reduces the amount of computing overhead involved in populating the probability database table and deriving variables (such as the accessibility term). The key steps involved in the proposed methodology are illustrated in Figure 2.

*Table 3: Summary of predictor variables used in trip-end models*

<i>Predictor variable</i>	<i>Description</i>	<i>Data source</i>	<i>Derivation</i>
<i>Resident population (no.)</i>	Population at unit-level postcode	2011 census. E&W: NOMIS, S: Scotland's Census.	
<i>Workplace population (no.)</i>	Residents (16-74) in employment within x minutes of the station by road network	Workplace zones (E&W, NOMIS); output areas (Scotland, Scotland's Census)	Service area analysis (1,2,3 and 4 minutes) and spatial join in ArcGIS
<i>Daily service frequency (no.)</i>	No. of trains on typical weekday	GB train schedule (GTFS format)	Imported into PostgreSQL, then a SQL query
<i>Nearest Category A-D station (km)</i>	Nearest Category A,B,C or D station by road network	NAPTAN	OD cost matrix analysis in ArcGIS
<i>Car park spaces (no.)</i>	Number of car park spaces at station	NRE Knowledgebase XML feed	Parsed in R
<i>Electric services (y/n)</i>	Whether station is served by electric trains	GB train schedule (GTFS)	Adjusted to take account of schemes since 2011/12
<i>Travelcard boundary (y/n)</i>	Whether station is at the boundary of a city or regional travelcard scheme	Schemes identified for: Strathclyde, London, West Midlands, Merseyside, Greater Manchester, West Yorkshire, Tyne & Wear, and South Yorkshire	Various web sites
<i>Terminus (y/n)</i>	Whether a station is at the end of a line	Previous work by Blainey (2010)	Updated for Scottish stations and new stations with reference to rail maps
<i>Accessibility term</i>	See section 2.1		Distances between station pairs measured using an OpenTripPlanner instance.

Table 4: Results of trip-end model calibrations

Variable	TE1			TE2			TE3			TE4		
	B	t-value	Sig	B	t-value	Sig	B	t-value	Sig	B	t-value	Sig
	Population assigned to nearest station (by distance)			Population assigned to nearest station (by distance); distance decay function			Population probability-weighted; distance decay function			Population probability-weighted; distance decay function; accessibility term		
Intercept	2.58	14.37	***	2.37	13.53	***	3.67	38.50	***	3.65	38.30	***
$\ln(\text{population})^{\wedge}$	0.23	15.06	***	0.34	17.30	***	0.37	20.14	***	0.37	20.38	***
$\ln(\text{daily train frequency})$	1.43	50.90	***	1.36	48.40	***	1.14	41.47	***	1.13	41.21	***
$\ln(\text{dist. to Cat A-D station})$	0.15	6.20	***	0.21	8.64	***						
$\ln(\text{work pop. 1 min})^{\wedge}$	0.09	13.48	***	0.06	7.66	***	0.05	7.75	***	0.05	7.70	***
$\ln(\text{car park spaces})^{\wedge}$	0.13	13.43	***	0.15	15.44	***	0.13	14.14	***	0.13	14.07	***
Electric services	0.20	4.61	***	0.22	5.09	***	0.24	5.93	***	0.24	5.97	***
Travelcard boundary	0.31	3.30	***	0.30	3.26	**	0.30	3.29	**	0.30	3.26	**
Terminus	0.90	10.31	***	0.86	10.03	***	0.78	9.37	***	0.78	9.34	***
McFadden's adjusted R <sup>2</sup>			0.8378			0.8434			0.8500			0.8506
AIC			3911.7690			3848.2150			3770.2630			3762.5480
Delta AIC			149.2210			85.6670			7.7150			0.0000
Akaike weight			0.0000			0.0000			0.0207			0.9793

Notes:  $\wedge \log(\text{variable} + 1)$  used due to presence of zero values

## 4.2. Demand predictions

The methodology was initially applied to forecast demand for three new stations that opened in 2012 (Fishguard & Goodwick) and 2013 (Conon Bridge and Energlyn & Churchill Park). The forecasts obtained using the three trip-end models (TE2, TE3 and TE4) are shown in Table 5, along with actual station usage data for 2015/16. All three models produced an accurate forecast for Energlyn & Churchill Park, within +/- 2% of actual trips. While the probabilistic models under-forecast demand by 18% at Fishguard & Goodwick, this represents a 10 percentage-point adjustment (in the desired direction) compared to the deterministic catchment model. All the models over-forecast demand at Conon Bridge, by around 60%, although this is more accurate than the original project forecast of 36,000 trips (Railfuture, 2017). These initial findings, while not conclusive, suggest that probabilistic catchments have the potential to adjust, and correct, forecasts produced using simple catchments.

Figure 2: Methodology for generating demand forecast for new stations

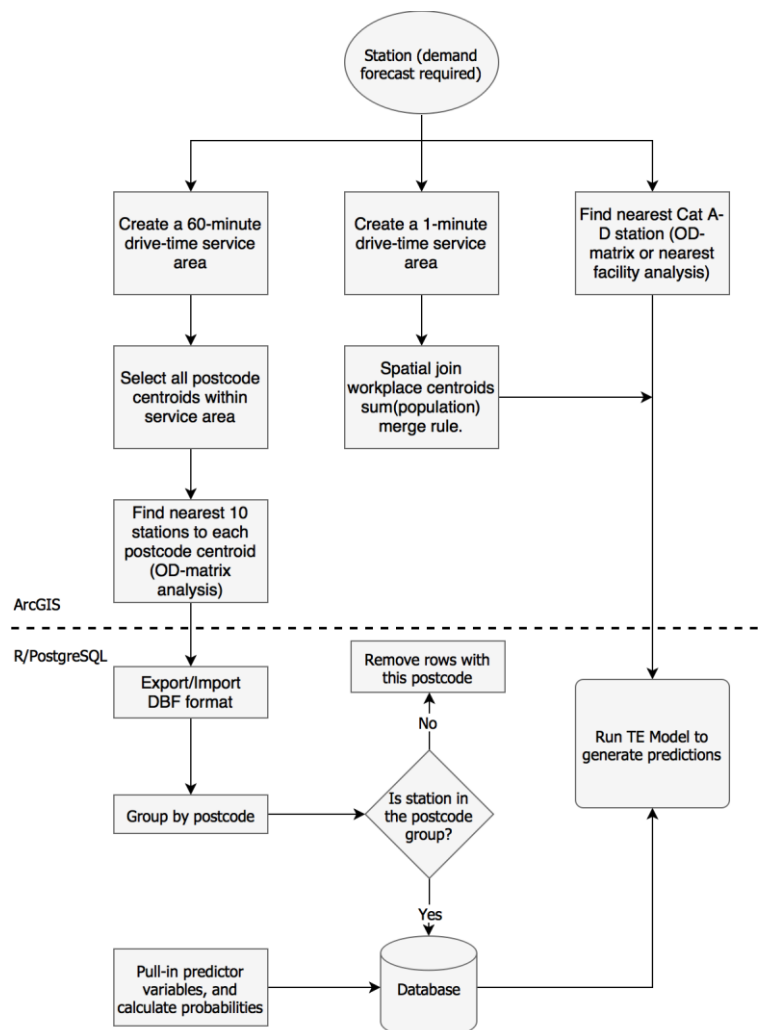


Table 5: Demand forecasts for three new stations and comparison with actual trips in 2015/16

Station	Weighted catchment pop. TE2	Weighted catchment pop. TE3	Weighted catchment pop. TE4	ORR trips 2015/16	TE2 trip forecast	% diff from 15/16	TE3 trip forecast	% diff from 15/16	TE4 trip forecast	% diff from 15/16
Conon Bridge (opened Feb. 2013)	1249	859	856	15276	24453	60	25091	64	25090	64
Energlyn & Churchill (opened Dec 2013)	3864	1183	1180	74206	73015	-2	75467	2	75329	2
Fishguard & Goodwick (opened May 2012)	1992	1429	1416	19946	14345	-28	16317	-18	16387	-18

Table 6: Demand forecast for Borders Railway (new stations only) and comparison with actual trip data in 2016/17

Station	Weighted catchment population			Actual trips				Trip forecasts				
	TE2	TE3	TE4	First year from opening	Lennon data^ 2016/17	Final business case forecast	% diff from 16/17	Simple catchment (TE2)	% diff from 16/17	Probability-based catchment (TE3)	Probability-based catchment (TE4)	% diff from 16/17
Tweedbank	2476	2015	2010	337864	474000	43242	-91	806146	70	485965	481432	2
Galashiels	4737	3742	3749	201666	342000	46862	-86	200381	-41	146705	147335	-57
Stow	700	476	478	48282	66000	11686	-82	96263	46	67705	67136	2
Gorebridge	3189	2182	2198	74891	93000	180038	94	254489	174	204831	204737	120
Newtonrange	3538	2204	2197	96735	137000	105836	-23	239277	75	196977	196190	43
Eskbank	5230	2628	2588	133121	228000	261050	14	312784	37	242701	240717	6
Shawfair	1323	320	324	16853	21000	123720	489	106627	408	65467	64979	209
Totals				909412	1361000	772434	-43	2015967	48	1410350	1402525	3

Notes: ^ Trip data read from graphs provided in the Borders Railway Year 1 Evaluation report, therefore figures are only indicative of actual values

The methodology was next applied to forecast demand for the seven stations that were built as part of the new Borders Railway in Scotland, which opened in September 2015 (see Figure 3). The results are shown in Table 6 and summarised in Figure 4, along with the final business case forecast for the first 12 months, and actual usage in 2016/17. The predictor variables for each of the stations are summarised in Table 7.

Figure 3: The Borders Railway. Source: Wikipedia

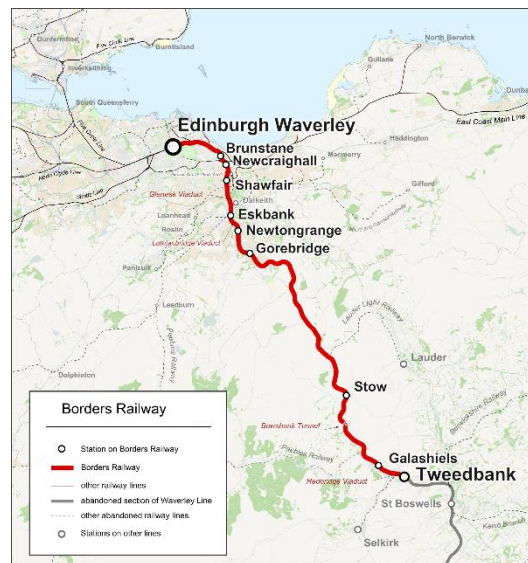


Table 7: Predictor variables for Borders Railway (new stations only)

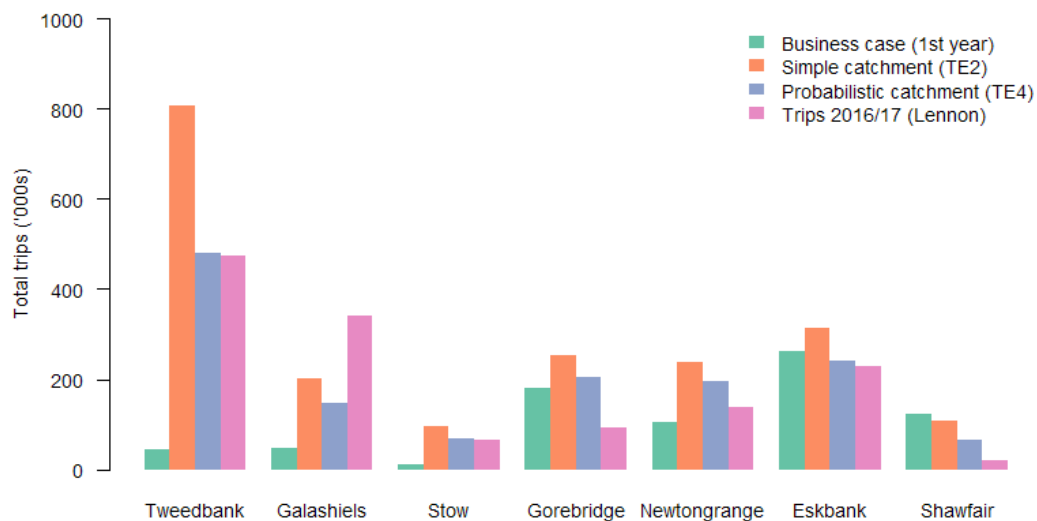
Station	Workplace pop. (1 min)	Daily service freq.	Car park spaces	Nearest Cat A-D station	Terminus station
Tweedbank	1120	66	235	54.89	1
Galashiels	3746	66	0	50.62	0
Stow	718	47	33	39.08	0
Gorebridge	2330	66	73	16.58	0
Newtonrange	1965	66	56	13.17	0
Eskbank	819	66	248	11.39	0
Shawfair	0	66	59	10.16	0

The results show that model TE4 (probabilistic catchments) has performed reasonably well across the stations, and in all but one case has produced more accurate forecasts than model TE2 (simple catchments). The forecasts for three of the stations, Tweedbank, Stow and Eskbank, are within 10% of actual trips. This is substantially better than the performance of model TE2. Model TE4 has noticeably corrected the large over-prediction for Tweedbank station, reducing it from +70% to +2% of actual trips. However, TE4 has under-predicted



Galashiels by 57%, performing worse than TE2 (-41%). Unlike the other new stations, Galashiels has no station car park. It is possible that the station choice model is penalising Galashiels excessively, attributing higher probabilities to Tweedbank and Stow than justified for some postcodes. The under-forecasting of Galashiels by all the models may also be the consequence of no car parking spaces appearing in the model. This could be due to the trip-end model performing less well in these circumstances, or it may indicate that alternative parking opportunities are available for station users. Model TE4 has over-forecast Gorebridge (+120%), although this is rather better than TE2 (+174%). There is anecdotal evidence that competition from local bus services might be suppressing demand at Gorebridge, something that the models would not be sensitive to. Considering all seven stations together, model TE4 predicts a total of 1.40 million trips, which compares to 1.36 million actual trips in 2016/17. Despite some shortcomings, it is particularly encouraging that the models have performed substantially better than the business-case projections for the three Scottish Borders stations (Tweedbank, Galashiels and Stow).

Figure 4: Comparison of demand forecasts and actual trips in 2016/17 for the new stations on the Borders Railway



## 5. CONCLUSIONS AND FUTURE WORK

This research has shown that it is possible, through the use of a station choice component, to incorporate more realistic representations of station catchments into the type of aggregate demand model that is commonly used in the UK to forecast demand for new local stations. The trip-end models that define probability-based catchments perform better than those with traditional deterministic catchments, both in terms of measures of model performance, and their predictive ability when applied to real-world forecasting scenarios.



Although the models that have been developed need to be validated more extensively, this research has important policy implications. The findings suggest that it is possible to develop a robust and transferable national forecasting model for new local railway stations. Such a model, which could be re-calibrated and refined on a regular basis, may be preferable in some circumstances to models that are developed on an ad hoc basis when a local need arises. A national model would also be a useful comparator tool that could help assess the reliability of forecasts produced by locally developed models.

Previous work has established that station choice models which include predictor variables relating to the rail leg, such as journey time or number of transfers, perform considerably better as predictive models than those used in this research (Young and Blainey, in press). Such station choice models are suitable for incorporating into flow models, which forecast trips between origin-destination station pairs. The calibration of flow models with these enhanced probabilistic catchments is therefore an important avenue for future research. Further work is also needed to extend the model application methodology to enable the impact of passenger abstraction from existing stations to be assessed.

## **ACKNOWLEDGEMENTS**

The author wishes to thank the Welsh Government and Transport Scotland for providing passenger survey data. This work was supported by the EPSRC under DTG Grant EP/M50662X/1. 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.

## **BIBLIOGRAPHY**

Alderson, J., & McDonald, I. (Eds.). (2017). Britain's growing railway (6<sup>th</sup> ed.). Railway Development Society.

Blainey, S. (2010). Trip end models of local rail demand in England and Wales. *Journal of Transport Geography*, 18(1), 153–165.

Blainey, S. (2017). A new station demand forecasting model for Wales. (Unpublished working paper).

Blainey, S., & Preston, J. M. (2013). A GIS-based appraisal framework for new local railway stations and services. *Transport Policy*, 25, 41–51.



Campaign for Better Transport. (2017). Re-opening rail lines. Webpage. Retrieved 5 September 2017, from <http://www.bettertransport.org.uk/re-opening-rail-lines>

Campaign for Borders Rail. (2016, December 3). Rail monitoring group attacks Borders Railway forecasting failure. Webpage. Retrieved 5 September 2017, from <https://campaignforbordersrail.wordpress.com/2016/12/03/rail-monitoring-group-attacks-borders-railway-forecasting-failure/>

Department for Transport. (2011). Guidance note on passenger demand forecasting for third party funded local rail schemes.

Fotheringham, A. S. (1986). Modelling hierarchical destination choice. *Environment and Planning A*, 18(3), 401–418.

Green, C., & Hall, P. (2009). Better rail stations (An independent review presented to Lord Adonis, Secretary of State for Transport).

Ho, C. Q., & Hensher, D. A. (2016). A workplace choice model accounting for spatial competition and agglomeration effects. *Journal of Transport Geography*, 51, 193–203.

Lythgoe, W., & Wardman, M. (2002, September). Estimating passenger demand for parkway stations. Paper presented at AET European Transport Conference.

Lythgoe, W., & Wardman, M. (2004). Modelling passenger demand for parkway rail stations. *Transportation*, 31(2), 125–151.

Lythgoe, W., Wardman, M., & Toner, J. (2004, October). Enhancing rail passenger demand models to examine station choice and access to the rail network. Paper presented at AET European Transport Conference.

Pellegrini, P. A., & Fotheringham, A. S. (2002). Modelling spatial choice: a review and synthesis in a migration context. *Progress in Human Geography*, 26(4), 487–510.

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

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

Young, M., & Blainey, S. (in press). Development of railway station choice models to improve the representation of station catchments in rail demand models. *Transportation Planning and Technology*.

Young, M., & Blainey, S. (2017). Railway station choice modelling: a review of methods and evidence. *Transport Reviews*. (Advance online publication.)



## NOTES

- <sup>1</sup> This analysis is based on the survey data described in Section 2.
- <sup>2</sup> In the UK a unit postcode represents the most detailed spatial unit available from postcode data. For small postal users (i.e. not business addresses), a unit postcode typically represents around 15 addresses, though it is possible to contain up to 100 addresses in densely populated areas.
- <sup>3</sup> 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.
- <sup>4</sup> There are approximately 3000 addresses in a postcode sector.
- <sup>5</sup> Many of these stations are served by so-called 'parliamentary trains', a bare minimum service to avoid invoking the costly formal process of closing a station.