Logistic Regression Models for the Nearest Train The Nearest Station Choice: A Comparison of Captive and Non-captive stations

Changying Shao[a], Jianghong (Cecilia) Xia[a], Ting (Grace) Lin[a], Konstadinos G. Goulias[b] and Chunmei Chen[a]

[a]Department of Spatial Sciences, Curtin University, Kent St Bentley, Western Australia
[d]Department of Geography, University of California, Santa Barbara, USA

Abstract: *We usually assume that each commuter is an efficient traveller, which means they maximize trip utility. From a spatial optimization perspective, a commuter might therefore choose the nearest station to reach their destination. However, based on a survey at seven train stations in Perth, Western Australia, only between 30 and 80 percent of commuters choose the nearest station to their origin. Many factors could affect this travel behaviour. From a logistic regression model, five factors were found to be significant (p-value < 0.05), indicating that commuters are more likely to choose the non-nearest station for longer commutes, while traveling further away from origins and destination if the chosen stations are at, or near, the end of train lines (captive stations). If the chosen stations are along the train line (non-captive stations), longer distance, longer wait times and lower costs from the chosen station to a destination were found to be significant. The results of the study are important for public transport policy makers to understand transit choice behaviours. Therefore public transport policies such as adjustments of travel fees and improving station service and facilities, could be developed.*

Keywords: Logistic regression model, the nearest station choice, captivity

1. INTRODUCTION

Why don't commuters always choose the nearest train station to their origin, such as home, to reach their destination? This has been an interesting question for transport geographers and planners. In this case, the nearest station means the station is located nearest to the origin based on network distance. The assumption is that distance is one of the key variables for commuters' station choice. However, based on our survey conducted in Perth, Western Australia from July 31 to August 1, 2012, the probability of the nearest station choice varies by the location of stations (Desfor 1975). For example, the station at the end of railway lines is more likely to be chosen as the nearest station than a station somewhere along the line. For example (see Figure 1), the probability of choosing Midland station, at the end of the train line, as the nearest station is 68.8%, while for Cannington (located at the middle of the train line), it is only 26.9%, which means 73.1% people didn't choose the nearest station to their origin, instead driving a longer distance to use Cannington station. Therefore, in addition to location and distance, other variables could be important for these choice behaviours and they are classified in three groups:

- The objective and latent characteristics of commuters bring demographic effects to choice models, such as age and gender. Nordlund and Westin (2013) identified that values, beliefs and age can influence train use decisions. For example, younger people are more likely to use trains than other age groups.
- The characteristics of stations also play an important role in station choice decisions. Stations with better accessibility, such as intermodal connectivity, higher train frequency, service quality and diverse land use are more likely to be chosen as a travel alternative (Debrezion, Pels, and Rietveld 2009, Givoni and Rietveld 2007, Brons, Givoni, and Rietveld 2009, Rietveld 2000).
- Trip characteristics, such as travel time, cost, trip direction (inbound or outbound) and motivation, can influence station choice (Desfor 1975, Boyce and School 1973).

Understanding variables affecting the nearest station choices is essential to railway planning, design and management. For example, by understanding the reasons why commuters do not choose the nearest stations, some drawback of stations could be identified and some intervention strategies can be developed to improve the railway service and facilities, which could ultimately encourage more public transit usage.

The aim of the paper is to develop a method for predicting the nearest station choice and identifying significant variables affecting commuters' nearest station choice using logistic regression models. Train stations are categorised into two types: captive and the non-captive stations (Beimborn, Greenwald, and Jin 2003b). A captive station can be defined as a station located at the end or near the end of a railway line and these have a location advantage to capture a bigger pool of transit users because of less competition from surrounding train stations. In contrast, a non-captive station is located at the middle of a railway line with more competition from the surrounding train stations and so has a smaller pool of transit users. According to Beimborn, Greenwald, and Jin (2003b), transit modal split models that do not consider captive conditions could underestimate variation in mode choice behaviours for captive users, whereas overestimating the use of transit for choice users. Similarly, the nearest train station choice model lacking captive conditions could underestimate the probability of captive station choice behaviours, while overestimating the nearest non-captive station choice behaviours. Therefore, for this project logistic regression models are used to model the nearest station choice behaviours separately for captive and non-captive stations. We used a set of complete and disaggregated travel data collected from an intercept survey, in Perth, Western Australia, and the station characteristics data obtained from the Western Australian Department of Transport, WA.

The next section reviews the literature on station choice behaviours. Section 3 presents a logistic model of the nearest station choice. Section 4 describes our case study of Perth including the study area, data collection and analysis results, and section 5 discusses the findings, contributions and limitations.

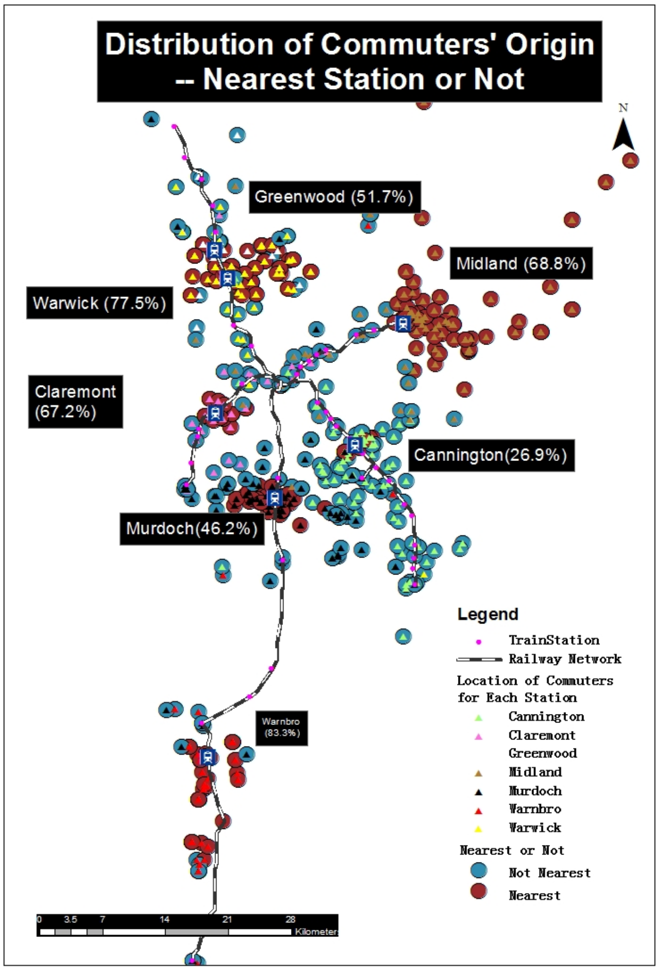

Figure 1 Distribution of Nearest Station Choice

2. LITERATURE REVIEW

Several studies have investigated the relationship between the proximity to rail stations and the propensity to use rail (Lindsey et al. 2010, Cervero 2006, 2007, Bernick and Cervero 1997, Beimborn, Greenwald, and Jin 2003a). Beimborn, Greenwald, and Jin (2003a) categorised the transit market as choice and captive users. Transit captive users do not own a car or have limited access to a car. They heavily rely on public transport and tend to live within walking distance of a transit stop. Automobile captive users must complete their trips using an automobile due to the inconvenience and inflexibility of transit services. Choice users have freedom to travel modes without these constraints. By integrating these constraints into a transit forecasting model, the decision making patterns of transit users can be better understood.

Cervero (2006) states that only 20% Californian commuters who live near a rail station actually travel that way, which could be due to continued decentralization of office buildings and the inflexibility and inconvenience of public transport (Cervero 2007, Lindsey et al. 2010). However, he also pointed out that U.S. residents, living within half mile circle of a transit stop, are four times more likely to use train services than commuters living between one-half and three-miles of a stop, and five to six times more likely than commuters living beyond the three mile circle (Cervero 2007). Based on Cervero's work, Lindsey et al. (2010) developed a method to understand the relationship between proximity to transit stops and ridership for commuter trips in Chicago in order to estimate energy consumption reduction if commuters whose origin/destination was within one mile of commuter rail stations could shift their travel model from a private car to public transport. However, they admitted that distance is only one factor affecting rail transit choice. Other factors such as the ease of accessing train stations, the price of parking, and parking availability are also important factors. Krygsman, Dijst, and Arentze (2004) discovered that if the distance to the station goes beyond a certain threshold, commuters will not take transit alternatives into consideration.

One of the early rail transit station choice models was developed by Kastrenakes (1988) in an effort to prepare a basis for forecasting railway travel in the New Jersey area. With origin-destination pair data, he analysed the choice process for a departure station and identified *location of station, access time, frequency of service and generalised cost* were uncorrelated factors affecting station choice for commuters. Then, Wardman and Whelan (1999) studied railway station choice for the London area and determined *parking availability and other station facilities* are also important factors and should be introduced into a choice model. Later, *travel time to station* and *access mode* were discovered as important factors affecting station choice by many authors (Davidson and Yang 1999, Fan, Miller, and Badoe 1993, Wardman and whelan 1999, Debrezion, Eric Pels, and Rietveld 2007). Recently, Debrezion, Pels, and Picard (2009) discovered that *rail service quality* and *accessibility to the station* are indispensable factors for station choice models during a study of Dutch railway users for access mode and departure railway stations.

The most common approach for modelling station choice is based on discrete choice framework such as logistic regression model, binary logit model, multinomial logit model, cross-nested logit model and nested logit model for station choice (Mcfadden 1974, Hensher, Rose, and Greene 2005, Train 2002, Tversky 1972, Wardman, Lythgoe, and Whelan 2007, Cervero 2007, Debrezion, Pels, and Rietveld 2009). This paper focuses on understanding *whether* the chosen station is the nearest station or not and *why* these choices were made. In other words, if distance is not a dominating optimising factor for a station choice, and we think travellers are trip utility maximisers, then what other factors could be included in their choice utility? Therefore, as a first approximation we estimate binary choice using logistic regression to model the nearest station choice process and identify significant determinants.

3. METHODS

3.1 Study Area

The study area of this paper is metropolitan Perth, the state capital of Western Australia and the fourth largest Australian city, after Sydney, Melbourne and Brisbane. In Perth, there are five train lines (Armadale, Fremantle, Joondalup, Midland and Mandurah lines), with one spur line (the Thornlie line), giving an overall starlike shape to this system (see Figure 2). The total railway is 173.1 kilometres in length and covers 69 train stations (Department of Infrastructure and Transport 2012) .

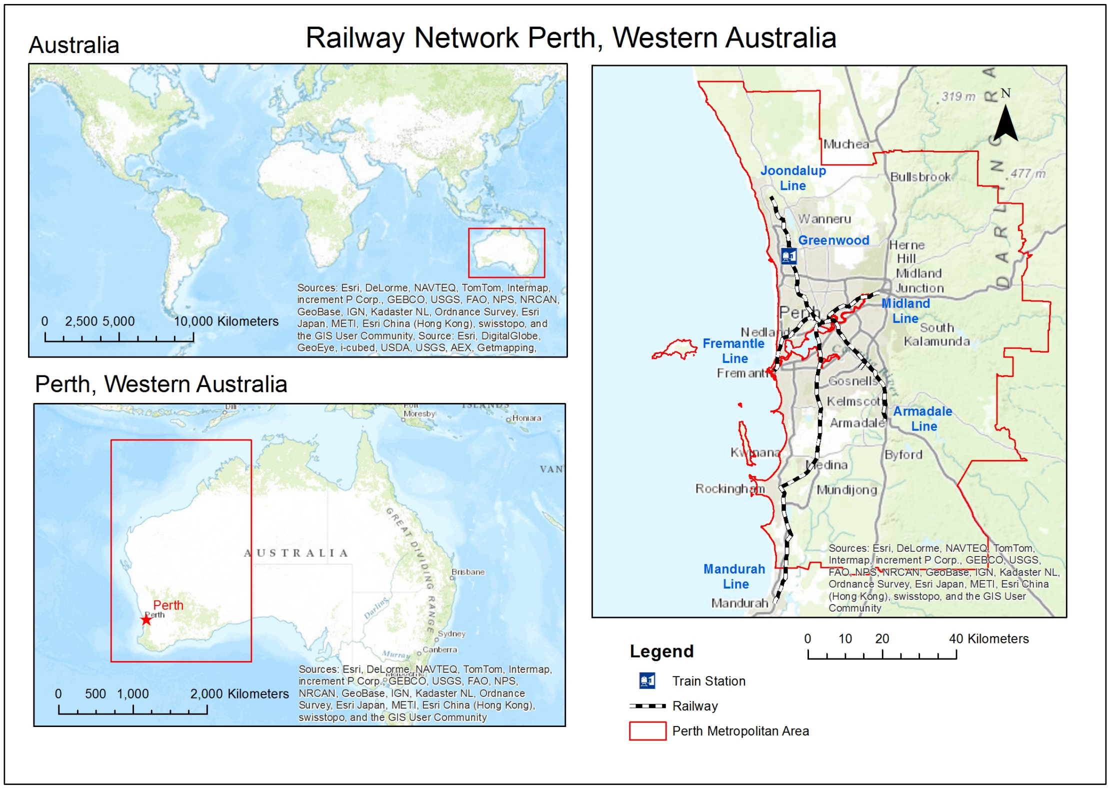
Railway Network Perth, Western Australia
Australia
0 2,500 5,000 10,000 Kilometers
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, and the GIS User Community, Source: Esri, DigitalGlobe, GeoEye, i-cubed, USDA, USGS, AEX, Getmapping,
Perth, Western Australia
AUSTRALIA
Perth
0 500 1,000 2,000 Kilometers
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, and the GIS User Community
N
Muchea
Joondalup Line
Bullsbrook
Wanneru
Greenwood
Herne Hill
Midland Junction
Midland Line
Perth
Nedla
Fremantle Line
Fremantle
South Kalamunda
Gosnells
Kelmscott
Armadale
Armadale Line
Kwinana
Medina
Byford
Rockingham
Mundijong
Mandurah Line
Mandurah
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, and the GIS User Community
0 10 20 40 Kilometers
Legend
Train Station
Railway
Perth Metropolitan Area

Figure 2. The railways of Perth Metropolitan Area

3.2 Data Collection and preparation

The data used in this paper were collected from two sources: a field survey and archival data from related government departments. Data such as geographical data (e.g., land use, Railway, Road, Strategic Transport Evaluation Model (STEM) zone) and previous survey data (e.g., 2008 P&R survey) were from the Department of Transport (DoT), Public Transport Authority (PTA) and The Department of Planning (DoP), Western Australia. Intercept surveys were conducted to collect commuters' trip diaries and their attitudes to station facilities and service quality. Seven train stations were selected— Warwick, Greenwood, Murdoch, Warnbro, Midland, Cannington, and Claremont— and the data were collected from 6:00AM to 6:00PM on July 31 and August 1, 2012. At these stations, respondents aged 18 and over were asked to complete a self-administered questionnaire. In total, 940 survey responses were used in this study.

Based on the research framework, 11 factors were used for station choice modelling. The four factors in green colour in Table 1 were derived directly from the survey data, while the others needed further calculation. Table 2 describes traveller and trip characteristics of the sample used in this study.

Table 1. The definition of variables used in the analysis

Name	Definition
Nearest station (NS)	The nearest train station to the commuter's origin based on network distance, the nearest station (1), Non-nearest station (0)
Distance	The shortest network distance between the origin and the chosen station
Travel time	A period of time that a commuter travelling from the origin to the chosen station calculated based on the distance and travel mode
Waiting time	A period of time between a commuter's arrival on the platform of the train station and boarding on the train
CostSD	Travel cost from the chosen station to the destination including fees such bus and train fares and driving cost per km.
CostOS	Travel cost from the origin to the chosen station including fees such as bus and train fares and driving cost per km
Gender	Male (0) or Female (1)
Age	Young (0), Middle (1) or Elderly (2)
Further-away station	The chosen train station is further away from the origin and the destination instead of between the origin and the destination, further away station (1) and non-further-away (0)

Inbound-out-trip	Inbound trip: trip towards the Perth CBD Area (1) Outbound trip: trip away from the Perth CBD Area (0)
Trip purpose	Home (0), Work (1), Education (2), Shop (3), Gym (4), Pub (4)
Travel mode	Park and ride (0), kiss-and-ride (1), Bus and ride (2), Walk and ride (3), Cycling and ride (4)

Table 2 Traveller and trip characteristics of the sample

Characteristics	*Percentage*			*Characteristics*	*Percentage*		
	All stations (833)	The captive (274)	The Non-captive 559)		All stations	The captive	The Non-captive
Gender,%				**Age, %**			
Female	50.42	54.38	48.48	Young	45.26	43.43	46.15
Male	44.78	41.97	46.15	Middle	36.25	35.40	36.67
Missing values	4.80	3.65	5.37	Old	12.73	15.33	11.45
Nearest station, %				Missing values	5.76	5.84	5.73
The NS	61.46	80.66	52.06	**Trip purpose, %**			
The non-NS	38.54	19.34	47.94	Home	9.96	4.01	12.88
Missing values	0	0	0	Work	51.50	45.62	54.38
Travel Mode, %				Education	16.69	19.71	15.21
Park and Ride	32.53	34.31	31.66	Personal business	7.44	8.03	7.16
Kiss and Ride	21.85	29.56	18.07	Shopping	3.24	4.01	2.86
Cycling and Ride	1.32	0	1.97	Social	0.84	1.09	0.72
Bus and Ride	29.65	26.64	31.13	Accompany anyone	1.68	4.74	0.18
Other (e.g.walk, taxi mode)	13.45	9.49	15.38	Attending events	0.24	0.36	0.18
Missing values	1.20	0	1.79	other	2.40	5.84	0.72
				Missing values	6.00	6.57	5.72

3.3 Data analysis framework:

The aim of this paper is to understand why the chosen station is or is not the nearest station by using logistic regression analysis. Figure 3 summarises the data analysis procedure. If the chosen station is the nearest station, the dependent variable is one, otherwise it is zero. The independent variables (Xs in Equation 1) are the characteristics of chosen stations, individual respondents and their trips. The form of the logistic regression equation is (Ralph B. D'Agostino, Sullivan, and Beiser 2006):

$$\ln\left\{ {}^{p}\!\Big/_{1-p} \right\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \qquad (1)$$

where p is the proportion of successes
X_1 to X_p are independent variables
β_0 is the intercept
β_i $(i = 1, \cdots, p)$ are the regression parameters
ε is the random error term

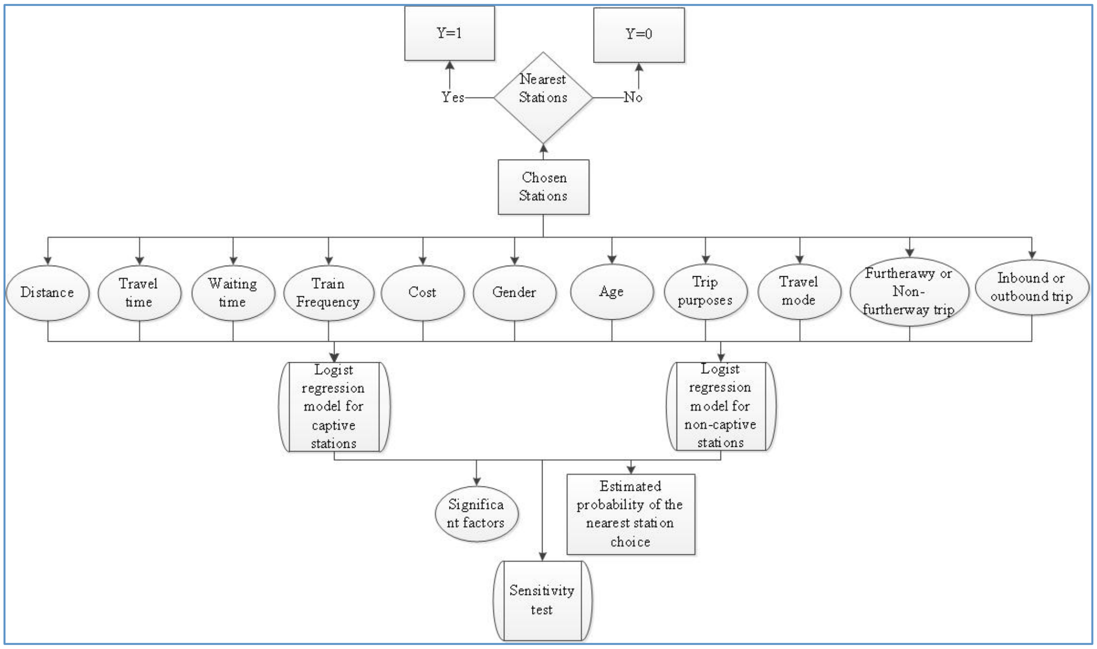

Figure 3 Data Analysis Framework

In this study, p represents the proportion choosing the nearest station. And so $1-p$ is the proportion not choosing the nearest station. ${}^{p}\!\big/_{1-p}$ is the "odds" of choosing the nearest station. $\ln\left\{ {}^{p}\!\big/_{1-p} \right\}$ is called the "log odds" or the "logit" of Y. Regression parameters, β_i $(i = 1, \cdots, p)$, reflect the change in the log odds (or logit) of Y relative to a one unit change in X_i. The independent variables X_i can be continuous or categorical variables. The logistic regression model was used for captive and non-captive stations respectively to identify the significant factors affecting the nearest station choice of train users. Sensitivity tests were also conducted based on established the models to understand the influences of independent variables, such as distance, on the nearest station choice.

In addition, we used R Package 'polycor' to compute a heterogenous correlation matrix, consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables.

4. DATA ANALYSIS RESULTS

4.1 Logistic regression overall models for all chosen stations

The correlation between *distance* and *travel time*, *travel mode* and *time*, *purpose* and *InboundOut* and *Waiting time* and *InboundOut* are 0.53, -0.45, -0.44 and -0.36 respectively (See Table 3). Therefore, variables, *travel time* and *InboundOut,* were removed before model selection. In addition, *travel purpose* was identified to have 95% confidence interval (0, inf). Therefore, it was not considered in the modelling process.

Table 3. Correlation matrix for all station

Correlation	Distance	Travel Time	Further away	InBound Out	Purpose	Gender	Age	Mode	Waiting Time	Cost OS	Cost SD
Distance	1	A	B	**B**	**B**	**B**	**B**	**B**	A	A	A
Travel Time	0.53*	1	**B**	**B**	**B**	**B**	**B**	**B**	A	A	A
Further away	-0.31*	-0.15*	1	**C**	**C**	**C**	**C**	**C**	**B**	**B**	**B**
InBound Out	0.08*	-0.09*	0.14	1	**C**	**C**	**C**	**C**	**B**	**B**	**B**
Purpose	0.07*	0.18*	-0.00*	-0.44*	1	**C**	**C**	**C**	**B**	**B**	**B**
Gender	0.05*	0.03*	-0.05	-0.03	0.02	1	**C**	**C**	**B**	**B**	**B**
Age	0.00*	-0.03*	0	0.15	-0.05*	-0.01	1	**C**	**B**	**B**	**B**
Mode	0.02*	-0.45*	-0.02	0.28	-0.26*	0.005	0.02	1	**B**	**B**	**B**
Waiting Time	0.08*	0.07*	-0.29*	-0.36*	0.1*	0.09*	-0.02*	-0.05*	1	A	A
CostOS	0.1*	0.1*	-0.03*	0.04*	0.06*	0.01*	-0.01*	-0.06*	0.05*	1	A
CostSD	-0.03*	-0.09*	-0.00*	-0.32*	0.04*	0.03*	0.04*	0.02*	0.01*	0.00*	1

* *P*-value <0.05. A:Pearson B:Polyserial C: Polychoric

Table 4 presents the best-fitting logistic regression model for predicting the nearest station choice for all seven stations. There are 833 records for all the stations (Table 2), but the sample size for this regression model is 732 with 101 missing records being removed for the purpose of the analysis. Three significant variables in the model were found to be statistically significant. The travel fee from a chosen station to a destination was generally the most influential one. The less the cost of travelling from a chosen station to a destination, the more likely a chosen station would be a non-nearest station. For example, a commuter could choose a transit station along the way towards their destination instead of using the nearest station in order to save ticket fares on trains. This suggests the effect of a big fare price jump between zones (Jansson and Angell 2012). The shortest network distance from an origin to a station was also found to have an important influence on the nearest station choice. The shorter the distance from origin to station, the more likely a chosen station is the nearest station. In addition, as revealed by the model, the shorter the waiting time at a chosen station, the more likely that station is a non-nearest station.

Table 4. Logistic regression models for all stations.

Variables	Estimate (SE)	OR (95%CI)	z value	Pr(>|z|)
(Intercept)	-0.14 (0.38)	0.87 (0.41-1.82)	-0.38	0.7
Distance (km)	-0.19 (0.02)	0.83 (0.8-0.86)	-10.52	< 2e-16 ***
Cost (station to destination)	0.41 (0.09)	1.51 (1.28-1.79)	4.8	1.62e-06 ***
Cost (origin to station)	0.07 (0.05)	1.08 (0.97-1.19)	1.44	0.15
Waiting time	0.05 (0.02)	1.05 (1.01-1.1)	2.52	0.01 *
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Null deviance: 972.03 on 731 degrees of freedom Residual deviance: 760.99 on 727 degrees of freedom (*p*-value =0.19) AIC: 770.99 Number of Fisher Scoring iterations: 5				

4.2 Logistic regression models for captive stations

According to Table 5, the correlation between *distance* and *travel time*, *mode* and *travel time*, *distance* and *InboundOut* and *furtherAway* and *travelFeeD* are 0.5, -0.5, 0.38 0.4 respectively. Therefore *travel time*, *travelFeeD* and *InboundOut* were not considered for model selection.

Table 5. Correlation matrix for captive station

Correlation	Distance	Travel Time	Further away	InBound Out	Purpose	Gender	Age	Mode	Waiting Time	Cost OS	Cost SD
Distance	1	**A**	**B**	**B**	**B**	**B**	**B**	**B**	**A**	**A**	**A**
Travel Time	0.55*	1	**B**	**B**	**B**	**B**	**B**	**B**	**A**	**A**	**A**
Further away	-0.3*	-0.01*	1	**C**	**C**	**C**	**C**	**C**	**B**	**B**	**B**
InBound Out	0.39*	0.17*	-0.37	1	**C**	**C**	**C**	**C**	**B**	**B**	**B**
Purpose	-0.004*	0.14*	0.08*	-0.27*	1	**C**	**C**	**C**	**B**	**B**	**B**
Gender	0.01*	0.06*	0.14	0.06	-0.05	1	**C**	**C**	**B**	**B**	**B**
Age	0.00*	-0.08*	-0.15	0.31	-0.06*	-0.07	1	**C**	**B**	**B**	**B**

Mode	0.06*	-0.51*	-0.14*	0.28*	-0.22*	0.01	0.09	1	**B**	**B**	**B**
Waiting Time	0.1*	0.07*	-0.02*	-0.12	0.05*	0.18*	-0.1*	-0.02*	1	**A**	**A**
CostOS	0.09*	0.04*	-0.14*	0.07*	0.06*	0.06*	-0.15*	-0.01*	0.05*	1	**A**
CostSD	-0.17*	-0.29*	0.43*	-0.37*	-0.04*	-0.14*	0.02*	0.04*	-0.16*	-0.09*	1

* *P*-value <0.05. A:Pearson B:Polyserial C: Polychoric

As defined in the introduction, a captive station is the one located at the end or near the end of train line, which means this type of the station has a bigger catchment area and transit users have less choice of other stations. Therefore, fewer variables would influence the station choice. The derived model for captive stations provided evidence for this definition by identifying only two significant variables in the best-fitting logistic regression model(See Table 6). There are 274 records for the captive stations (Table 2), but the sample size for this regression model is 245 with 29 missing records being removed for the purpose of the analysis. Similar to the model discussed in section 4.1, distance has a negative influence on the nearest station choice. In addition, the further-way station choice suggested by the model has negative influences on the nearest station choice. The captive station attracted transit users who are willing to either drive or take buses to reach a station which is further away from their destination, and the nearest station is chosen for reasons such as seat availability.

Table 6. Logistic regression models for captive stations.

Variables	Estimate (SE)	OR (95%CI)	z value	Pr(>|z|)
(Intercept)	3.42 (0.4)	30.55 (13.91-67.05)	8.52	< 2e-16 ***
Distance (km)	-0.16 (0.03)	0.85 (0.8-0.89)	-6.12	9.42e-10***
Further_away	-1.44 (0.51)	0.24 (0.09-0.65)	-2.8	0.005 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Null deviance 245.20 on 244 degrees of freedom Residual deviance: 183.03 on 242 degrees of freedom (*p*-value =0.998) AIC: 189.03 Number of Fisher Scoring iterations: 5				

4.3 Logistic regression models for non-captive stations

It can be seen from Table 7 that the correlation between *distance* and *travel time*, *travel mode* and *time*, *purpose* and *InboundOut* and *InBoundOut* and *travelFeeD* are 0.5, -0.45, -0.5 and -0.35 respectively. Therefore, *travel time* and *InboundOut* were removed for model selection.

Table 7. Correlation matrix for non-captive station

Correlation	Distance	Travel Time	Further away	InBound Out	Purpose	Gender	Age	Mode	Waiting Time	CostOS	CostSD
Distance	1	**A**	**B**	**B**	**B**	**B**	**B**	**B**	**A**	**A**	**A**
Travel Time	0.53*	1	**B**	**B**	**B**	**B**	**B**	**B**	**A**	**A**	**A**
Further away	-0.29*	-0.21*	1	**C**	**C**	**C**	**C**	**C**	**B**	**B**	**B**
InBound Out	-0.04*	-0.23*	0.3	1	**C**	**C**	**C**	**C**	**B**	**B**	**B**
Purpose	0.09*	0.20*	-0.02	-0.52	1	**C**	**C**	**C**	**B**	**B**	**B**
Gender	0.06*	0.02*	-0.06	-0.07	0.05	1	**C**	**C**	**B**	**B**	**B**
Age	-0.00*	-0.0*	0.08	0.06	-0.04	-0.01	1	**C**	**B**	**B**	**B**
Mode	0.01*	-0.45*	0.04	0.29*	-0.31*	-0.02	-0.02	1	**B**	**B**	**B**
Waiting Time	0.03*	0.08*	-0.24*	-0.51*	0.13*	0.00*	-0.01*	-0.11*	1	**A**	**A**
CostOS	0.12*	0.19*	0.06*	0.03*	0.08*	-0.05*	-0.09*	-0.16*	-0.03*	1	**A**
CostSD	0.03*	0.12*	0.07*	-0.36*	0.10*	-0.16*	0.02*	-0.06*	0.09*	0.11*	1

* *P*-value <0.05. A:Person B:Polyserial C: Polychoric

Three variables were identified to be significant from the best fitting logistic regression model for non-captive stations (See Figure 8). There are 559 records for the non-captive stations (Table 2), but the sample size for this regression model is 486 with 73 missing records being removed for the purpose of the analysis. The most influential variable is travel cost (from a chosen station to a destination). The less the cost of travelling from a chosen station to a destination, the more likely a chosen station will be a non-nearest station, which is consistent with the results from the overall model. However, different from the model for all chosen station, cost (origin to station) was found to be significant. The less the cost from origin to the chosen station, the less likely it is that chosen stations will be the nearest station.

Table 8. Logistic regression models for non-captive station.

Variables	Estimate (SE)	OR (95%CI)	z value	Pr(>|z|)
(Intercept)	0.3 (0.7)	1.35 (0.34-5.35)	0.42	0.67
Distance (km)	-0.34 (0.03)	0.71 (0.67-0.76)	-10.3	< 2e-16 ***
Cost (station to destination)	0.53 (0.18)	1.7 (1.18-2.44)	2.88	0.004**
Cost (origin to station)	0.17 (0.09)	1.19 (1-1.41)	2.004	0.045*
Waiting time	-0.05 (0.03)	0.95 (0.9-1.01)	-1.702	0.089**.**
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Null deviance: 672.12 on 485 degrees of freedom Residual deviance: 467.12 on 481 degrees of freedom (*p*-value =0.67) AIC: 477.12 Number of Fisher Scoring iterations: 6				

5. SENSITIVITY TEST FOR POLICY IMPLICATION

A sensitivity test was conducted using the overall model shown in Table 3 to identify the influence of travel distance (from an origin to a chosen station) and travel cost (from a chosen station to a destination) on the nearest station choice by holding other independent variables at their mean. The resulting sensitivity plot for all stations (Figure 4a) indicates that the predicted probabilities of choosing nearest station decrease as travel distance increases for all five different travel fees, which are travelling over one zone ($2.70), two zones ($4.00), three zones ($4.90), four zones ($5.80) and five zones ($7.10). Generally, the closer the chosen station to the destination, the lower probability of a chosen station is the nearest train station to the origin, except when travelling over four zones ($5.80). In the Perth metropolitan area, only the Mandurah line extends over five zones. After closely examining the travel patterns of respondents who travelled by trains over four zones, we identified that 24% of them chose stations, mostly Murdoch station on the Mandurah line, even though they came from a location near other train lines. Interestingly, about 82% of these travelled by bus feeder services to the train station, which is a good example of commuter/work-based transit service.

Figure 4a also shows that the likelihood of respondents choosing a station that is the nearest station to their origin is over 80% if travel distance from an origin to the chosen station is less than 800 meters. However, when they have to travel over 10 km from an origin to the chosen station, the estimated probability of choosing the nearest station is still over 80% only for respondents travelling over five zones. For respondents travelling over less than five zones, the estimated probability dropped sharply, especially for respondents who travelled on trains within one zone, the likelihood decreased to 39%. The station, which belongs to the travelling-over-five-zone category, is Warnbro: a captive station. While stations belonging to the travelling-within-one-zone category are non-captive stations. Lack of competition with surrounding stations has led to a bigger catchment area for Warnbro station, leaving the train users with less travel options. This demonstrates a certain level of transport disadvantage for the train users.

The resulting sensitivity plot (Figure 4b) for non-captive stations calculated based on the model in Table 7 indicates the same trends shown in Figure 4. However, the probability of choosing the nearest stations for non-captive stations decreased more quickly than the captive station model. The non-captive model shows that there was less than a 50% chance for chosen stations to be the nearest station if travel distance was 10 km. No travel over five zones ($7.10) was identified for the transit trip involving in non-captive stations. This result could be interpreted that expect for a distance minimisation strategy, others such as cost and travel time minimisation strategies and multi-trip purpose utility maximisation could play important roles to the decision maker of this type trips. In addition, more travel uncertainty, such as availability of parking, could be involved in non-captive station choice than captive station choice.

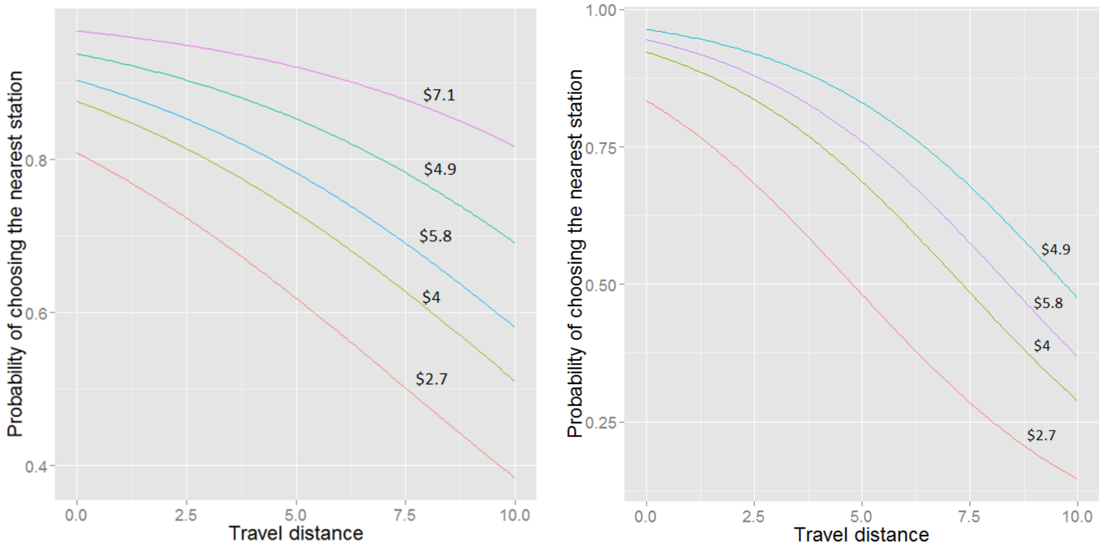

Figure 4 Predicted probabilities for travel distance to reach all station (4a left), non-captive stations (4b right): traveling over one zone ($2.7), two zones ($4.0), three zones ($4.9) and four zones ($5.8).

A sensitivity test was also conducted using the logistic regression model shown in Table 5 to identify the influence on the nearest station choice for travel distance (from an origin to a chosen station) and whether the station is further away from origins and destinations. The resulting sensitivity plot for captive stations, shown in Figure 5, indicates that the predicted probabilities of choosing the nearest station decrease as travel distance increases, whether the station is further-away or not. Generally speaking, the predicted probability of choosing the nearest station is the higher for non-further-away stations than further-away stations. According to Figure 5, at 10 km travel distance, the predicted probability of the nearest station choice is over 85% for non-further-away stations. In comparison, it is less than 60% for further-away stations.

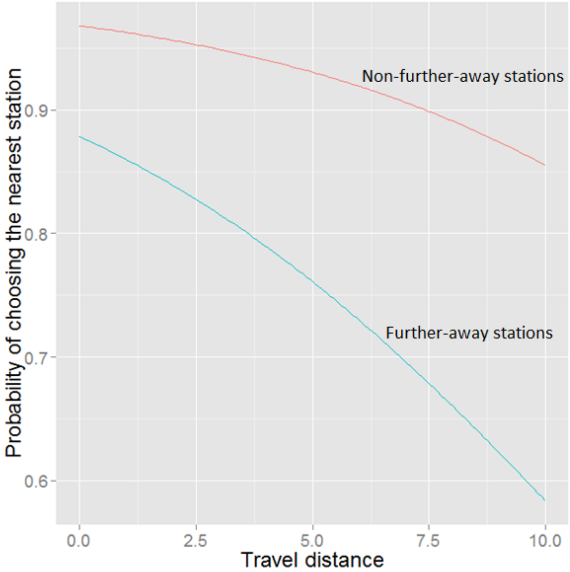

Figure 6 Predicted probabilities for travel distance for captive stations: further-away and non-further-away.

6. DISCUSSIONS AND CONCLUSIONS

This paper applied Logistic regression models to understand the nearest station choice behaviours of transit users. The study revealed that the nearest station choice depended on the location of station, and characteristics of stations and transit users. When the chosen station was located at the end or near the end of a train line (captive stations), this left transit users with much less station choice than a station located along the line (non-captive stations). This also means that less variables influence the nearest station choice. For example, only two variables— distance and station location in terms of destination direction (further-away or non-further away)— were relevant for the captive model. Based on our survey, the reasons why users choose a station further away from their origin and destination are seat- and parking-availability, particularly the former. This suggested that crowding on trains was becoming an issue in Western Australia, which can be managed by increasing train capacity and frequency (Li and Hensher 2012), providing better service and design, such as improving air quality and circulation (Thompson, Hirsch, and Rainbird 2012)

If the chosen station was located further along the train line, more variables would affect the nearest station choice of transit users. Based on our model, except distance, travel cost and waiting time were found to significantly influence the choice. Interestingly, transit users were willing to drive or take a bus to travel a little bit further towards their destination in order to decrease transit waiting time and travel cost. Two more variables— traffic congestion and travel comfort— although not captured by the model, were found to be applicable to this situation from our survey interviews. Many of the respondents preferred driving rather than using public transport due to convenience and comfort. However, when there is a trade-off between convenience and travel time, they optimised their trips by choosing a transit station along their trip (Debrezion, Pels, and Rietveld 2009). In addition, land use diversity could affect station choice (Badoe and Miller 2000, Cervero 1996). For example, a large shopping center, Westfield Carousel, was indicated by our respondents as one of reasons they chose Cannington station. The nearest station choice rate for Cannington station is only 26.9%.

Some limitations of this study are that limited variable data were collected (12 variables considered). In addition, due to correlation between variables, some variables such as travel time (correlated with distance), inbound and out bound trip (correlated with distance), train frequency (correlated with waiting time) were manually removed from model. More variables such as traffic congestion and land use diversity should be considered in the future study.

Another limitation of the study is that some of respondents had a misunderstanding of the survey question of “Where and when did you start that trip” which was used to collect the location of the respondents’ origin, especially those who were interviewed in the afternoon. They thought the place they departed in the morning was their origin, while some of them filled in the location of the activity immediately before they left for a station. Although we have removed some unreasonable results manually, there was still some ambiguity in the data. The other aspect is related to geocoding. Due to missing data of the landmark and street data from the survey, we have used the centroid location streets or suburbs as a substitution, which could reduce the accuracy of geocoding. In next survey, we will improve the questionnaire design to cater for this.

From a public transport policy point view, the result of the paper indicates that attention should be paid to the transit users who chose non-captive stations because more uncertainty was involved in non-captive station choice than captive station choice. This randomness could due to reasons such as, late departure, less likelihood to get parking in the nearest station and multi-trip purpose. Our future work will further investigate how much randomness could be involved in the station choice behaviour.

This study provides evidence as to why some transit users don’t choose the nearest station from their origin. The results of this study will be of importance to public transit policy makers, urban planners and researchers, particularly the Public Transport Authority, to understand transit choice behaviours. Therefore public transport policies such as adjustments of travel fees and improving station service and facilities, could be developed. The major contribution of this study is the development of a systematic approach for identifying variables affecting the nearest station choice. The method is reproducible and generalisable internationally to other studies.

Reference

Badoe, Daniel A., and Eric J. Miller. 2000. "Transportation–land-use interaction: empirical findings in North America, and their implications for modeling." *Transportation Research Part D: Transport and Environment* no. 5 (4):235-263. doi: http://dx.doi.org/10.1016/S1361-9209(99)00036-X.

Beimborn, Edward A, Michael J Greenwald, and Xia Jin. 2003a. "Accessibility, connectivity, and captivity: impacts on transit choice." *Transportation Research Record: Journal of the Transportation Research Board* no. 1835 (1):1-9.

Beimborn, Edward, Michael Greenwald, and Xia Jin. 2003b. "Accessibility, Connectivity, and Captivity: Impacts on Transit Choice." *Transportation Research Record: Journal of the Transportation Research Board* no. 1835 (-1):1-9. doi: 10.3141/1835-01.

Bernick, Michael, and Robert Cervero. 1997. *Transit villages in the 21st century*. New York: McGraw-Hill

Boyce, David, and Wharton School. 1973. *Impact of access distance and parking availability on suburban rapid transit station choice; analysis of the Philadelphia - Lindenwold High-Speed Line*. Philadelphia; [Springfield, Va.]: Regional Science Dept., Wharton School, University of Pennsylvania [Distributed by National Technical Information Service, U.S. Dept. of Commerce].

Brons, Martijn, Moshe Givoni, and Piet Rietveld. 2009. "Access to railway stations and its potential in increasing rail use." *Transportation Research Part A: Policy and Practice* no. 43 (2):136-149. doi: http://dx.doi.org/10.1016/j.tra.2008.08.002.

Cervero, Robert. 1996. "Mixed land-uses and commuting: evidence from the American Housing Survey." *Transportation Research Part A: Policy and Practice* no. 30 (5):361-377.

Cervero, Robert. 2006. "Office development, rail transit, and commuting choices." *Journal of Public Transportation* no. 9 (5):41-55.

Cervero, Robert. 2007. "Transit-oriented development's ridership bonus: a product of self-selection and public policies." *Environment and Planning A* no. 39 (9):2068-2085.

Davidson, B., and L. Yang. 1999. Modeling commuter rail station choice and access mode combinations. In *the Transportation Research Board Annual Meeting*. Washington, DC.

Debrezion, G., E. Pels, and N Picard. 2009. "Modelling the joint access mode and railway station choice." *Transportation Research Part E: Logistics and Transportation Review* no. 45 (1):270-283.

Debrezion, Ghebreegziabiher, Eric Pels, and Piet Rietveld. *Modelling the joint access mode and railway station choice* 2007. Available from http://www.tinbergen.nl/discussionpapers/07012.pdf.

Debrezion, Ghebreegziabiher, Eric Pels, and Piet Rietveld. 2009. "Modelling the joint access mode and railway station choice." *Transportation Research Part E:*

Logistics and Transportation Review no. 45 (1):270-283. doi: http://dx.doi.org/10.1016/j.tre.2008.07.001.

Department of Infrastructure and Transport. 2012. Understanding Australia's urban railways, Research Report 131. Canberra ACT 2601, Australia: Bureau of Infrastructure, Transport and Regional Economics (BITRE).

Desfor, Gene. 1975. "Binary station choice models for a rail rapid transit line." *Transportation Research* no. 9 (1):31-41. doi: http://dx.doi.org/10.1016/0041-1647(75)90018-0.

Fan, K., E. Miller, and D. Badoe. 1993. "Modeling rail access mode and station choice " *Transportation Research Record* no. 1443:49-59.

Givoni, Moshe, and Piet Rietveld. 2007. "The access journey to the railway station and its role in passengers' satisfaction with rail travel." *Transport Policy* no. 14 (5):357-365. doi: http://dx.doi.org/10.1016/j.tranpol.2007.04.004.

Hensher, D. A., M.John. Rose, and H.Willian. Greene. 2005. *Applied Choice analysis*. New York, America: Camberidge university Original edition, Cambridge University Press.

Jansson, Kjell, and Truls Angell. 2012. "Is it possible to achieve both a simple and efficient public transport zone fare structure? Case study Oslo." *Transport Policy* no. 20 (0):150-161. doi: http://dx.doi.org/10.1016/j.tranpol.2011.07.005.

Kastrenakes, C.R. 1988. "Development of a rail station choice model for NJ transit." *Transportation Research Record* no. 1162:16-21.

Krygsman, Stephan, Martin Dijst, and Theo Arentze. 2004. "Multimodal public transport: an analysis of travel time elements and the interconnectivity ratio." *Transport Policy* no. 11 (3):265-275. doi: http://dx.doi.org/10.1016/j.tranpol.2003.12.001.

Li, Zheng, and David Alan Hensher. 2012. "Crowding in public transport: a review of objective and subjective measures."

Lindsey, Marshall, Joseph L. Schofer, Pablo Durango-Cohen, and Kimberly A. Gray. 2010. "Relationship between proximity to transit and ridership for journey-to-work trips in Chicago." *Transportation Research Part a-Policy and Practice* no. 44 (9):697-709. doi: 10.1016/j.tra.2010.07.003.

Mcfadden, D. 1974. *Frontiers in econometrics*. New York: Academic Press.

Nordlund, A., and K. Westin. 2013. "Influence of values, beliefs, and age on intention to travel by a new railway line under construction in northern Sweden." *Transportation Research Part A: Policy and Practice* no. 48 (0):86-95. doi: http://dx.doi.org/10.1016/j.tra.2012.10.008.

Ralph B. D'Agostino, Sr., Lisa M. Sullivan, and Alexa S. Beiser. 2006. *Introductory Applied Biostatistics*.

Rietveld, Piet. 2000. "The accessibility of railway stations: the role of the bicycle in The Netherlands." *Transportation Research Part D: Transport and Environment* no. 5 (1):71-75. doi: http://dx.doi.org/10.1016/S1361-9209(99)00019-X.

Thompson, K., L. Hirsch, Muller, S. , and S. Rainbird. 2012. A socio-economic study of carriage and platform crowding in the Australian railway industry: Final Report,. Brisbane, Australia: CRC for Rail Innovation.

Train, Kenneth. 2012. *Discrete Choice Methods with*
Simulation 2002 [cited 17 April 2012]. Available from http://elsa.berkeley.edu/books/choice2nd/Ch0_Front_Quotes_Contents.pdf.

Tversky, Amos. 1972. "Elimination by aspects: A theory of choice. ." *Psychological Review,* no. 79 (4):281-299.

Wardman, M., and G.A whelan. 1999. using geographical infomation systems to improve rail demand models. In *Final report to Engineering and Phy sical Science Research Council*

Wardman, Mark, William Lythgoe, and Gerard Whelan. 2007. "Rail Passenger Demand Forecasting: Cross-Sectional Models Revisited." *Research in Transportation Economics* no. 20 (0):119-152. doi: http://dx.doi.org/10.1016/S0739-8859(07)20005-8.