Defining probability-based rail station catchments for demand modelling

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

Dr Simon Blainey
Lecturer
Transportation Research Group, University of Southampton

Abstract
The aggregate models commonly used in the UK to estimate demand for new local rail stations require the station catchment to be defined first, so that inputs into the model, such as the population from which demand will be generated, can be specified. The methods typically used to define the catchment implicitly assume that station choice is a deterministic process, and that stations exist in isolation from each other. However, studies show that pre-defined catchments account for only 50-60 percent of observed trips, choice of station is not homogeneous within zones, catchments overlap, and catchments vary by access mode and station type. This paper describes early work to implement an alternative probability-based approach, through the development of a station choice prediction model. To derive realistic station access journey explanatory variables, a routable multi-modal network, incorporating data from OpenStreetMap, the Traveline National Data Set and National Rail timetable, was built using OpenTripPlanner and queried using an API wrapper developed in R. Results from a series of multinomial logit models are presented and a method for generating probabilistic catchments using estimated parameter values is described. An example probabilistic catchment is found to provide a realistic representation of the observed catchment, and to perform better than deterministic catchments.

1 Introduction

In Great Britain (GB), travel by rail has experienced a resurgence in recent decades, with a rapid growth in passenger journeys replacing the declines of the 1960s and 1970s and the modest growth of the 1980s. The average annual growth in passenger journeys was 4% between 1997/98 and 2013/14 (compared to 0.33% between 1980/81 and 1996/97), and has substantially out-paced growth in GDP (Rail Delivery Group 2014). The rail network has also expanded, with some 370 stations either reopened or newly built during the last 50 years, and many more currently under construction, proposed or being campaigned for by local communities (Railfuture 2015). Against this backdrop, the potential to meet local or regional transport needs, and also economic growth objectives, by investing in new rail stations, routes or services, is increasingly being recognised by UK local authorities, Passenger Transport Executives and Local Enterprise Partnerships (Department for Transport 2011).

To assess whether a particular scheme or intervention will achieve the required objectives, it is necessary to produce accurate forecasts of the effect on demand. The models typically used for new local rail stations (trip rate, trip end, and flow models) rely on aggregating relevant data, such as population, and need a station catchment to be defined. Two main methods are commonly used. The first is a buffer around a station, such as the 0.8km and 2km radial catchments proposed by Preston & Aldridge (1991); and the second divides the population into zones and allocates each zone to its nearest station. For example, Blainey (2010) assigned census output areas based on road travel time. Both methods produce deterministic catchments, where a particular trip origin is assumed to fall within the catchment of a single station, and competition between stations is not accounted for.

1.1 Catchments in reality

Previous studies have explored how representative defined catchments are of real station catchments. Blainey & Evens (2011) found that 2km non-overlapping radial catchments based on straight-line distance accounted for only 57% of observed trips, with large

variations at the individual station level. Neither do all passengers choose their nearest station. For example, Mahmoud et al. (2014) found that over 30% of commuters accessing a station by car did not choose their nearest; and Blainey & Preston (2010) found that only 53% of trip ends were located within catchments defined by assigning census areas to their nearest station by road access time. Several studies have used Geographical Information Systems to visualise observed catchments and reported significant overlap (Fan et al. 1993; Mahmoud et al. 2014), and others have found that station choice is not homogeneous within catchment zones (for example, Givoni & Rietveld (2014)). It is also intuitive to expect catchments to vary by access mode, with walk catchments smaller than those for motorised vehicles, and public transport catchments reflecting the routes serving a station (Givoni & Rietveld 2014). The size and shape of catchments will also depend on the type of station, with passengers willing to travel further to stations that offer inter-urban services, where the access journey is a smaller part of the total journey (Lythgoe & Wardman 2004).

It is clear that deterministic catchments do not reflect reality, suggesting that the aggregate demand models could be improved by using probability-based catchments. This paper will now briefly review prior station choice research and then describe the development of a station choice model and its application to generate probabilistic catchments.

2 Previous station choice research

A consensus has formed around using relatively simple closed-form discrete choice models to model station choice, with multinomial logit commonly used to model station choice alone (for example, see Blainey & Evens (2011); Mahmoud et al. (2014)), and nested logit used to model combined access mode and station choice (for example, see Debrezion et al. (2009); Givoni & Rietveld (2014)). These models are based on the concept of Random Utility Maximisation, where an individual is assumed to choose the one alternative, from a group of alternatives known as the choice set, that provides them with maximum utility. The researcher attempts to measure utility by identifying attributes of the alternatives and/or the individual. That part of utility that the researcher cannot measure is called the unobserved portion of utility and is treated as a random component. The utility that an individual derives from an alternative is expressed using the following formula:

$$U_{ni} = V_{ni} + \varepsilon_{ni} \quad (1)$$

Where U_{ni} is the utility for individual n of alternative i, V_{ni} is the utility measured by the researcher, and ε is the unobserved portion of utility. In practice V, which is known as the representative or observed utility, will be a function consisting of the selected attributes and their respective parameters. The parameters, if unknown, are obtained statistically, for example by maximum likelihood estimation.

The effect of a range of factors related to the access journey and service levels has been consistently reported. Station utility decreases as the access journey becomes further or longer, as the rail leg journey time increases, and when the journey involves more transfers or has a higher fare; and utility increases as departures become more frequent. The effect of station facilities, such as car parking, is more problematic, potentially due to endogeneity issues, and only limited attention has been given to land-use factors. Recently, more complex model frameworks have been proposed, for example using mixed logit (Chen et al. 2014), but it remains unclear how well even the simple models can predict station choice in real-world scenarios or how the models can be used to improve industry standard rail demand forecasting methods. Most prior studies have focussed on developing models to better understand the factors that influence station choice, and there have been few attempts to create a transferable station choice model and integrate it into one of the aggregate models used to predict demand at new stations. Wardman & Whelan (1999) attempted to incorporate probabilistic station catchments into a direct demand model by apportioning population to one of five competing stations for each postal sector, but due to time and computer resource constraints they had to use a subset of the data which resulted in the model failing to converge. In the work of Lythgoe & Wardman (2004) station choice is an intrinsic component of a spatial interaction model, but the method is limited to forecasting demand for inter-urban journeys in excess of 40km.

3 Data sources and processing

Methods and practices that support research reproducibility and automated workflows were adopted for the manipulation and analysis of data, with open-source data management and analytical tools being used wherever it was practical to do so. This approach enables processes to be easily repeated on the same or a new dataset, and allows process modifications to be readily applied. The R environment was used for data processing, descriptive analysis, developing the choice models, and graphical output. Data was stored in a series of related tables in a PostgreSQL database and the PostGIS spatial extension was used for spatial analysis. Data visualisation was carried out in QGIS.

3.1 Revealed preference data

In order to develop disaggregate models of station choice, information is required about individual trips via the rail network. This data needs to include the access station where a train was first boarded, the final egress station, and the ultimate origin of the trip, such as home or work address. The data must also be at a spatial resolution that is sufficient for the variability in explanatory factors between individual decision makers, such as access distance, to be revealed. For UK-based research, the unit postcode area boundary is probably the maximum spatial unit of address aggregation appropriate for this type of analysis. Suitable trip data was obtained from an on-train survey carried out on the Cardiff Central to Rhymney line in South Wales (Blainey 2009). This was an ideal dataset for developing data processing techniques and estimating initial choice models, as it was already in a "clean" state and of a manageable size, consisting of 513 responses. Observations without data for both the origin and destination parts of the trip were removed, reducing the number of observations to 284.

3.2 Supporting data

Details on railway stations in GB were obtained from the National Public Transport Access Node (NaPTAN) database. The centroid of each unit postcode in the UK was obtained from the Ordnance Survey Code-Point dataset which was downloaded from EDINA Digimap (Code-Point 2015). Polygons representing the area covered by each postcode, used to visualise station catchments, were obtained from the Ordnance Survey Code-Point with Polygons dataset which was downloaded from EDINA Digimap (Code-Point with Polygons 2015).

3.3 Developing a routable multi-modal network

As the access journey is such an important factor influencing station choice, a key objective of this research was to generate realistic representations of the access journey made by survey respondents. This required a multimodal network that could generate routes for a range of motorised and non-motorised transport modes.

3.3.1 Selecting a suitable routing tool

A review of commercial and open source tools identified three candidates: Google Maps API, Visography TRACC and OpenTripPlanner (OTP). The Google Maps API limits the number of API calls and has restrictive usage conditions. It is also limited to current timetables and it is not possible to query historic timetable data to match the date of origin-destination surveys, nor to add new public transport routes, adjust frequencies or add station stops to assess the impact of potential service changes. Visography TRACC is a commercial application that can import the standard UK public transport data formats: NaPTAN, TransXChange, and ATCO CIF. However, given that the UK public transport data is freely available under open data initiatives, a solution that is not reliant on commercial software was considered preferable. OTP is an open-source and cross-platform multi-modal route planner written in JAVA that uses imported OpenStreetMap (OSM) data for routing on the street and path network and supports multi-agency public transport routing through imported GTFS feeds. OTP has a web front-end and a sophisticated API, and was considered the most promising platform.

3.3.2 Building the multi-modal network

As OTP has a high random access memory (RAM) requirement when graph building[1], this stage was carried out on a Microsoft Azure Linux cloud server with 56 GB of RAM. The graph was then transferred to a local server for normal operation of the trip planner. The initial graph build included OSM data for GB obtained from Geofabrik[2] and a GTFS feed for GB National Rail services[3]. Bus timetable data for Wales was obtained from the Traveline National Dataset (TNDS) in TransXChange format, a UK XML standard, and an attempt was made to convert this to GTFS using the open source TransXChange2GTFS converter[4]. This failed when processing most of the TNDS XML files, despite the files passing validation in the official TransXChange Publisher tool, and it was rejected as a plausible solution. The only alternative was Visography TRACC, which is able to import TransXChange files and export a GTFS feed.

3.4 Deriving explanatory variables

3.4.1 Access journey

Measures of the access journey, from origin postcode centroid to station, were obtained by running an R script to query the OTP API, processing the JSON response and then writing the results to a database table. A set of functions were developed to query the OTP API, and these are the beginnings of an API wrapper for OTP which has the potential to be released as an R package in the future. For the bus time variable, which consists of walk time, on-bus time and waiting time (in the case of transfers), a desired trip start time of 09:00 on Monday 5 October 2015 was set, which corresponds with the Rhymney Line survey which was carried out on weekdays in early October. Several other parameters were set for the bus trip, including a “soft” maximum walk distance (for walk to and from the bus stop) of 1600m, a minimum time to allow for transfer between buses of 10 minutes, and a walk reluctance parameter that ensured a realistic balance between the walk and bus components of the multi-modal trip.

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 ’directness’ is obtained by dividing the drive distance from the trip origin to the station by the straight line (euclidean) distance, with the value of the ratio increasing from one as the route becomes more circuitous and deviates from the straight line.

3.4.2 Station facilities

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 and the XML response was processed using an R script and the results written to a database table. The variables recorded were: car park spaces (number), station CCTV (y/n), ticket machine (y/n), waiting room (y/n), station buffet (y/n), toilets (y/n), cycle storage (y/n), taxi rank (y/n), bus services available (y/n), and staffing level (unstaffed, part-time, full-time).

3.4.3 Train journey

For every unique origin station:destination station pair, a single train journey itinerary was obtained by querying the OTP API. The JSON response was processed and required variables written to a database table. A minimum transfer time of six minutes was specified, corresponding to the suggested connection time for a medium-sized interchange station. The desired trip start time was set to 09:00 on Monday 5 October 2015, which corresponds with the Rhymney Line survey. The variables used in the choice models are the journey duration, consisting of on-train time and waiting time (in the case of transfers); the number of

[1] The trip planner graph specifies every location in the region covered and how to travel between them. It is compiled from the OSM and GTFS data.
[2] see http://download.geofrabrik.de
[3] see http://www.gbrail.info
[4] see https://code.google.com/p/googletransitdatafeed/wiki/GoogleTransitDataFeed

transfers; and the difference between the desired departure time (09:00) and the actual departure time. The latter measure should, to an extent, capture effects related to frequency and headway.

Fares data is available direct from ATOC, but it is provided in a large number of flat text files that would have required a considerable investment of time and effort to produce any meaningful data from. Fortuitously, the independent BR Fares website[5] provides a fares lookup service, and permission to use the associated API was obtained. The API was queried using an R script, the JSON response processed, and required variables were written to the database. The cheapest off-peak and anytime return fares for each unique origin and destination station pair were extracted. As off-peak return fares are not an available ticket option on the Rhymney line, only the anytime return fare was used in model estimation.

3.4.4 Land use and built environment

A land-use mix measure was generated using the Ordnance Survey Points of Interest dataset, obtained for the study region from the EDINA Digimap service (Points of Interest 2015). The number of points of interest for each of the nine top level classifications within a euclidean distance of 400m of each station were counted using a spatial query. The Herfindahl-Hirschman Index (HHI) was then calculated for each station. It indicates the extent to which one land use type dominates in an area, and is influenced by the number of land uses and their relative size. It is calculated by squaring the percentage share of each classification, and then summing the squares. In this study, with a possble nine classifications, the HHI can range from 1,111, where each is equally represented in an area, to 10,000 where only a single classification is present.

4 Choice models

4.1 Defining choice sets

It was decided, given the spatial nature of the choice alternatives, to define a separate choice set for each origin postcode based on selecting the nearest *n* stations to that postcode. To establish an appropriate value for *n*, for each unique origin postcode in the survey the 50 nearest stations by euclidean distance were identified, using the efficient PostGIS indexed nearest neighbour query. For each postcode:station pair the drive distance was obtained from an API call to OTP, and for each postcode the stations were then ranked by drive distance. The 15 nearest stations by drive distance account for all the observed choice, with the nearest 10 stations accounting for 98.94% of observations. It was therefore decided to use the nearest 10 stations to each postcode as the choice set and these were added to a database table which was then populated with various access journey variables obtained from OTP API calls as described in Section 3.4.1. Populating the variables for each origin postcode, rather than for each observation, eliminates duplication and minimises the number of API calls required to populate the variables.

4.2 Model estimation

Prior to estimating any models a correlation matrix was produced. Apart from expected high correlation between the various access distance measures, there is a very high correlation between several of the station facility variables. For example, there is a correlation of 1 between full time staffed stations and the presence of a station buffet; and a correlation of 0.95 between toilets and waiting room. There is also a very high correlation (0.98) between the presence of station CCTV and CCTV covering cycle parking areas, as might be expected. There is a moderate to high correlation between the duration of the train leg and the fare paid (0.69), and between the number of car parking spaces and the presence of a taxi rank (0.78). In view of the very high correlation between many of the station facility variables, it was decided to include the staffing level categorical variable as the main station facilities measure. Summary statistics for the choice dataset used in the models that follow are shown in Tables 1 and 2. As the car parking spaces parameter was only estimated against observations that access the station by car, its summary statistics only relate to

[5] see http://www.brfares.com/

those observations. As there are so few observations for bicycle or taxi as access mode, no variables specifically relating to those modes were included in the models.

4.2.1 Models with basic choice sets

A series of models were estimated using the R package mclogit (Elff 2014), with the choice sets as defined in Section 4.1 and using an additive linear utility function. Explanatory variables were entered in a manual forward selection procedure, and the results are shown in Table 3. In Models 1 to 3, access drive distance, staffing level dummies (with full-time excluded as the reference), and train journey time are entered into the models. All the variable parameters are significant at the 99.9% confidence level and all have a negative effect as would intuitively be expected. The staffing level parameters have to be interpreted with reference to the full-time staffing level, and the results indicate that the utility of a station is lower for part-time or unstaffed stations. This would be expected, especially as the level of staffing is also an indicator of a range of station facilities.

	min	max	mean	var	sd	n
cardist(km)	0.08	14.70	5.05	9.17	3.03	2800
train_time(mins)	2.00	144.00	34.87	364.89	19.10	2800
fare(£)	2.40	22.40	6.21	2.53	1.59	2800
transfers	0.00	1.00	0.33	0.22	0.47	2800
headway(mins)	0.00	55.00	15.15	168.38	12.98	2800
carspaces	0.00	402.00	36.96	3279.04	57.26	519

Table 1: Summary statistics for numeric variables - Models 1-10

Access mode	Staffing level	CCTV
NA: 20	fullTime : 221	FALSE: 611
Walk: 1981	partTime : 521	TRUE: 2189
Bus: 240	unstaffed: 2058	NA's: 0
Car(driver): 320		
Car(passenger): 199		
Bicycle: 20		
Taxi: 20		

Table 2: Summary statistics for logical and categorical variables - Models 1-10

The fare variable is introduced in Model 4, and while this is significant at the 99% level and results in a small but significant (at 95% level) reduction in the log likelihood (LL) function[6], the positive effect on utility is counter-intuitive. It would be expected that given a choice of stations, all else being equal, an individual would choose the station with the lowest fare. Fare has a moderate to high positive correlation with train time (0.69), and correlation between two variables can result in the parameter for one of the variables having the wrong sign. To confirm this another model was run (4a) with train time removed, and the parameter for fare is significant and negative in this model[7]. As fare results in a much smaller reduction in LL than train time, it is removed from the next model (5) and replaced with the nearest station dummy variable. As other studies have found (see, for example, Adcock (1997); Fan et al. (1993)), this variable does improve the model. When the number of transfers for the train journey was added to the model (6), an extremely high standard error was reported for the parameter (1,016), and further investigation revealed that this model is invalid due to complete separation, with all chosen stations having zero transfers. Previous studies have found the number of transfers to have a negative effect on station utility, but the nature of the data in this study, with both chosen origin and destination stations limited to a single rail line, has resulted in no journeys involving transfers. In Model 7, the transfers variable was removed and replaced with the headway measure, the difference in minutes between the desired departure time (09:00) and the actual departure time. The parameter is significant at the 95% level and has the expected sign, and the small reduction in the LL (compared to

[6] Calculated using log likelihood ratio test.
[7] In view of this it may be preferable to use a generalised journey time variable that combines both fare and train time.

Model 5) is also significant at the 95% level. However, once the CCTV variable is added in Model 8, headway is no longer significant, while the presence of CCTV has a strong and significant positive effect on station utility, and also has little impact on the other parameters. This result is surprising as it has not been included in previous studies of station choice. However, the main source of advice on passenger demand forecasting for the rail industry in GB, the Passenger Demand Forecasting Handbook (ATOC 2013), does recommend a demand uplift when adding CCTV to a station of 8% for business and leisure trips and 5% for commuter trips. In Model 9, headway is removed and replaced with the car park spaces variable. As the availability of car parking spaces is only relevant to travellers using a car as access mode, the car park spaces variable was interacted with a dummy variable indicating whether either of the car access modes (driver or passenger) was used.

	Drive distance	Staffing level		Train time	Fare	Nearest station	Transfers	Headway	CCTV	Car spaces	logLik	Adj R^2
		PT	None									
1	-1.00***										-349	0.46
2	-0.93***	-3.40***	-4.50***								-249	0.62
3	-1.10***	-2.20***	-2.70***	-0.21***							-212	0.67
4	-1.10***	-1.30*	-1.90***	-0.25***	0.73***						-209	0.68
4a	-0.97***	-3.50***	-4.50***		-0.41**						-247	0.62
5	-0.82***	-2.20***	-2.80***	-0.21***		0.98***					-203	0.69
6	-0.79***	-1.60**	-2.30***	-0.18***		0.98***	-15.00				-201	0.69
7	-0.84***	-2.10***	-2.60***	-0.20***		0.94***		-0.03**			-201	0.69
8	-0.82***	-2.50***	-2.60***	-0.20***		0.98***		-0.02	1.40***		-195	0.70
9	-0.81***	-2.60***	-2.70***	-0.20***		1.00***			1.40***	0.002	-196	0.70
10	-0.81***	-2.60***	-2.70***	-0.20***		0.99***			1.40***		-196	0.70

Table 3: Model results - basic choice sets

The utility function for this model, for individual *i* choosing station *k*, is therefore:

$$V_{ik} = \alpha D_k + \beta Spt_k + \gamma Sno_k + \delta T_k + \varepsilon Ns_k + \zeta C_k + \eta(Dcar_i \times Ps_k) \quad (2)$$

where *D* is drive distance, *Spt* is staffing level (part time), *Sno* is staffing level (unstaffed), *T* is train time, *Ns* is nearest station, *C* is CCTV, *Dcar* is a dummy variable with value 1 if individual *i* uses the car as access mode, and zero otherwise; *Ps* is the number of parking spaces and α, β, γ, δ, ε, ζ, and η are parameters to be estimated. The parameter for car park spaces was not significant in the model. Car parking is the most common station facility attribute considered in prior studies, and in most cases the presence of a car park or the number of parking spaces has a positive effect on station choice, although there have been conflicting results and counter-intuitive coefficient signs in some cases. It may be that in this study car parking is not a limiting factor. Of the 52 observations where the station was accessed by car, only three used one of the central Cardiff stations where parking is likely to be difficult. It is also possible that the number of spaces is not the most appropriate measure. The presence or not of a car park, the availability of spaces or on-street parking and level of fee may all be relevant factors. Car park spaces is removed from Model 10, which is the final and best fitting of the models using the basic choice sets.

4.2.2 Models with threshold-based choice sets and access mode specific parameters

A feature of logit models is that an alternative can never have a probability of zero, and if an alternative has no realistic prospect of being chosen it can be excluded from the choice set (Train 2009). If an individual has chosen to walk to a station, then there must be a cut-off distance at which a station is no longer considered a feasible alternative; and if bus is used as the access mode, the choice of stations should be restricted to those that can realistically be accessed by bus from the individual's trip origin. A further potential limitation of the earlier models is the assumption that the negative effect on utility of increasing access distance is the same irrespective of access mode. If only a single parameter is estimated for access distance this will represent an average effect on utility across the different access modes. In reality, a 1km increase in access distance would be expected to have a greater negative effect on utility for walking or cycling modes than it would for driving or bus modes. The basic models also assume that the access distance is the same for all modes, when this is unlikely

to be the case. Pedestrians can use off-road pathways and are not governed by restrictions such as one-way working, while buses are likely to take a longer more circuitous route than a car driver. Furthermore, it could be argued that access distance is not the most appropriate measure, as it is the access time that is important, rather than the distance which may not be known to the individual when making a choice. To address these potential limitations, with the view to producing more accurate predictive models, the following adjustments were made to the dataset:

- Observations where the access mode was not recorded (two observations) and where bicycle or taxi/minicab was the access mode (too few observations, two for each mode) were removed.
- For observations where the access mode is bus, stations were removed from the choice set where no bus route was available (i.e. not returned by OTP), or where the bus walk time was equal to the total bus duration (i.e. where OTP advised walking to the station rather than catching a bus due to its close proximity)
- The survey data shows that the maximum walk time to access a chosen station was 97.25 minutes. This appears to be an outlier, as the next highest is 43.77 minutes, a more reasonable figure. Therefore, 45 minutes was taken as the maximum walk time, and any stations where the access time exceeded this were removed from the choice set for those individuals that used walk mode. The outlier was also removed.

The amendments reduced the number of individuals in the dataset from 281 to 274, and reduced the average choice set size from 10 alternatives to six. Summary statistics for the new dataset are shown in Table 4. As the mode specific access time parameters were only estimated against observations that accessed the station using that mode, the summary statistics for those variables only relate to those observations.

	min	max	mean	var	sd	n
cardist(km)	0.08	14.70	3.96	8.82	2.97	1659
train_time(mins)	2.00	133.00	34.40	282.94	16.82	1659
poihhi	1400.00	3066.67	2016.76	134089.17	366.18	1659
directness	0.81	9.06	1.68	0.54	0.74	1659
time_walk(mins)	1.60	44.93	24.17	118.31	10.88	903
time_bus(mins)	1.05	105.22	34.48	428.39	20.70	237
time_car_p(mins)	0.57	26.75	12.07	26.03	5.10	199
time_car_d(mins)	0.30	25.13	12.22	25.26	5.03	320

Table 4: Summary statistics for numeric variables - models 11-14

A series of models were estimated using the amended dataset, and the results are shown in Table 5. It should be noted that the measures of model fit in these models, LL and McFadden's adjusted R^2, cannot be compared directly with those reported in Table 3. As a reference point, Model 11 was run with the same parameters as Model 10. The change in the choice sets does not have a major impact on the model, but the negative effect of drive distance is reduced somewhat, and the positive effects of nearest station and CCTV are increased. In Model 12 access mode specific parameters are estimated by including an access time variable for each mode (obtained from OTP) which is interacted with a dummy variable for that mode. The utility function for this model, for individual *i* choosing station *k*, is as follows:

$$V_{ik} = \sum_{m=1}^{4} \alpha_{mtime}(Dmode_{im} \times time_{km}) + \beta Spt_k + \gamma Sno_k + \delta T_k + \varepsilon Ns_k + \zeta C_k \quad (3)$$

where $Dmode_{im}$ is 1 if individual *i* used access mode *m*, and zero otherwise; $time_{km}$ is the access time to alternative *k* using mode *m*; and α_{mtime} is the parameter to be estimated for access time by mode *m*. This model is an improvement over the reference model (11), with a significantly lower LL, suggesting that this is also an improvement over Model 10. The access time parameters for the two car modes are similar, as might be expected. The negative effect on utility when walking to the station is less than half the size of the car effect, which may at first appear counter intuitive. However, these parameters represent the change in utility for each additional minute of access time, and the distance covered by car within a

minute will be considerably more than that covered on foot. Assuming an average walk speed of 3mph and an average drive speed of 40mph, the parameters indicate that an additional half mile of access distance reduces the utility of a station for car access by 0.22 units, and reduces the utility of a station for walk access by 1.3 units. The larger effect for walk access is what would be intuitively expected. Nevertheless, the results indicate that one minute of extra travel time is a greater cost to car drivers and passengers than to bus passengers or pedestrians. The HHI is introduced in Model 13, and the parameter estimate is negative and significant, at the 90% level. The LL is reduced from -159 to -157 and this difference is significant at the 95% level. This suggests that a station is more likely to be chosen if it is surrounded by greater land use diversity. The parameter is very small, but as the HHI ranges from 1,111 - 10,000 this is potentially more important than it may at first seem. At 10,000 (where only one point of interest group is represented) utility would be reduced by 10 units. In the data HHI ranges from 1400 to 3067, representing a potential effect on utility between -1.4 and -3.07. In the final model (14) the directness measure is added. The parameter is significant and improves the model, but the direction of the effect was not expected. It was thought that a more circuitous access journey would make a station less attractive than a station with a more direct route, especially for walk mode, but the model suggests the opposite (a higher directness value signifies a less direct route). A plot of the directness measure against station access drive distance revealed that shorter access journeys tend to be less direct. This could be due to shorter journeys being confined to urban areas around trip origins, where the road network is dense and the layout more complex, while longer journeys are likely to include roads in non-built up areas which have longer and straighter stretches. This variable could therefore be responding to the preference for nearer stations, rather than a desire for less direct routes, although the correlation between directness and car distance is fairly low (-0.21***). It should also be noted that a one unit change in the directness measure will represent a considerable deviation from a straight line (the standard deviation for directness in the dataset is only 0.74).

	Drive distance	Access time (car drive)	Access time (car pas)	Access time (bus)	Access time (walk)	Staff level		Train time	Nearest station	CCTV	HHI	directness	logLik	AdjR2
						PT	None							
11	-0.60***					-2.70***	-2.60***	-0.21***	1.10***	1.70***			-178	0.61
12		-0.29***	-0.32***	-0.18***	-0.13***	-3.00***	-3.00***	-0.20***	0.78***	1.80***			-159	0.65
13		-0.31***	-0.33***	-0.19***	-0.13***	-3.10***	-2.70***	-0.22***	0.78***	1.80***	-0.001*		-157	0.65
14		-0.28***	-0.30***	-0.18***	-0.11***	-3.00***	-2.60***	-0.24***	1.10***	2.00***	-0.001*	0.40***	-153	0.66

Table 5: Model results - threshold based choice sets and access mode specific parameters

5 Generating probabilistic catchments

As the purpose of developing a predictive model of station choice is to enable probabilistic catchments to be incorporated into station demand models, it was considered important to assess the practicability of generating such catchments using the results of this early modelling work. To reduce the complexity and the amount of data processing involved, the best performing of the basic choice set models, Model 10, was selected to generate the catchments. The utility function with the parameters estimated in Model 10 is as follows:

$$V_{ik} = (-0.81 \times D_k) + (-2.6 \times Spt_k) + (-2.7 \times Sno_k) + (-0.2 \times T_k) + (0.99 \times Ns_k) + (1.4 \times C_k) \quad (4)$$

As the utility function contains the time of the train journey, the probabilistic catchment will depend upon the destination station, and each station will have a different catchment for each destination. As an example, the process of generating the catchments for Ystrad Mynach and surrounding stations on the Rhymney line, using Cardiff Central as the destination station, involved the following steps:

- For each unit postcode in the area of interest, the 20 nearest stations (by euclidean distance) were identified. The drive distance from each postcode to each station was then obtained from OTP, and the stations then ranked by drive distance. The top 10 ranked stations were placed in a database table and the access journey variables obtained by querying the OTP API as described in Section 3.4.1.
- A train leg table was generated for each unique origin station:Cardiff Central pair, and populated with train time and fare variables as described in Section 3.4.3.

- A separate probability table was then generated. This pulls together the explanatory variables for each origin postcode:origin station pair and calculates the probability of each alternative station being chosen for each postcode (using the standard multinomial logit probability equation and the utility function described in Formula 4).
- The probabilistic catchments for a specific station were generated using a database view, which pulls data from the probability table and the Code-Point Polygons table. These were then visualised in QGIS.

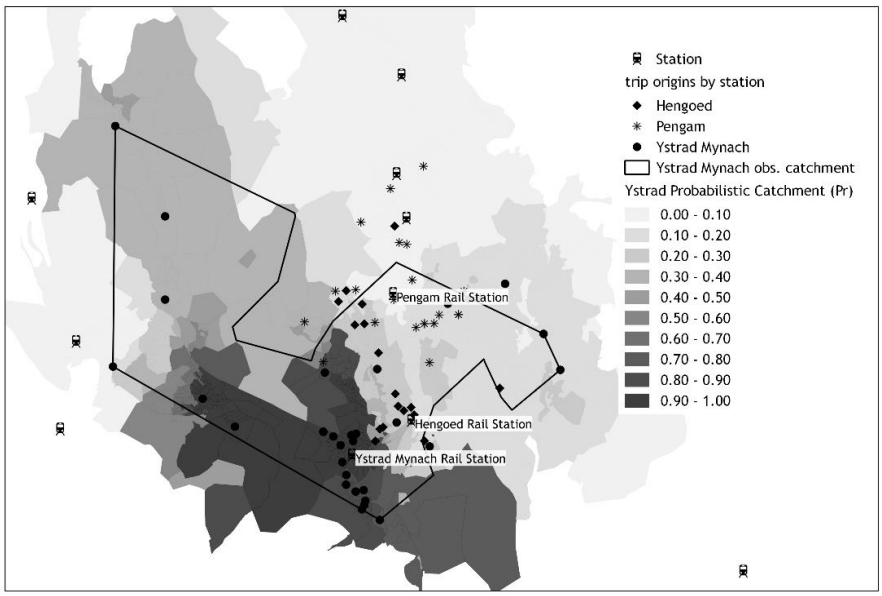

Figure 1: Probabilistic catchment for Ystrad Mynach station to Cardiff Central and observed catchment (all destinations)

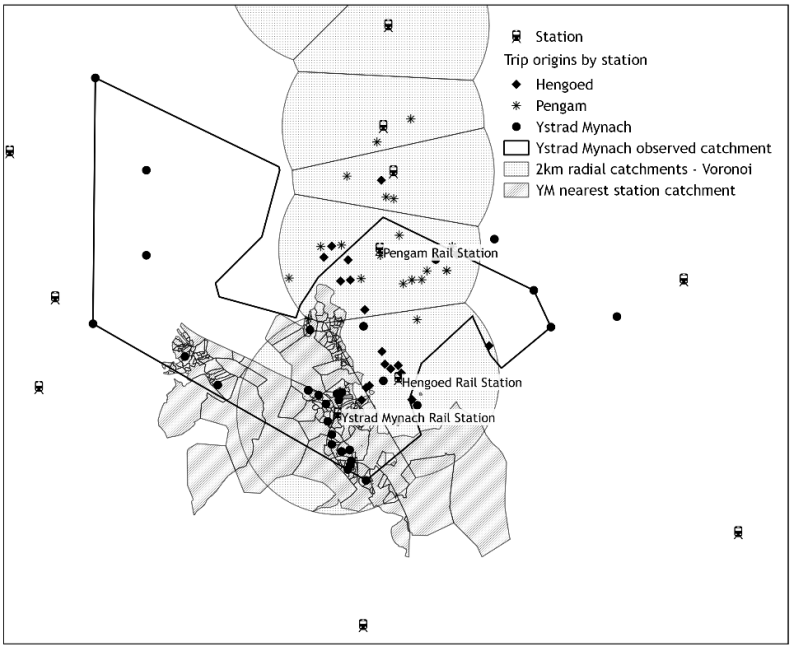

Figure 2: Nearest station, 2km radial and observed catchments for Ystrad Mynach station

Figure 1 shows the probabilistic catchment for Ystrad Mynach rail station (to Cardiff Central), along with its observed catchment (all destinations), and the trip origins for Ystrad Mynach, Hengoed and Pengam rail stations in the full survey dataset. Figure 2 shows catchments for Ystrad Mynach based on assigning unit postcodes to their nearest station (by drive

distance), and based on a 2km radial buffer around the station. As the buffers of nearby stations overlap with one another, Voronoi polygons have been used to generate a discrete catchment for each station. Whilst the very high probability postcodes match well with the nearest station catchment, the probabilistic catchment extends further to the north west to postcodes which have a nearer station. The postcodes here have a 30 - 40 percent probability of choosing Ystrad Mynach, and correspond well with the observed catchment. The on-train survey did not include the stations on the rail line to the west, so there is no data on trips that may have originated in this area and chosen one of those stations. The probabilistic catchment also extends to the north east of Ystrad Mynach station with probabilities in the 10 - 20 percent range. This also corresponds well with the observed catchment and the effect of "competition" from Hengoed and Pengam stations. The catchment derived from the 2km radial buffer captures many of the highest probability postcodes, but not all of them, and, like the nearest station catchment, misses many of the observed trip origins.

6 Conclusions and future work

This paper has shown that it is possible to calibrate a relatively simple station choice model that fits the observed data well. The estimated parameters can be used to generate probabilistic station catchments that are a realistic representation of observed catchments and perform better than the deterministic station catchments used in conventional aggregate demand models. This paper has also described a set of robust and reproducible methods for deriving explanatory variables using open source data and software tools. Future work will seek to apply these methods to much larger datasets, test additional variables and variable forms (for example, a generalised journey time measure), and develop more sophisticated choice models. A particular issue with the multinomial logit models described in this paper, is that they suffer from proportional substitution behaviour, and if a proposed new station is added to the choice set the probability of all existing stations will be reduced by the same percentage. However, it is more likely that a new station will have a greater effect on the probability of nearer stations. Ensuring a realistic representation of abstraction from pre-existing stations is an important consideration, and is a significant limitation of the existing aggregate demand models. Failure to account for abstraction can result in a new station having a smaller net effect on rail demand than predicted, and in some circumstances this could undermine the business case for the station. Once a suitable station choice model has been calibrated, a key component of future work will be to incorporate probabilistic catchments into the aggregate rail demand models. This will represent a novel application of station choice modelling, and should allow the demand impacts of opening new stations and of making amendments to existing rail services to be more accurately assessed.

Acknowledgements

The authors wish to thank Paul Kelly for permission to use the brfares.com API, and Dan Saunders at Basemap Ltd for providing a TRACC educational license. The work reported here forms part of a PhD funded by EPSRC DTG Grant EP/M50662X/1. Code.Point and Code.Point Polygons © Crown Copyright and Database Right 2015. Ordnance Survey (Digimap Licence). This work uses data licensed from PointX © Database Right/Copyright 2015 and public sector information licensed under the Open Government Licence v3.0.

References

Adcock, S.J., 1997. A Passenger Station Choice Model for the British Rail Network. In Proceedings of European Transport Conference PTRC. pp. 141–146.

ATOC, 2013. Passenger Demand Forecasting Handbook v5.1.

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

Blainey, S. & Evens, S., 2011. Local station catchments: reconciling theory with reality. In AET European Transport Conference.

Blainey, S.P., 2009. Forecasting the use of new local railway stations and services using GIS.

Blainey, S.P. & Preston, J.M., 2010. Modelling local rail demand in South Wales. Transportation Planning and Technology, 33(1), pp.55–73.

Chen, C. et al., 2014. Development of a Conceptual Framework for Modeling Train Station Choice Under Uncertainty for Park-and-Ride Users. In Transportation Research Board 93rd Annual Meeting.

Debrezion, G., Pels, E. & Rietveld, P., 2009. Modelling the joint access mode and railway station choice. Transportation Research Part E: logistics and transportation review, 45(1), pp.270–283.

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

Elff, M., 2014. mclogit: Mixed Conditional Logit, Available at: http://CRAN.R-project.org/package=mclogit.

Fan, K.-S., Miller, E.J. & Badoe, D., 1993. Modeling rail access mode and station choice. Transportation Research Record, 1413, pp.49–59.

Givoni, M. & Rietveld, P., 2014. Do cities deserve more railway stations? The choice of a departure railway station in a multiple-station region. Journal of Transport Geography, 36, pp.89–97.

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

Mahmoud, M.S., Eng, P. & Shalaby, A., 2014. Park-and-Ride Access Station Choice Model for Cross-Regional Commuter Trips in the Greater Toronto and Hamilton Area (GTHA). In Transportation Research Board 93rd Annual Meeting.

Code-Point [CSV], Coverage: GB, Updated: April 2015, Ordnance Survey (GB), Using: EDINA Digimap Service, http://digimap.edina.ac.uk, Downloaded: August 2015.

Code-Point with Polygons [Shapefile], Coverage: GB, Updated: May 2015, Ordnance Survey (GB), Using: EDINA Digimap Service, http://digimap.edina.ac.uk, Downloaded: October 2015.

Points of Interest [CSV], Scale 1:1250, Items: 97398, Updated: September 2015, Ordnance Survey (GB), Using: EDINA Digimap Service, http://digimap.edina.ac.uk, Downloaded: October 2015.

Preston, J. & Aldridge, D., 1991. Greater Manchester PTE New Railway Station Demand Prediction Model, Institute of Transport Studies, University of Leeds.

Rail Delivery Group, 2014. GB rail: dataset on financial and operational performance 1997-98 - 2012-13

Railfuture, 2015. New Stations, Available at: http://www.railfuture.org.uk/New+stations.

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.