UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

A Trip End Model for the Manchester Metrolink

by

Emine Tugba Yazici

Supervisor Dr Simon Blainey

A Dissertation Submitted in Partial Fulfilment
of the Degree of MSc (Transportation Planning and Engineering)
by Instructional Programme

September 2018

Abstract

While light rail systems have been proposed for many cities in the UK in recent decades, relatively few have been constructed. The most successful of these systems has arguably been the Manchester Metrolink, which has been significantly expanded in recent years, and with further extensions planned. Similar extensions (and new systems) are proposed for other cities, but there is currently no transferable methodology in the UK for forecasting the number of passengers who would be expected to use individual stations on light rail systems. Therefore, this study aims to fill this gap by developing a trip end passenger demand model for stations on the Manchester Metrolink. The project used daily passenger count data for existing 93 light rail stations in Greater Manchester as a dependent variable, and factors affecting the station usage were investigated. Also, Geographic Information System (GIS) software for data manipulation and Statistical Analysis Software (SPSS) for model calibration were used. In general, investigated factors considered the influence of station characteristics, service area characteristics and socio-economic factors. For the measurement of some of these factors, the study examined the three different catchment area definitions based on the previous researches and chose the best-fit among these three definitions for each independent. After that, many different combinations of these explanatory variables were examined. As a result, the study provides a statistically significant model (adjusted $R^2 = 0.855$) with five variables including the number of the tram in an hour, workplace population (employment), resident population, distance to city centre and the size of car parking space around the station. These findings show similarities with previous researches. The final model might be useful to forecast demand at new light rail stations on the Manchester Metrolink.

Statement of originality

I, (Emine Tugba Yazici) declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research, and the content of this thesis has not previously been submitted for a degree in any University.

Acknowledgements

I would like to thank my supervisor and personal tutor, Dr Simon Blainey for his continued support, patience, and enormous knowledge. His guidance helped to complete researching and writing of this thesis.

I would like to express my gratitude to my sponsor, The Republic of Turkey, Ministry of National Education, giving me this opportunity.

I would like to thank my family and friends, Mustafa Yazici Muazzez Yazici, Edanur Yazici, Mucahit Atalan, Mustafa Ekmekci, Sara Saglam, Paraskevi Sarri, Keli Mallmann, Sreedevi Padmakumar who encourage and support me during all year.

Table of Contents

List of Figures

List of Tables

Nomenclature

CDRC	Consumer Data Research Centre
GIS	Geographic Information Systems
GJT	Generalised Journey Time
GWR	Geographically weighted regression
LSOA	Lower Layer Super Output Area
MSOA	Middle Layer Super Output Area
OA	Output Area
OS	Ordnance Survey
PDFH	Passenger Demand Forecasting Handbook
TfGM	Transport for Greater Manchester
WZ	Workplace Zone
NaPTAN	National Public Transport Access Nodes

1. Introduction

1.1. Transport demand

There is an increasing interest towards the topic of demand estimation and modelling among many transport planners because forecasting demand is able to help to assess project sustainability and viability over time. With accurate estimates, the transport facilities can be designed in proper capacity, and by good service quality. Therefore, operators and public administration can alleviate their substantial transport investments (Zhang and Wang, 2014).

Demand estimation is essential for all transport modes, but the methods applied would be changed for public transport compared to transport by private car. Mainly, the four-stage models at an aggregate level (McNally, 2007) or activity-based models at a disaggregate level (Hildebrand, 2003; Bowman & Ben-Akiva, 2001) are used to forecast car travel. However, according to public transport operators, these models are cumbersome, expensive and require excessive data to establish models (Preston, 2015). Due to this plausible reason, elasticity-based approaches are applied to examine public transport demand. However, when the base demand of facilities is zero, this method cannot be suitable because the calculation of demand is equal to the multiplication of base year demand and the proportional change, so the demand at new stations remains always zero (Preston, 2015). Further detail regarding to elasticity based method will be explained in Section 2.1. Moreover, to handle this issue, different demand models have been improved for new public transport services. These are trip end model, flow level demand and choice model (see Figure 1.1). These methods will be explained in Section 2.2. In the UK, these models were particularly applied to heavy rail (Preston 1987, 1991a, 1991b; Lythgoe, 2004; Wardman et al., 2007; Blainey, 2010; Young, 2017). However, the presence of models for heavy rail is not useful to forecast demand at light rail stations because firstly the researchers did not incorporate the light rail stations in their calibration dataset, and secondly, both rail systems have different characteristics in term of passenger capacity, speed, user profile and journey distance between stations. Therefore, separated modelling might be more realistic as same as that Lane, (2006) has done separately for both light rail and heavy rail stations in U.S regions. However, there is no existing study for light rail in the UK so that this study will be mainly concerned with demand at light rail.

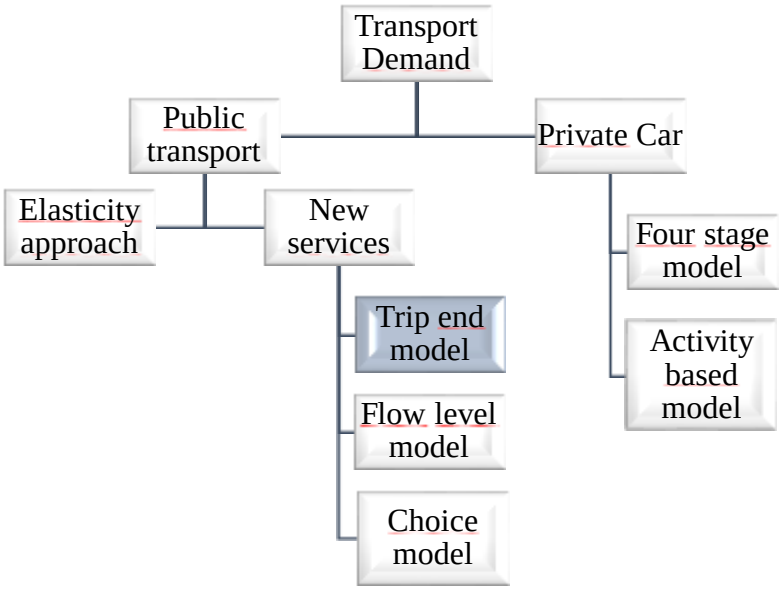

Figure 1.1: The different models used for forecasting transport demand

1.2. Light rail and tram system in Greater Manchester

The principal concern of this study is the light rail system. There are eight light rail and tram systems in England; their locations on the map are shown in Figure 1.4. Since 1995, there has been an almost twofold increase in the number of light rail and tram stations in England, from 207 to 400 (see Figure 1.3). The largest increase has been seen in the Manchester Metrolink system, known simply as 'Metrolink', with the construction of 67 new stations (GOV.UK, 2018). Indeed, in recent years, Metrolink has been significantly expanded, with yet further extensions also being planned. Thus, this study focuses solely on Metrolink, which is the largest tram system in the United Kingdom with a network of 57 miles (92 km) of track. Metrolink has 93 stations, located throughout Manchester (42 stations) and its environs: Trafford (12), Salford (10), Oldham (10), Tameside (7), Rochdale (6), and Bury (6). With a new line presently under construction to Trafford Park, the Metrolink network would reach 62 miles (100 km) in length and have 99 stations, as shown in Figure 1.4 (Transport for Greater Manchester, 2018a; GOV.UK, 2018). The construction of this new line would make the network one of the longest tram systems in Europe. Transport for Greater Manchester (TfGM) aims to create a fully integrated public transport system and connect these links to important business centres. Thus, there is an expansion proposal for the new western loop, which aims initially to extend to Manchester Airport's Terminal 2 development and then connect to the proposed High Speed 2 link (HS2) (Transport for Greater Manchester, 2018b). As a result, it is thought that the presence of a demand-forecasting model for Metrolink can help to assess the new stations' sustainability and viability over time.

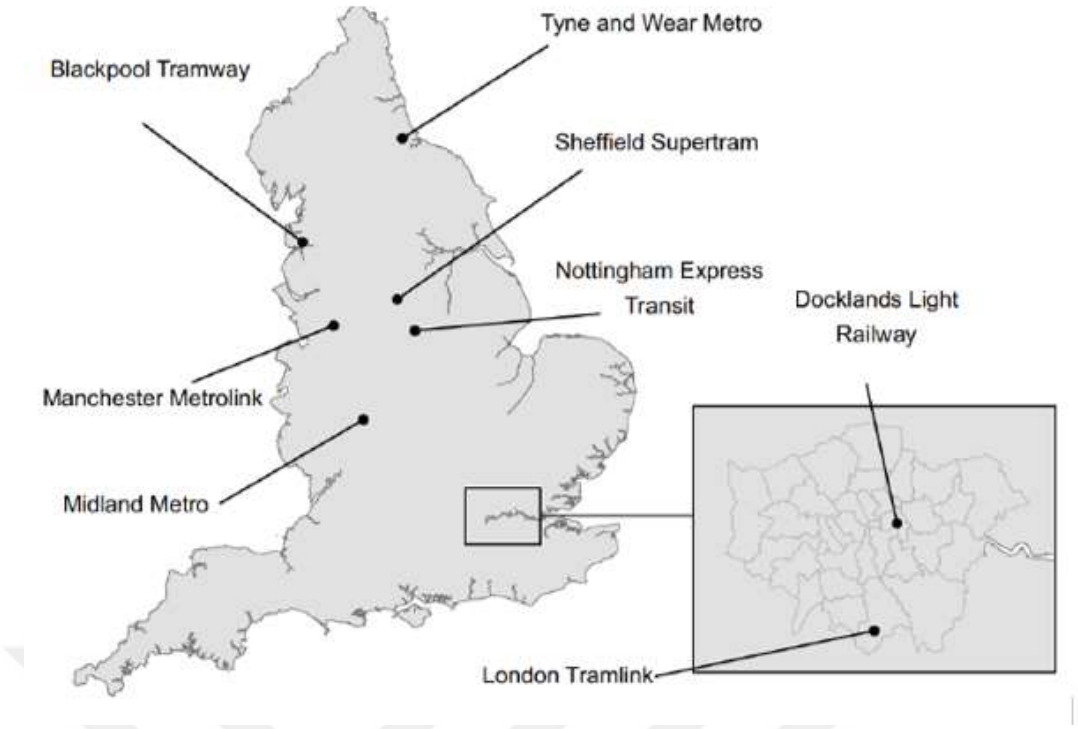

Figure 1.2: The location of all light rail and tram system in England (Department for Transport, 2018)

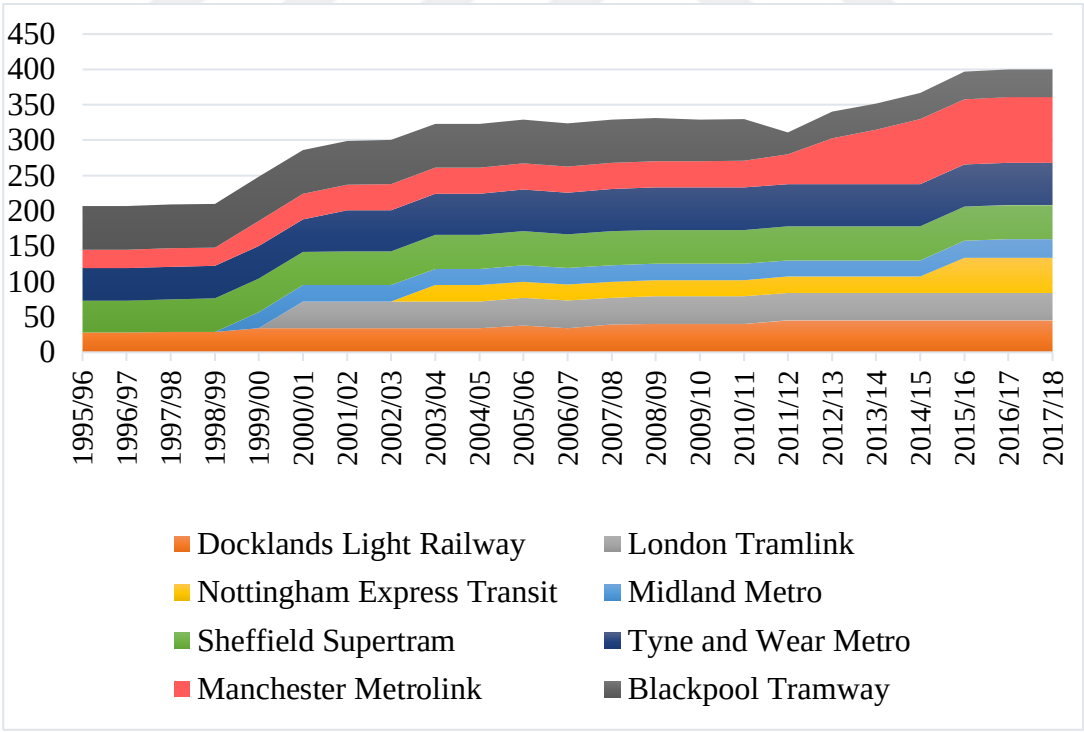

Figure 1.3: The number of stations on light rail lines in England from 1995 to 2018 (GOV.UK, 2018)

Figure 1.4: The geographical map of Manchester Metrolink (Transport for Greater Manchester, 2018a)

1.3. Research aim and objectives

There is currently no transferable methodology for forecasting the number of passengers who would be expected to use stations on light rail systems in the UK. This study aims to fill this gap by developing a trip end demand model for new stations on the Manchester Metrolink. The objectives are outlined below.

1. Reviewing literature to understand differences among models used for rail demand estimation. This is presented in Section 2.
2. Investigating the influence on station usage of a wide range of factors obtaining from available data sources. This is presented in Section 4.
3. Examining the methodologies previously applied to station level demand estimation across the world and then adapting the most convenient method into the study area given. This is presented in Section 4.1.
4. Using ArcGIS software for data manipulation, and the visual presentation of data and models. This is presented in Section 3 and 4.

1.4. Thesis structure

This section describes how the aim and objectives listed in Section 1.3 were achieved during a Master dissertation period. After the introduction part (see Section 1), previous studies regarding rail demand modelling and the factors affecting this demand are reviewed in Section 2. The methodologies and data preferred for this study are described in Section 3. The development of the trip end model and the final model chosen are presented in Section 4. Finally, Section 5 concludes the findings of this study and contains some limitation of the study and recommendation for future work.

2. Literature Review

This section includes a review of previous studies conducted in the field of forecasting rail demand. Forecasting demand methods vary because of the available data sources, the station characteristics, and the condition of the station being new or already exist. Passenger Demand Forecasting Handbook (PDFH) which is the primary source for rail demand modelling in the UK suggests that each method cannot be suitable for every study area due to transferability problems, data limitations and the characteristics of the study area, but the more detailed examination of methods makes a model the more ideal.

There are two main categories of rail demand modelling, with the elasticity-based approach discussed in Section 2.1 and the models for new services discussed in Section 2.2. These new service models will be categorised as trip end model, direct demand model, and mode choice model.

2.1. Elasticity based approach

The basic principle of this approach is that viewing public transport as a typical consumer good. Thus, like other consumer good, the usage of public transport mode is based on the price, income and price of the rival products. In addition to these monetary factors, time related factors such as access and egress time, wait time, and in-vehicle time have an impact on demand for public transport (Preston, 2015). For example,

- the rail demand goes up (down) when the fare falls (increases)
- the rail demand goes up (down) when income increases (falls)
- the rail demand goes up (down) when the price of rival products (e.g. car travel) increases (falls)
- the rail demand goes up (down) when the time related factors fall (increases) (Preston, 2015).

The proportionate change in demand in response to a change in these factors is defined as elasticity. These are fare elasticity, income elasticity and generalised journey time (GJT) elasticity and cross elasticities (observes the effects of other transport modes namely, bus, car and air). In the UK, the Passenger Demand Forecasting Handbook (PDFH) is one of the sources of rail elasticities (ATOC, 2013). In addition to the elasticities of factors above, PDFH includes many different elasticities such as employment elasticity, fuel cost elasticity and car time elasticity. Therefore, the demand can be calculated by using the following equation 2.1.

$$J_{new} = I_E \times I_T \times I_F \times J_{base} \tag{2.1}$$

where I_E, I_F, and I_T are demonstrating the proportional change in demand because of the external factors, generalised journey time, and fares see following equations 2.2, 2.4, 2.5 (Preston, 2015, p.197).

$$I_E = \left(\frac{GDP_{new}}{GDP_{base}}\right)^g x \left(\frac{POP_{new}}{POP_{base}}\right)^p x \exp(n(\mathrm{NC}_{\mathrm{new}} - NC_{base})) x \left(\frac{FUELCOST_{new}}{FUELCOST_{base}}\right)^f x \left(\frac{CARTIME_{new}}{CARTIME_{base}}\right)^c x$$

$$\left(\frac{BUSCOST_{new}}{BUSCOST_{base}}\right)^b x \left(\frac{BUSTIME_{new}}{BUSTIME_{base}}\right)^t x \left(\frac{BUSHEAD_{new}}{BUSHEAD_{base}}\right)^b x \left(\frac{AIRCOST_{new}}{AIRCOST_{base}}\right)^a x \left(\frac{AIRHEAD_{new}}{AIRHEAD_{base}}\right)^f \tag{2.5}$$

$$I_{GJT} = \left(\frac{GJT_{new}}{GJT_{base}}\right)^j \tag{2.3}$$

$$I_F = \left(\frac{F_{new}}{F_{base}}\right)^f \tag{2.4}$$

The equation contains 10 external elasticities, "covering Gross Domestic Product per capita (replaced by Employment for key commuter markets), Population, (Non-) Car Ownership, Car Costs and Journey Time, Bus Costs and Headway and Air Costs and Headway. In addition, there are fares elasticities and generalised time (GJT) elasticities, where GJT includes journey time, along with interchange and service interval penalties, with the latter being a function of headways" (Preston, 2015, p.198).

2.2. New services

Using the elasticity-based approach for the new station is not possible because their demand at base year does not exist, so the proportional change cannot provide an estimate. Therefore, for the newly opened station, demand can be estimated based on current demand at existing stations and local factors. The models improved are described in Sections 2.2.1; 2.2.2; 2.2.3 below. In the UK, a trip end model is commonly used so that this study will apply trip end model for light rail service in Greater Manchester.

2.2.1. Trip end model

One of the cheapest and easiest methods to estimate rail demand is trip rate model which incorporates only population variable. However, the model accuracy is questionable because of the omission of other important factors. Thus, trip end models which are an advanced version of the trip rate model with additional explanatory variables were applied by Preston (1991a) for new local stations in the UK.

Basically, trip end models are multiple regression models to forecast demand at station level (see equation 2.2). Current demand at existing station is used as the dependent variable, and factors affecting the station usage are investigated. These factors might be different for every region due to differentiation in station accessibility, the people attitude toward public transport use, passenger's travel behaviour, and land use around the station. Even the study area is the same, the model calibration should be separately done for light rail and commuter rail stations. One of the good examples can be seen in a study by Lane et al. (2006) over the 17 U.S. regions. This study improved two multiple regression equations; one is for commuter rail lines while other is for light rail lines. The data variability for models will be detailly examined in Section 2.3.

Moreover, the form of the trip end model can vary. Kuby et al. (2004), Blainey (2010), de Andrade et al. (2014) and Zhao et al. (2014) improved a multiple linear regression for demand modelling in U.S cities, England and Whales, Brasil, and China, respectively. Currie et al. (2011) also used linear regression, but to find demand at route level in Australia, North America and Europe. Other studies have applied multiplicative format for demand modelling at the station level (Lane et al., 2006;

Blainey, 2010; Choi et al., 2012). The multiplicative format is obtained from a logarithmic transformation of linear models. Blainey (2010) noted that the multiplicative form of model gave the better predictive power than linear format as shown 2.5 below. The application of quadratic regression was only observed in the study of Sun et al. (2016).

Some studies have examined spatial influence on demand, using geographically weighted regression (GWR), distance/time-decay weighted regression, and network kriging regression (Blainey, 2010; Gutiérrez et al., 2011; Cardozo et al., 2012; Blainey and Mulley, 2013; Zhang and Wang, 2014; Young, 2017). It is important to note that Blainey (2010) used the GWR model in local railway stations in England and Wales (national scale) while Cardozo et al. (2012) implemented it to Madrid Metro stations (local scale). However, both stated that GWR developed model fit over the standard linear model for demand estimation. Similarly, Zhang and Wang (2014) stated that network kriging models increased demand estimation accuracy over linear models. Furthermore, Huang (2017) investigated the impact on demand of the accessibility, using accessibility-weighted regression model. He noted that the accessibility model at route level improved the accuracy of demand estimation.

An example of the trip end model improved to forecast demand at a local rail station in England and Wales is presented below (Blainey, 2009):

$$\widehat{V}_i = \alpha \left(\sum_{a}^{n} P_a w_a \right)^{\beta} F_i^{\delta} T_i^{\lambda} J_{i4}^{\tau} Pk_i^{\rho} Te_i^{\kappa} B_i^{\eta} El_i^{\nu} \qquad (2.5)$$

Where:

V_i is the predicted total demand at station i

P_a is the resident population in output area a

(a,. . . ,n) are the output areas whose nearest station by car travel time is station i

w_a is weight function attached to population unit *a*, given by $(t + 1)^{-3.25}$

F_i is the train frequency at station i on a normal weekday

T_i is the distance in km from station i to the nearest category A–D station

J_{i4} is the number of jobs located within 4 minutes drive of station i

Te_i is dummy variable taking the value e^1 if Station *i* is a terminus, and e^0 otherwise

B_i is dummy variable taking the value e^1 if Station i is a Travelcard boundary station, and e^0 otherwise

El_i is dummy variable taking the value e^0 if Station i is served by electric trains, and e^0 otherwise

$\alpha, \beta, \delta, \lambda, \tau, \rho, \eta, \kappa, \nu$ are parameters produced by calibration

2.2.2. Direct demand model (flow level)

Direct demand models are econometric models which combine the generation, distribution and mode choice stages of the four-stage model. While the trip end models forecast demand per station, direct demand model can estimate rail demand at the flow level between pairs of stations in other words station-to-station demand. The flow level demand estimation is superior to the station-based estimation due to the inclusion of the service pattern from a station to the destination (Preston, 2015). Therefore, it is possible to estimate the income generated by new stations and to observe how additional passengers affects rail services (Blainey and Preston, 2013). However, such models are more complicated and need to have flow level ticket sales data which is not in the open data sources (Preston, 2015). Thus, there are limited studies in the literature, following Choi et al., (2012); Blainey and Preston, (2013); Zhao et al., (2014); Kepaptsoglou et al., (2017). More detail related to these studies are summarised in Table2.1. The form of the direct demand modelling is preferred as the multiplicative model which allows interrelationships between variables of origin and destination. An example of the direct demand model to forecast demand at rail stations in South East Wales is presented below (Blainey and Preston, 2013):

$$\hat{V}_{ij} = \alpha \left(\sum_{a}^{n} P_a w_a \right)^{\beta} J_{i4}^{\tau} Pk_i^{\rho} Ex_j^{\gamma} D_{ij}^{\omega} Rs_{ij}^{\delta} Cs_{ij}^{\kappa} H_{ij}^{\eta} Rfkm_{ij}^{\lambda} \quad (2.6)$$

Where:

$\widehat{V_{ij}}$ is the predicted number of trips per year made from station i to station j

Ex_j is the total number of trips ending at station j in the year being modelled

D_{ij} is the straight line distance (in km) from station i to station j

Rs_{ij} is the rail journey time from station i to station j divided by D_{ij}

Cs_{ij} is the car journey time from station i to station j divided by D_{ij}

H_{ij} is the service headway in minutes between station i and station j

$Rfkm_{ij}$ is the fare per rail km for travel from station *i* to station *j*

$\alpha, \beta, \tau, \rho, \gamma, \omega, \delta, \kappa, \eta$ and λ are parameters determined during calibration

2.2.3. Choice model

Trip end and direct demand models are the aggregate models using zone-based data whereas choice models are the disaggregate models using household and individual data. Depending on the different aim, several choices approach has been improved such as station choice model, access mode and main travel mode choice. For example; Young and Blainey (2018) have used the station choice models to increase the accuracy of the trip end model. Lythgoe (2004) stated that the inclusion of a station choice decision in cross-sectional direct demand models improved model predictive power. Debrezion et al. (2009) explained the factors that impact Dutch traveller's choice with access mode and departure rail station choice models. The result shows that service frequency, parking places and bicycle stands have a positive whereas travel time and distance has a negative influence on the choice of departure stations (Debrezion et al. 2009).

The format of station choice model is the multinomial logit (MNL) model. However, when the analysis focused on simultaneously access mode and station choice models, the hierarchical (nested) logit model is necessary. The Ortuzar and Willumsen, (2011, p.234) pointed out the problem that "the MNL model is affected by the independence of irrelevant alternative axiom that prevents differential patterns of substitution" (as cited in Preston, 2015, p.204). Therefore, the hierarchical logit models classify similar alternatives in nests to deal with this problem.

2.3. Factors affecting rail demand

Many different factors might affect rail demand. Previous studies show that these factors vary depending on the study area and mathematical technic. Basically, they can be categorised into three categories; socioeconomic factors, service area characteristic and station characteristics (Gutiérrez et al., 2011; Sun et al., 2016).

- *Socioeconomic factors*

Car ownership, household income, age and race/ethnicity might affect the rail demand. The idea is that an increase (decrease) in real income and car ownership might decrease (increase) the use of public transport such as rail. Blainey and Mulley, (2013) have observed the negative influence on rail demand of the car ownership in Sydney Region. However, Blainey, (2010) and Gutiérrez et al. (2011) found out that car ownership and age factor do not have a significant impact on rail station demand in England, Whales and Madrid, respectively.

- *Station characteristics*

The demand is associated with station type (interchange, intermediate, terminus, or intermodal). Terminus stations have a bigger service area than intermediate stations, so tend to attract more demand. Interchange stations and intermodal stations receive more passenger demand than other stations because they can catch a passenger from other rail line and other transport modes (Gutiérrez et al. 2011). Moreover, the connectivity of station with bus line, the provision of car parking and bicycle stand area at station significantly affects rail demand (Kuby et al., 2004; Lane et al., 2006; Blainey, 2010; Gutiérrez et al., 2011; Cardozo et al., 2012; Choice et al., 2012 Blainey and Mulley, 2013; Zhao et al., 2014; Young, 2017). Service frequency has an important influence on demand estimation (Blainey, 2010). However, Gutiérrez et al. (2011) and Cardozo et al. (2012) note that it was not included due to the endogeneity problems.

- *Service area characteristic*

Land use patterns in service areas are important factors affecting rail demand. Some previous studies described these patterns just with population and employment variables (Kuby et al. 2004; Blainey, 2010; Currie et al., 2011; Young, 2017). However, for more detailed examination, some study separated land use in accordance with commercial, education, residence, office, shopping, leisure activities (Gutiérrez et al., 2011; Sun et al., 2016; Zhao et al., 2014; Choi et al., 2012; Huang et al., 2017). These all studies prove that land use variables are statistically significant.

Moreover, accessibility to the station is another important parameter. The expression of accessibility in the model can be done with station location, station facilities, the connectivity with other transport modes, street design and street density around the station. For example; more walkable environment increases the passenger demand at the rail station (Choi et al. 2012). The most detailed examination of this factor has been done by Huang et al. (2017).

2.4. Modelling issue: station catchment (service) area

Many factors affect the station usage as described in Section 2.3. However, the analysis of the impact on demand of these factors (employment, population, car ownership accessibility) is not straightforward. The catchment area which contains the potential users of a station is required. This catchment area should best represent the real travel pattern of passengers. In literature, the most general method is used that creating the buffer zone around the station to catch users. However, the decision of distance threshold of buffer zone varies depending on the spatial feature of the study area, explanatory variables and access/egress mode. For example; 800 meter network distance with distance-decay function gave the best-fit description of variables around Madrid metro stations (Gutiérrez et al. 2011). 0.5 mi based on shortest network paths was chosen by Kuby et al. (2004) for

light rail stations in U.S cities. Moreover, also again, in light rail and commuter rail in U.S regions, 120 demographics variables were determined by circle catchment area with radii of 0.25 mi, 0.5 mi, 1 mi, and 2 mi (Lane et al. 2006). Sometimes, time-based buffer zone definition was preferred rather than distance-based. For example; Blainey (2010) used 4 minutes threshold to define his buffer zone for employment. These all works used to GIS software to determine distance and explanatory variables.

In addition to this single buffer zone method, Sun et al. (2016) determined multi-level catchment areas in accordance with access/egress mode. As shown in Figure 2.1, the population within inner circle reach to the station by walking, the population within the outer circle by bicycle or bus, and the region outside the outer circle represents the population who think that accessibility is inconvenient. The radii of the inner (770 m) and outer (3800 m) areas are defined by a resident travel survey, while the region outside the outer circle is defined using the Thiessen polygon (Sun et al., 2016).

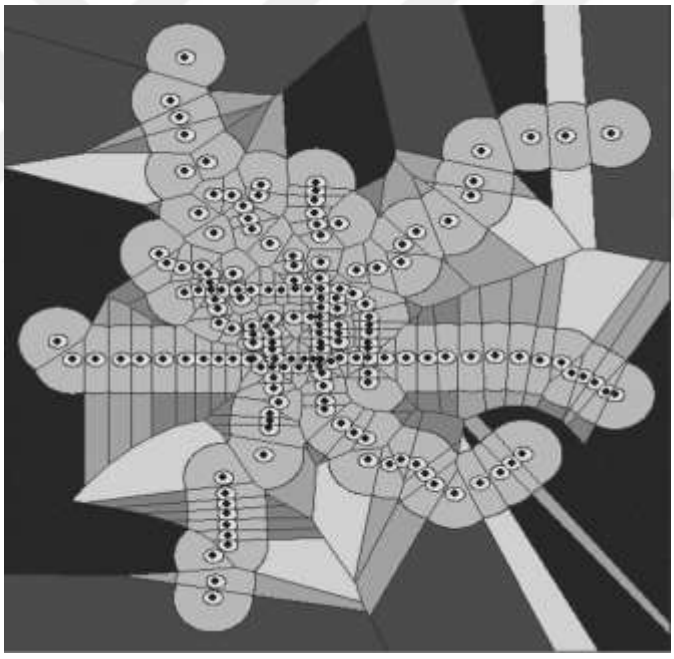

Figure 2.1: The multi-level catchment areas (Sun et al., 2016, p.143)

Instead of the buffer zone-based approach, Blainey (2010) defined the new station catchment that allocates all census output areas in England and Whales to their closest rail station by network distance or travel time. Figure 2.2 displays the allocation of the output area centroids to their closest station. Also, the distance/time decay weighted function was used to represent the traveller behaviour which users' willingness to use station gradually decrease as distance or time increase. See equation (2.7) is the format of the allocation based catchment, and equation 2.8 is a time decay weighted function.

$$V_i = \alpha + \beta \sum_{a}^{n} P_a \, w_a \tag{2.7}$$

$$W_a = (t+1)^{-3.25} \tag{2.8}$$

Where:

V_i is the predicted total demand at station i

P_a is the resident population in output area a

(a,. . . ,n) are the output areas whose nearest station by car travel time is station

w_a is the decay function (time based) attached to the output area a

t the travel time from the output area a to the nearest station

α, β are parameters produced by the calibration

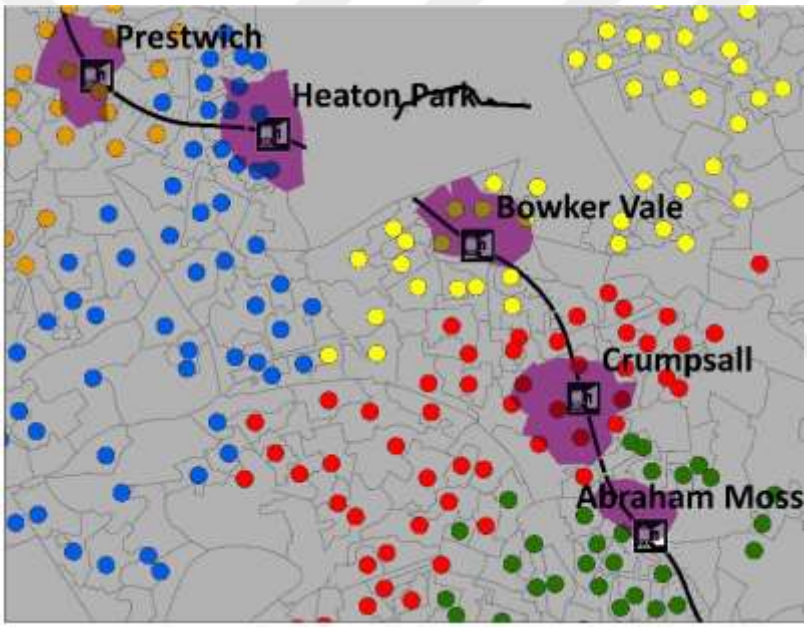

Figure 2.2: Showing the allocation of output area centroids to their closest station in Manchester Metrolink

In recent years, Young and Blainey (2018a, p.81) stated that the deterministic catchment definition could not "capture the patterns of abstraction and competition between railway stations". Therefore, they developed a probabilistic catchment using the station choice models which are based on the disaggregated data. The probability of a station being selected, amongst a group of alternative stations is calculated by a multinomial logit model (see equation 2.10). The utility function here was calibrated by revealed preference passenger survey data. Moreover, then, the probability is inserted into trip end models (equation 2.9).

$$V_i = \alpha + \left(\sum_{z}^{Z} Pr_z P_z w_z \right)^{\beta} \tag{2.9}$$

$$Pr_z = \frac{e^{V_z}}{\sum_{j=1}^{J} e^{V_j}} \tag{2.10}$$

Where: Pr_z is a station probability, V_z is utility of station z, V_j is the utility of competing stations, these utilities include explanatory factors which are nearest station, access time for walking, cycling, car and public transport separately, full-time and part-time staff, train frequency, the presence of CCTV, car parking space, free car park, ticket machine, toilets, and train leg time (Young and Blainey, 2018).

The probabilistic catchment approach using disaggregate population data was also applied by Huang et al. (2017). However, the probability is derived from distance-decay function (see equation 2.12). The underlying principle of the function is that the potential demand gradually decreases as distance or time to station increases. Therefore, the multiplication of a probability with population gives the potential demand at the station (see equation 2.11).

$$D_k = \sum_{i} P_{ik} S_i \tag{2.11}$$

$$P_{ik} = e^{a-bd_{ik}} / (1 + e^{a-bd_{ik}}) \tag{2.12}$$

Where: "D_k is the potential demand around station k, i is any location within the searching distance of station k, S_i is the population at location i, P_{ik} is the probability of making a rail trip from location i to station k, a and b are parameters for adjusting the negative logistic curve, and d_{ik} is the distance (in metres) between location i and station k" (Huang et al., 2017, p.628).

Table 2.1: The summary of previous studies

Publication	Study area	Station	Model format	Dependent variable	Explanatory variables in final model
Huang et al. (2017)	Wuhan, China	Light-rail system	Accessibility-weighted model and Distance-decay weighted function	Route level	**Service area characteristic** (accessibility-related variables, population, employment, and land use variables)
Young (2017)	England and Whales	Heavy rail system	Distance or time-decay weighted regression with station choice model	Station level	**Service area characteristic** (population, employment, train frequency, distance from the station to the nearest Category A-D station) **Station characteristic** (travelcard boundary station, terminus station, the station served by electric trains, car parking space)
Kepaptsoglou et al. (2017)	Cyprus	Light rail system	Multiplicative	Statin-to-station	Travel times by car and light rail, travel cost by car and light rail depend on survey data
Sun et al. (2016)	Beijing, China		Quadratic regression	Station level	**Service area characteristic** (land use variables which are education, commuting, residence, shopping, leisure areas)
Zhang and Wang (2014)	New York	Rapid rail systems	Network Kriging regression	Station level	**Service area characteristic** (population, employment, retail area, storage area, attractions) **Station characteristic (**number of the line) **Socioeconomic factors** (income)
Zhao et al. (2014)	Nanjing, China	Metro network	Linear, multiplicative	Station level and Statin-to-station	**Service area characteristic** (population, employment, business area, office area, number of major educational buildings, entertainment venues and shopping centres, CBD* dummy variable, road length within catchment) **Station characteristic** (feeder bus lines, bicycle P&R spaces, and transfer)
De Andrade et al. (2014)	Rio de Janeiro, Brasil	Rail network	Linear	Station level	**Service area characteristic** (employment) **Station characteristic** (the number of lines)
Blainey and Mulley (2013)	Sydney	Rail network	Geographically weighted regression	Station level	**Service area characteristic** (population, distance from Sydney to CBD, distance to nearest metropolitan strategy centre) **Station characteristic** (bus headway, train frequency, car parking, bike racks, bike lockers) **Socioeconomic factors** (car ownership)
Choi et al. (2012)	Seoul, South Korea	Metro network	Multiplicative, Poisson regression	Statin-to-station	**Service area characteristic** (Population, employment, commercial area, office area, pedestrian accessibility) **Station characteristic (**feeder bus) **Other** (travel time)

Cardozo et al. (2012)	Madrid, Spain	Metro network	Geographically weighted regression	Station level	**Service area characteristic** (workers, employment) **Station characteristic** (bus lines, number of track lines)
Gutiérrez et al. (2011)	Madrid, Spain	Metro network	Distance-decay weighted regression	Station level	**Service area characteristic** (employment in the commercial sector, employment in the educational sector, land use mix, nodal accessibility, foreign population, workers) **Station characteristic** (car parking space, urban bus lines, suburban bus lines, number of lines)
Currie et al. (2011)	Australia, North America, Europe	Light rail system	Linear	Route level (demand/route km)	**Service area characteristic** (catchment employment density, **Station characteristic** (service level, integrated ticketing) **Other** (routes being in Europe, speed)
Blainey, (2010)	England and Whales	Heavy rail system	Linear, multiplicative, Geographically weighted regression	Station level	**Service area characteristic** (population, employment, train frequency, distance from the station to the nearest category A-D station) **Station characteristic** (travelcard boundary station, terminus station, the station served by electric trains, car parking)
Lane et al. (2006)	U.S. regions	Light rail and commuter rail	Multiplicative	Station level	**Service area characteristic** (households, employment, headways, bus availability, hours of operations, transportation centre or rail trunk, CBD* density, CBD* employment density, average household size, midday headway) **Station characteristic** (car parking space, bus line, distance to nearest station **Socioeconomic factors** (car ownership, fare) **Other** (speed, travel time)
Kuby et al. (2004)	U.S. cities	Light rail system	Linear	Station level	**Service area characteristic** (employment, population, renters, normalised accessibility) **Station characteristic** (stations that serve airport passenger terminals, international border- station location, park-and-ride spaces, bus feeder, terminal station, transfer station, normalised accessibility **Other** (heating and cooling degree-days)

*CBD is central business district

3. Methodology and Data Sources

3.1. Preference of methodology

Previous researches show that there is no single approach developed to estimate rail or metro demand at new stations. The methodology has been changed based on the different factors which are data availability, the spatial feature of case study area, travel behaviour of users, the general attitude of population toward public transport use, accessibility to the station, land use and most importantly technological methods used for data manipulation. For this study, the previous practices and their applicability in short dissertation term were considered, and some of these techniques were applied to Manchester Metrolink. Chapter 2 examined the all rail demand modelling calibrated by both aggregate and disaggregate data. The feature of data used throughout the modelling process are described, and their sources are presented in the following Section 3.2.

3.2. Data sources

3.2.1. Dependent variable

Daily passenger count data for each tram station was used as a dependent variable for model calibration. Transport for Greater Manchester provides this data via the website: http://www.gmtu. gov.uk/reports/transport2016.htm. The data was given separately for peak and off-peak intervals covering whole one day on weekdays, and at the weekend, so all intervals were summed to reach daily amount. However, only weekday data were used for calibration. The survey day was a different time when between October 2016- February 2017 for each metro line. More details of the obtained data are described below. Although the data collection does not have a standardised day in a defined month and season, it may help the model development.

1. The Bury Line Surveyed in February 2017 - Wednesday
2. The Altrincham Line Surveyed in February 2017 – Thursday
3. The Eccles Line Surveyed in December 2016 – Thursday
4. The East Didsbury Line Surveyed in September 2016 – Thursday
5. The Rochdale Line Surveyed in January 2017 – Wednesday
6. The Ashton-Under-Lyne Line Surveyed in November 2019- Thursday
7. The Manchester Airport Line Surveyed in November 2016- Wednesday
8. The City Zone Surveyed in October 2016 – Wednesday

3.2.2. Census aggregate data

Census data was obtained from Consumer Data Research Centre (CDRC) website that supplying the data separately for each Local Authority District, so the data of 10 different authorities (Manchester, Trafford, Oldham, Salford, Tameside, Bury, Rochdale, Bolton, Stockport, Wigan) within Greater Manchester boundaries are downloaded and combined. There is a wide range of data in this source, with the following variables inserted into the models improved here. All variables are at output area level for Greater Manchester (8684 in total), apart from workplace population data based on output geography, workplace zones (2534 in total). The differences between the two output geography can be seen in Figure 3.1 and Figure 3.2. Same data can also be obtained from Infuse by website http://infuse.ukdataservice.ac.uk in 2011 UK and 2001 England and Wales.

- Resident population (KS101EW)
- Workplace population - residents aged 16 to 74 (WP102EW)
- Car or van availability (QS416EW)
- Method of travel to work (QS701EW)

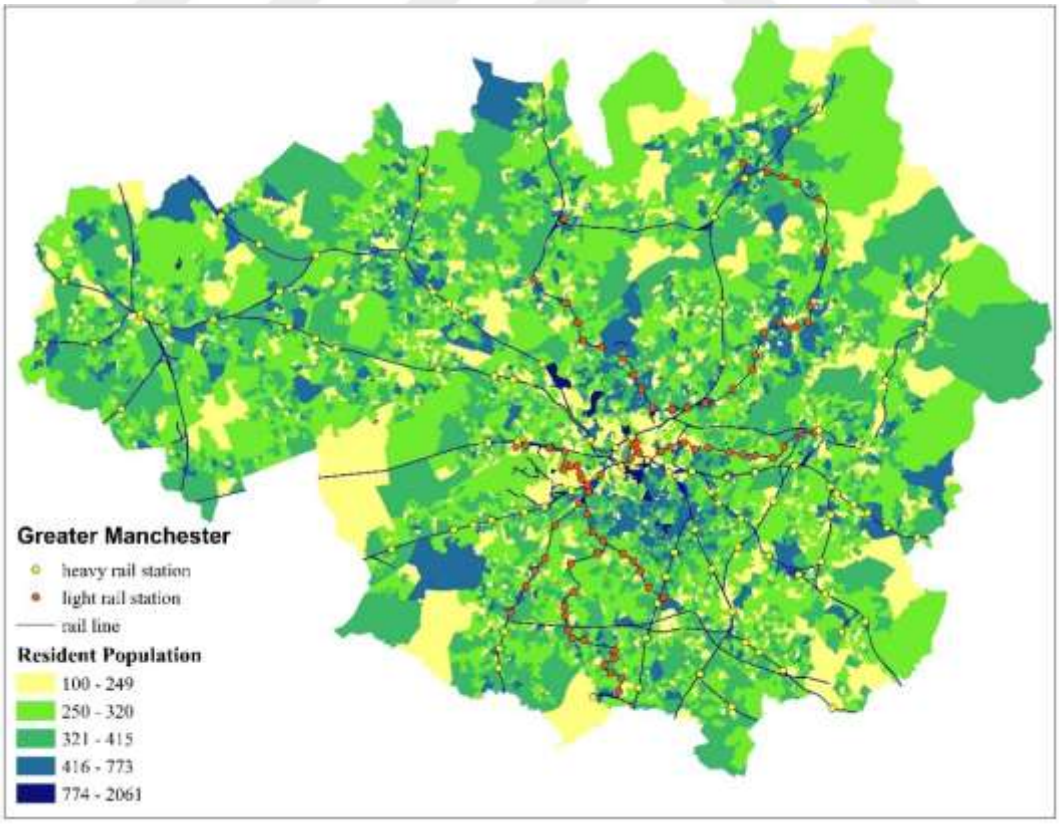

Figure 3.1 The resident population in Greater Manchester

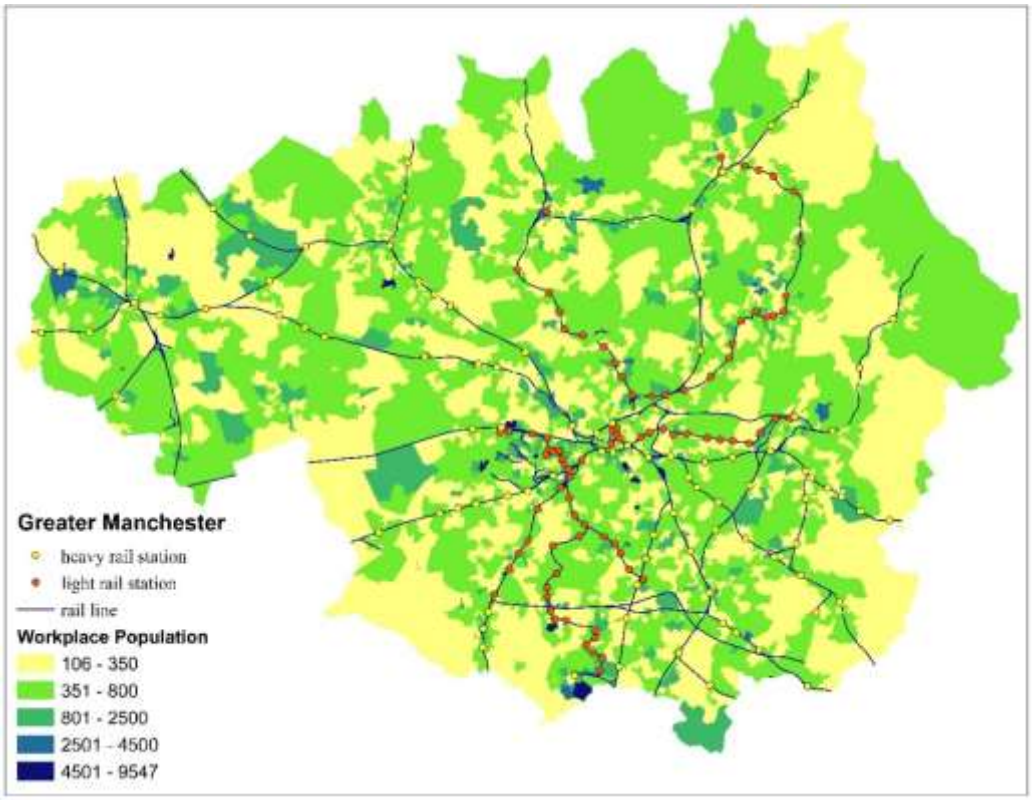

Figure 3.2 The workplace population in Greater Manchester

3.2.3. Weighted centroids of output

Weighted Centroids are a geographical point representing the spatial distribution of the population in each zone within OA, LSOA, MSOA or WZ geography. In other words, the population within each instance of geography is aggregated into a reference point. 2011 OA and WZ Population weighted centroids were obtained via the website: http://webarchive.nationalarchives.gov.uk/20160110200251/http://www.ons.gov.uk/ons/guide-method/geography/products/census/spatial/centroids/index.html

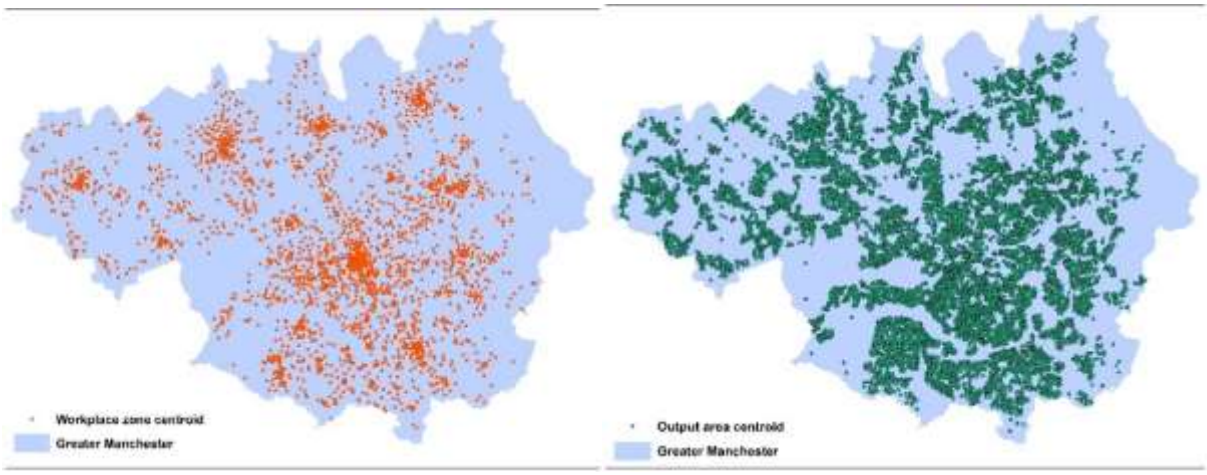

Figure 3.3 The visual comparison of the workplace and output area centroids

3.2.4. Station facilities

The information regarding the station facilities on Manchester Metrolink is available via website; Transport for Greater Manchester. These data are car parking capacity and bicycle standing area.

3.2.5. Tram frequency

The tram frequency of Manchester Metrolink for all service is given as 12 minutes on website Transport for Great Manchester. However, if a tram stop has several services, it means that the frequency of trams will increase. Through this approach, tram number in an hour is counted for each station and used instead of frequency value.

3.2.6. Spatial data

- *Metro and rail station XY coordinate*

The location of the Metrolink and rail stations in Great Manchester is available via the link: https://data.gov.uk/dataset/8faea7ee-eb7d-43dd-b1d4-f01aac4c44d3/metrolink-stops-and-railstations The comparison with Google map data shows that there is conflict among the four station name and their location and then they were fixed before data manipulation.

- *Road and metro network*

The Ordnance Survey (OS) via EDINA Digimap provides the wide range of data format. Firstly, the VectorMap Local data on road and rail networks were used, but on the process of GIS-based network analysis, found that the road network connection does not adequately represent real road link.

Moreover, the Integrated Transport Network of the OS MasterMap data was tried. While there is no rail network layer for this format, the road network data supplied the outstanding quality regarding representation of realism. As a result, road and rail network data were obtained respectively in the format of OS MasterMap and VectorMap Local.

- *Bus stops*

Accessibility to the tram station is an essential factor affecting passenger at the station. It was assumed that the number of bus stop around the station could be used to measure the accessibility factor. National Public Transport Access Nodes (NaPTAN) provides the location of bus stops. The threshold distance covered the bus stops will be defined in Section 4.1.

3.3. Geographical information systems (GIS)

The use of a GIS in transport-related research has been increasing due to its capability to manage, analyse and visualise geographic data, usually on a map. Transport systems have complex properties regarding their multimodal structures and natural integration with other systems, such as land use. However, GIS can allow users to plan and model transport networks of road, rail and buses by visualising data to understand transport trends and patterns (Miller and Shaw, 2001).

More specifically, GIS plays an essential role for demand-forecasting models by supplying a more flexible framework to define service areas. However, the inclusion of zone-based demographic data at the station level of a forecasting model is not straightforward; indeed, delineating a service area around the station is necessary. For instance, Gutiérrez et al. (2011) used GIS to define the threshold service area distance on a straight line or road network around a station to calculate the demographic data within the service area. Additionally, Blainey (2009) improved another service area definition by using a network analysis tool of GIS, which is based on the allocation of all output areas to the station.

With regard to the use of GIS in this project: firstly, the integration of the spatial data with aggregated census data was processed; subsequently, the different varieties of service areas for each station were defined. Thus, the explanatory variables within the service area or distance-based factors (listed below) were used in order to calibrate the trip end model. The visualising function of the GIS was used for the presentation of the data sources and model results as a map. The data groupings, with brief details about their source and derivation, are summarised in Table 3.1.

- Resident population
- Workplace population
- Non-car owning people
- The proportion of car owners
- The number of bus stop
- The distance between tram and nearest railway station
- Nearest distance between the metro station and city centre (Manchester central library)

Table 3.1 The summary of all data used and their sources

All variables	Description	Data source	Derivation
Boarding and alighting	Passenger count data at the station level	Transport Statistics Greater Manchester	Already station-based data
Resident population	The population at output area level	2011 census data from CRDC website	Service area analysis and spatial join in ArcGIS
Workplace population	Workforce, Residents (16-74) within workplace zone	2011 census data from CRDC website	Service area analysis and spatial join in ArcGIS
Car/ Van owners	The number of people having at least one car or van	2011 census data from CRDC website	Service area analysis and spatial join in ArcGIS
Non-car owning population	The number of the having no car	2011 census data from CRDC website	Service area analysis and spatial join in ArcGIS
Trams in an hour	Number of trains on a typical weekday	Transport for Greater Manchester website	Manually counted by using tram network map and timetable
Distance to the nearest rail station (km)	Road network distance from tram station to nearest rail station	Transport for Greater Manchester, data.gov.uk	The measured by using the closest facility tool in ArcGIS
Distance to the city centre (km)	Road network nearest distance from tram station to the city centre point (Manchester central library)	Transport for Greater Manchester, data.gov.uk	The measured by using the closest facility tool in ArcGIS
Car park spaces	Number of car park spaces at the tram station	Transport for Greater Manchester, data.gov.uk	Already station-based data
Bicycle standing area	Number of bicycle stands at the tram station	Transport for Greater Manchester, data.gov.uk	Already station-based data
Terminus (y/n)	Whether a station is at the end of a line	Transport for Greater Manchester website	Already station-based data
Bus stop number	The number of bus stop within the catchment area around the station	NAPTAN	Service area analysis and spatial join in ArcGIS

4. Trip End Model

This section aims to describe the procedure of how the trip end model was applied to light rail stations in Greater Manchester (shown in Figure 4.1). Previous implementations of the trip end model have been carried out by Preston (1991a) and Blainey (2010) for heavy rail stations across Great Britain. Therefore, the approach used in this study is building on Blainey's work and also inspired by Young (2017), and Gutiérrez et al. (2011). Firstly, the independent variables collected was decided based on previous researches. Secondly, the different catchment definitions were tested on independent variables in Greater Manchester, and then the definition which is more realistic or providing a more appropriate result for dataset was used in the process of the model calibration. Finally, several statistical analyses were run to develop the best fit model which forecasting passenger demand at metro stations.

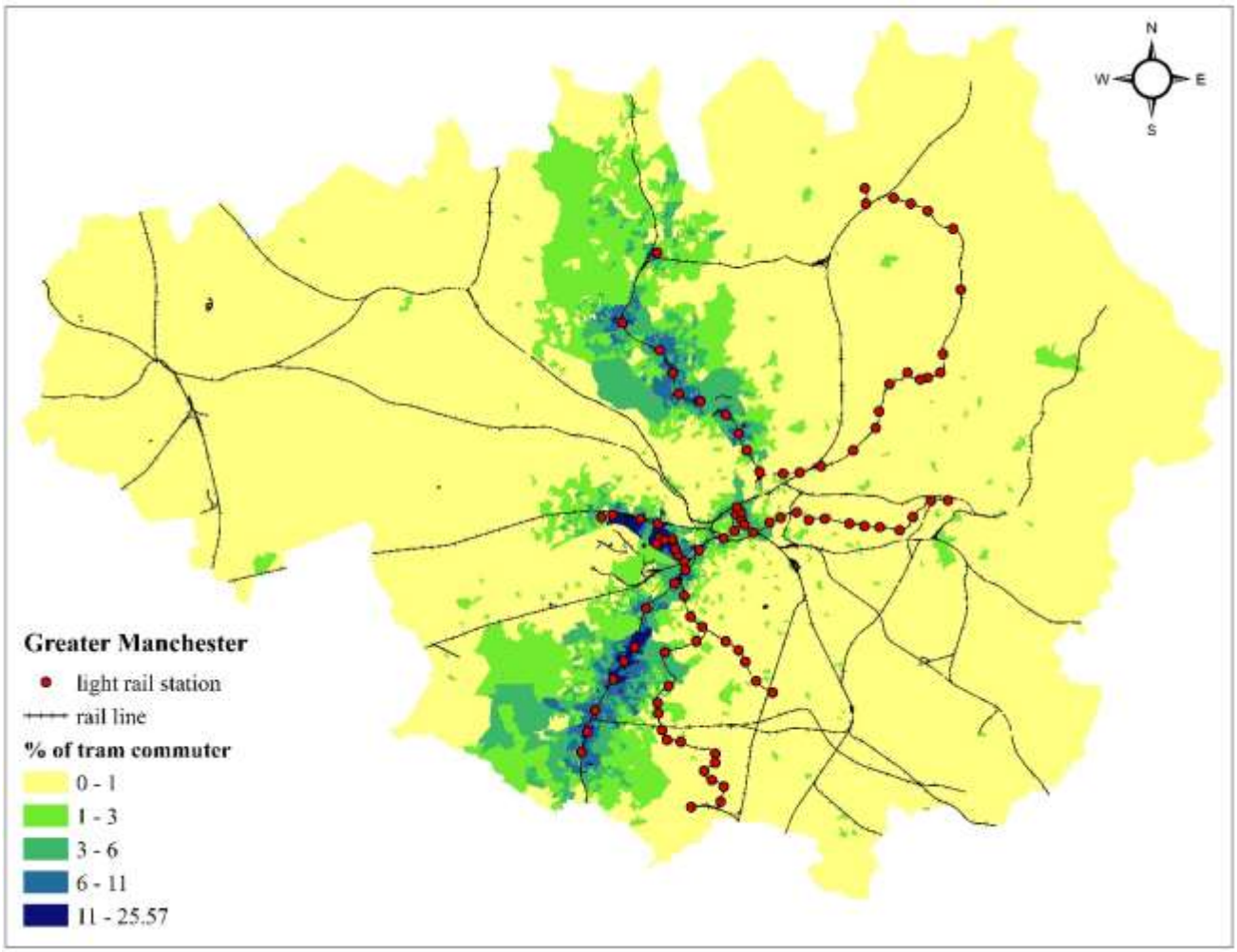

Figure 4.1 The percentage of tram commuter and light rail stations in Greater Manchester

4.1. Catchment (service area) definition

Station usage is affected by a wide range of variables. However, the measurement of the influence of some (employment, population, accessibility) is not straightforward, especially for station level analysis. Thus, it is necessary to define the catchment area around a station to find out whom the station serves or who the potential users of a station are. However, it is hard to analyse the real travel pattern of the passengers because "station choice is not homogeneous within zones, catchments overlap, and catchments vary by access mode and station type" (Young and Blainey, 2018, p.81). Thus, it was assumed that population aggregated to output areas would only prefer the nearest tram station (Blainey, 2009; Zhao et al., 2014). According to this assumption, the relationship between the distance of people to the nearest station and their tendency to use tram is shown in Figure 4.2. The data, number of people commuting by different transport modes from each output area was obtained from the Consumer Data Research Centre (CDRC) website. Then, the percentage of people commuting by tram in each output area in Greater Manchester (8684 in

total) was computed. Finally, these percentages were drawn against the road-network distance from each output area and its closest tram station. The graph seems to illustrate that it appears a correlation between distance to a tram station and the people's tendency to travel to work by train. In a detailed analysis with SPSS use, although Pearson correlation is at a weak level of -0.27, Spearman's rank correlation coefficient gave the moderate level result of -0.45.

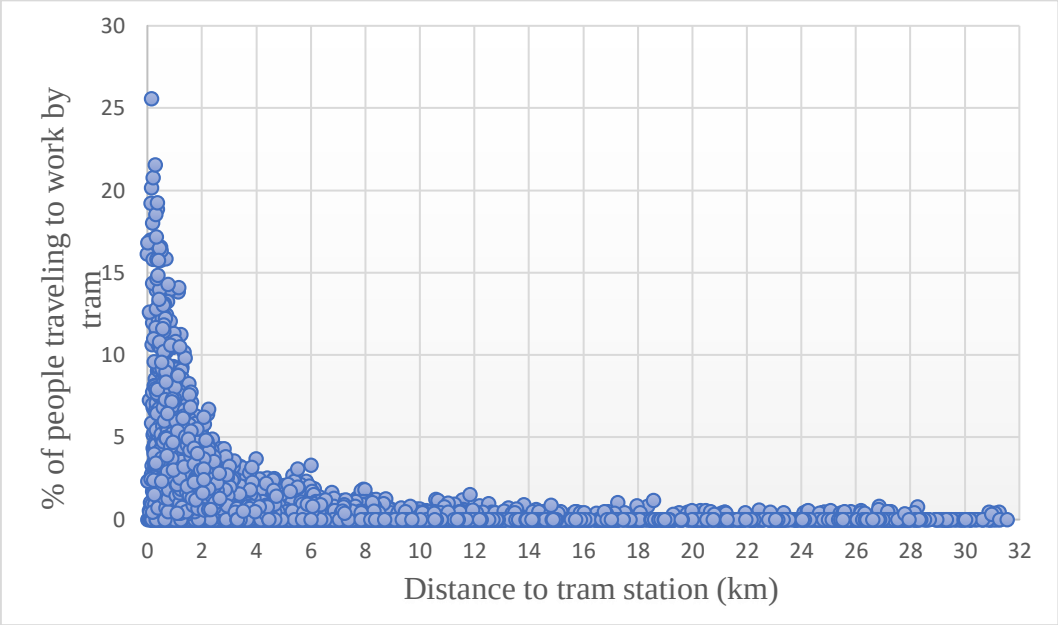

Figure 4.2: The graph showing % of people aggregated to output areas commuting by tram plotted against distance to the closest tram station

4.1.1. Method 1: catchment area definition based on walking distance

The catchment area definition based on walking distance assumes that most passengers prefer non-motorised mode to access/egress station, so this definition concerns with people just around a station (Gutiérrez and García-Palomares, 2008). The previous implementation shows that this catchment definition had been evolved by the new approaches and techniques. For instance, previously, the walking distance for passengers was measured as a straight-line distance in all directions from a station although the measurement on public footpaths represents the actual walking distance. However, the spatial data of public footpath is not available. Therefore, early papers has compared the influence of straight line distance with road network distance. They found that the straight-line measurement inclines to overestimate distance and the population served by the station (Gutiérrez and García-Palomares, 2008). In this perspective, the method based on the straight-line distance was ignored. Then, the network analysis ability of GIS was used to compute catchment areas around stations and to cover the population weighted centroids with resident population, workplace population, car owners, and the number of bus stops within the buffer zone as shown in Figure 4.3.

Apart from the measurement of walking distance, the choice of the distance threshold for a buffer zone is also essential. A very long-distance threshold includes the population who located far from the station and cannot access the station on foot, whereas, a very short distance threshold is likely to ignore the most of the station's passengers. Hence, the distance should be realistic about how far people are willing to walk to a tram station. National Travel Survey recommended that the acceptable walkable distances in the UK (excluding London) are 580 m for bus stops and 1010 m for rail stations (Wyg, 2015). Even though 1010 m can be a plausible maximum distance representing the behaviour of most travellers, this specific acceptation might not be suitable for every station and geography because walking distance varies based on the access route, a network density and station characteristics. For instance, passengers are willing to walk further to access a network with a less density of stations, when the closest line/station is crowded (El-Geneidy et al., 2010).

In this study process, the distance threshold was decided in accordance with statistical analysis, the one with the maximum correlation between light rail demand and each explanatory variable was used for multiple regression analysis. Table 4.1 displays several correlations obtained for explanatory variables within different buffer zone distances and the highest correlation value was highlighted. Also, the calibration results of the Model 4.1 for each the individual variable were presented on this table. Workplace population provides the strongest correlation coefficients than all other variables, whereas, not surprisingly, car owners within any service area gave the very weak and negative correlation coefficients. The negative sign means that an increase in car ownership declines the light rail usage, but the correlation number shows that it is not efficient. Non-car owning people is more likely to use public transport modes, so a strong correlation coefficient was expected, but the result is just 0.428. Moreover, it was thought that the connection of the tram station with other public transport mode could be evaluated by the bus stop number around the station. The outcome shows that the bus stop number has an almost strong relationship with a passenger at the tram station.

$$V_{id} = \alpha + \beta P_{id} \tag{4.1}$$

Where: V_{id} is the estimated passenger boarding and alighting per day at station *i;* P_{id} is the total resident population within *d* meter of station *i, or* P_{id} is the workplace population within *d* meter of station *i, or* Pid is the number people having no car within d meter of station i, or Pid is the car owners within d meter of station i divided by the resident population within d meter of station *i, or* P_{id} is the number of the bus stop within *d* meter of station *i*

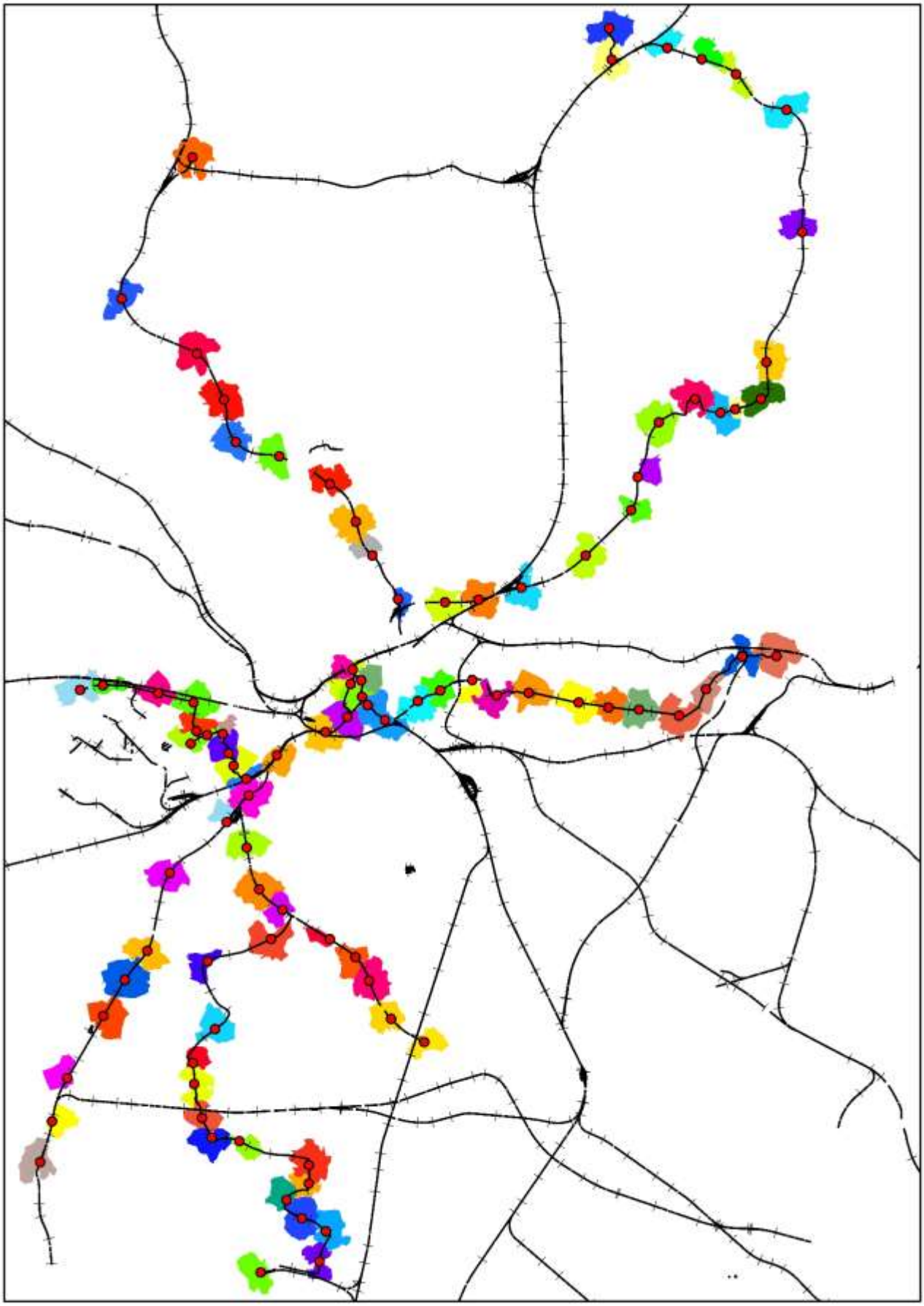

Figure 4.3 The buffer zones with a 600meter threshold distance

Table 4.1 The calibration result of the Model 4.1 and correlation coefficients

Explanatory variables	distance (meter)	correlation with the dependent value	*a* parameter		β parameter		R^2 adj
			value	t statistic	value	t statistic	
Resident population	450	0.091	2870.101	4.017	0.487	0.875	-0.0026
	500	0.045	3123.097	4.323	0.200	0.430	-0.0089
	550	0.026	3219.309	4.212	0.104	0.245	-0.0103
	600	0.031	3186.648	4.180	0.104	0.297	-0.0100
	650	0.006	3339.146	4.356	0.017	0.057	-0.0110
	700	0.158	2327.966	2.893	0.356	1.522	0.0141
	750	0.084	2830.954	3.564	0.195	0.808	-0.0038
	800	0.072	2911.093	3.662	0.149	0.689	-0.0057
Workplace population	450	0.631	1867.974	4.904	0.791	7.755	0.391
	500	0.654	1802.701	4.847	0.723	8.247	0.421
	550	0.632	1838.447	4.807	0.624	7.770	0.392
	600	0.629	1799.121	4.658	0.596	7.728	0.390
	650	0.600	1892.184	4.777	0.512	7.160	0.353
	700	0.638	1997.097	5.415	0.268	7.907	0.401
	750	0.604	1789.553	4.454	0.490	7.226	0.358
	800	0.593	1766.853	4.312	0.467	7.027	0.345
Non-car owning people	450	0.238	2263.558	3.605	5.556	2.338	0.046
	500	0.233	2223.096	3.415	4.793	2.282	0.044
	550	0.203	2266.365	3.254	3.950	1.979	0.031
	600	0.227	2083.392	2.929	3.943	2.226	0.041
	700	0.428	1071.066	1.680	4.109	4.511	0.174
	800	0.276	1609.962	2.113	3.242	2.737	0.066
Proportion of car owners	450	-0.040	3672.340	4.185	-1338.504	-0.386	-0.009
	500	-0.102	4182.768	4.523	-3465.581	-0.980	0.000
	550	-0.099	4182.768	4.338	-3490.974	-0.953	-0.001
	600	-0.035	3688.661	3.558	-1285.024	-0.331	-0.010
	700	0.007	3306.601	2.969	278.508	0.067	-0.011
	800	-0.025	3631.721	3.184	-1024.088	-0.242	-0.010
The number of bus stops	450	0.452	1113.540	1.855	101.529	4.838	0.196
	500	0.448	1071.837	1.750	94.051	4.777	0.192
	550	0.453	945.960	1.509	91.037	4.847	0.196
	600	0.449	921.706	1.450	84.647	4.797	0.193
	700	0.434	913.641	1.391	74.289	4.597	0.180
	800	0.410	871.117	1.245	65.851	4.288	0.159

In summary, service area distance was determined as the following distances for each variable: resident population (700 m); workplace population (500 m); non-car owning people (700 m); and bus stop (550 m). Car ownership will be ignored for the further process due to its very low influence, and the similar result given by previous researches see Blainey (2010) and Gutierrez et al. (2011).

In addition to the method applied above, the availability of the road network for the walking should be considered in terms of passenger's accessibility to the station. For instance, the

motorway is not suitable to walk, so if there were a motorway within the buffer zone, they would be eliminated. However, this project did not apply this approach because of the limited time available. Also, the service areas here did not include many motorways.

4.1.2. Method 2: wider service area definition

While variables just within walking distance gave the good result for workplace population, the correlation coefficients for others are still at a moderate or low level. It might be associated with the omission of the users accessing by cars, cycling or public transport. A further improvement of service area definition was therefore required to maximise the realism of catchment. The method developed by Blainey, (2009) allocates the all output area centroid to their nearest station. It is, therefore, possible to include also the people who are out of walking distance, but within the Greater Manchester boundaries. It was assumed that people within those boundaries are the potential users of the tram stations. However, all users' behaviour in the region is not the same. For example, people living near to the tram station might use the station more than those of living far away, and this tendency to use the tram station gradually decreases while access distance to the station increases (Blainey, 2009; Gutiérrez et al., 2011). Thus, this travel behaviour was represented in the models by inserting the weighting function. A range of different distance-decay weighting functions was tried (4.2 - 4.7), similar to those applied by Blainey, (2009), Young, (2017) to develop a demand forecasting model of local rail stations in the UK.

Power function: $W_a = (d+1)^{-w}$ (4.2)

Inverse exponential function 1: $W_a = e^{-wd}$ (4.3)

Linear function: $W_a = 1 - \frac{d}{w}$ where $\frac{d}{w} \leq 1$ or $W_a = 0$ otherwise (4.4)

Inverse exponential function 2: $W_a = exp[-0.5\ (\frac{d}{w})^2]$ (4.5)

Logit function: $W_a = \frac{1/\ exp[w\ (d+1)]}{1-(1/\ exp[w\ (d+1)])}$ (4.6)

Quadratic function: $W_a = 1 - wd^2$ where $wd^2 \leq 1$ or $W_a = 0$ otherwise (4.7)

Where: W_a is the weighting function attached to output area *a; d* is the distance (km) from output area *a* to its nearest station; *w* is a parameter defined by trying the number by 0.25 increase between 0.25 and 4.25

It is important to mention that allocating the output area centroids to only their closest tram station is unrealistic. Light rail stations have a compact structure in Greater Manchester whereas heavy rail station widely distributed in regions (see Figure 4.4). Some zones are likely to prefer using rail station instead of tram the station. Therefore, output area centroids were allocated to both light and heavy rail stations to increase the realism of servicthe e area. These weighting functions above were multiplied by each output area value and then summed for each tram station (93 in total) to find the people served as shown on model 4.8. However, there is one weakness of that this method did not allocate to four tram stations (Ashton Moss, Media City UK, Oldham Central, Pomona) when the output area centroids (8684 in total) are used. Additionally, when the method applied for workplace population centroid (2534 in total), there is no allocation to two tram stations (Kingsway Business Park and St. Werburgh's Road).

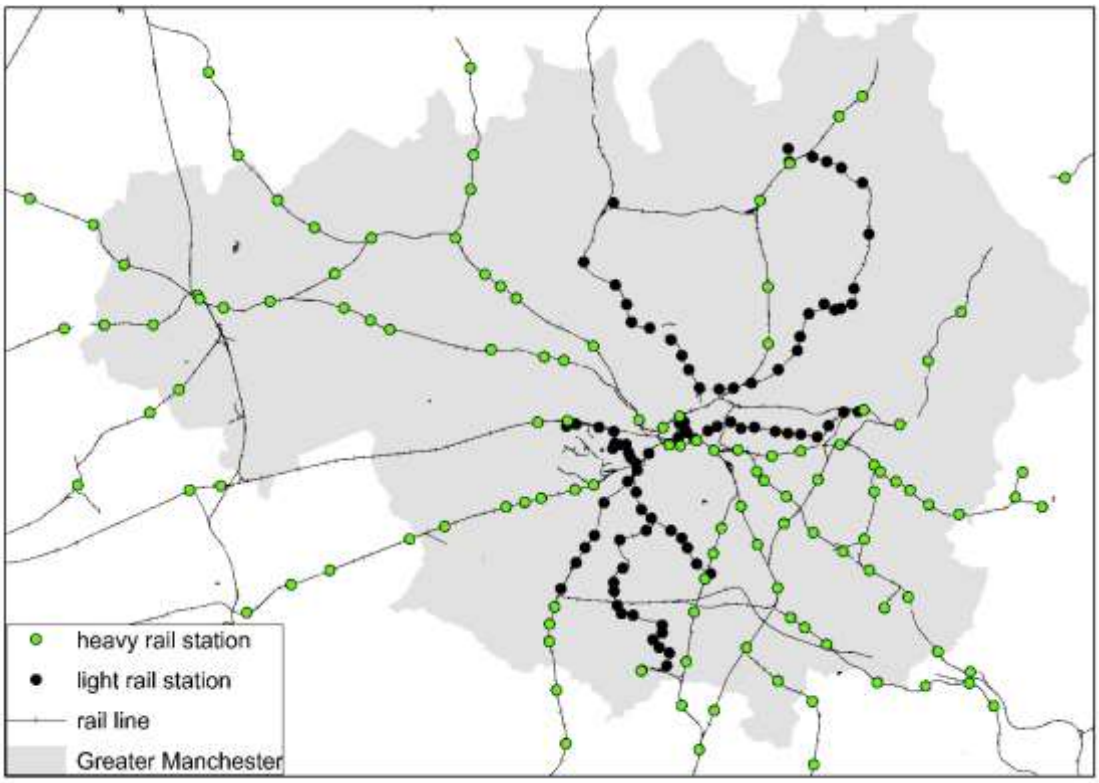

Figure 4.4 The light and heavy rail stations in Greater Manchester

Overall, the model calibration results are given in Table 4.2 with correlation coefficient. For the resident population, the weighting function 4.5 with a *w* value of 4.25 gave the best fit model (highest R^2 value 0.036). Also, it is obvious that the result is better than that of defined by the walking distance (highest R^2 adj value 0.0141). However, the value for non-car owning people has a decrease, and there is no change in workplace population.

$$V_i = \alpha + \beta \sum_{a}^{n} P_a W_a \qquad (4.8)$$

Where: W_a is the weighting function attached to output area *a;* P_a is the resident population in output area *a*, or P_a is the non-car owning people in output area *a, or* P_a is the workplace population in output area *a;* α, β are parameters produced by calibration

Table 4.2 The calibration result of Model 4.8 with the distance-decay weighting function

Explanatory variable	Weighting function	Best *w* value	correlation with the dependent value	a parameter		β parameter		R^2 adj
				value	t statistic	value	t statistic	
Resident population	4.2	0.25	0.119	2927.188	5.101	0.049	1.144	0.003
	4.3	0.25	0.087	2998.134	4.852	0.048	0.834	-0.003
	4.4	0.50	0.122	2807.185	4.381	1.790	1.172	0.004
	4.5	4.25	0.216	3141.507	7.360	0.203	2.110	0.036
	4.6	0.25	0.060	3069.382	4.522	0.020	0.575	-0.007
	4.7	4.25	0.139	2722.724	4.238	1.434	1.338	0.009
Non-car owning people	4.2	4	0.120	2627.769	3.396	5.483	1.150	0.003
	4.3	4	0.155	2514.954	3.540	10.383	1.496	0.013
	4.4	0.5	0.217	2492.565	4.253	14.732	2.119	0.037
	4.5	4.25	0.209	3099.834	7.137	2.384	2.040	0.033
	4.6	4.25	0.157	2526.869	3.614	798.313	1.512	0.014
	4.7	4.25	0.239	2392.499	4.080	11.429	2.345	0.047
Workplace population	4.2	3.75	0.650	1459.866	3.675	1.730	8.169	0.417
	4.3	3.00	0.650	1575.998	4.049	1.635	8.163	0.416
	4.4	0.75	0.650	1761.413	4.671	1.191	8.153	0.416
	4.5	0.25	0.432	1809.689	3.534	0.297	4.571	0.178
	4.6	3.00	0.650	1580.534	4.064	32.076	8.162	0.416
	4.7	2.75	0.653	1841.454	4.978	1.049	8.234	0.421

Blainey, (2009) also suggest that the approach used above can be enhanced by using access time rather than access distance, especially if there is a variation in the types of road. This appears more realistic than using road network distances. Therefore, the allocation of centroids to stations was also applied based on road travel time. The average speed variations used for different road type is displayed in Table 4.3, and the weighting functions (4.9 - 4.14) used for method were reorganised for travel time. Moreover, the model weakness still exists, but the problem station was partially changed. There is no allocation of any output area centroids (8684 in total) to four tram stations (Etihad Campus, Media City UK, Oldham Central, Pomona), and of the workplace population centroids (2534 in total) to five tram stations (Kingsway Business Park, St. Werburgh's Road, Media City UK, Oldham Central, and Pomona). The stations given unrealistic result increased at the application of workplace population centroid.

Power function: $W_a = (t+1)^{-w}$ (4.9)

Inverse exponential function 1: $W_a = e^{-wt}$ (4.10)

Linear function: $W_a = 1 - \frac{t}{w}$ where $\frac{t}{w} \leq 1$ or $W_a = 0$ otherwise (4.11)

Inverse exponential function 2: $W_a = exp[-0.5\,(\frac{t}{w})^2]$ (4.12)

Logit function: $W_a = \frac{1/\,exp[w\,(t+1)]}{1-(1/\,exp[w\,(t+1)])}$ (4.13)

Quadratic function: $W_a = 1 - wt^2$ where $wt^2 \leq 1$ or $W_a = 0$ otherwise (4.14)

Where: W_a is the weighting function attached to output area *a; t* is the time (minutes) from output area *a* to its nearest station; *w* is a parameter defined by trying the number by 0.25 increase between 0.25 and 4.25

Table 4.3 The speed variations used for travel time calculation (Blainey, 2009)

Road Type	**Average Speed (kph)**
Motorway	80
A Road	65
B Road	40
Minor Road	30
Others	25

Model calibration result based on the travel time is presented in Table 4.4 for each variable. The comparison of distance-based and time-based allocation method shows that the R^2 adj value of resident population increased to 0.093 with the weighting functions 3.11 with a w value of 2.5 whereas the R^2 adj value unexpectedly declined from 0.417 to 0.340 for workplace population. There is no specific reason found for the decrease in the value of the workplace population, but the number of the inaccurate stations encountered increases by 3 at time-based allocation, so the more error probably results in the lower R^2 adj value. Moreover, although there is an improvement for the value of non-car owning people, it is still less than the value obtained by walking distance definition (method 1).

Table 4.4 The calibration result of Model 4.8 with the time-decay weighting function

Explanatory variables	Weighting function	Best *w* value	correlation with dependent value	a parameter		β parameter		R^2 adj
				value	t statistic	value	t statistic	
Resident population	4.9	0.25	0.176	2732.87	4.87	0.07	1.70	0.020
	4.10	2.25	0.180	2426.37	3.55	1.96	1.75	0.022
	4.11	1.00	0.226	2454.00	4.19	2.27	2.21	0.040
	4.12	4.25	0.211	2980.41	6.55	0.15	2.06	0.034
	4.13	2.25	0.178	2440.22	3.58	17.83	1.73	0.021
	4.14	2.50	0.321	2243.10	4.22	2.35	3.23	0.093
Non-car owning people	4.9	3.5	0.278	2054.972	3.274	15.768	2.756	0.067
	4.10	2.25	0.295	2082.126	3.491	14.366	2.945	0.077
	4.11	1.25	0.332	2102.284	3.824	9.789	3.358	0.100
	4.12	4.25	0.225	2894.870	6.214	1.491	2.198	0.040
	4.13	2.25	0.293	2098.875	3.528	130.907	2.922	0.076
	4.14	2.5	0.377	2098.875	4.811	10.488	3.883	0.133
Workplace population	4.9	2.00	0.582	1799.539	4.350	1.013	6.828	0.331
	4.10	1.25	0.582	1901.699	4.687	0.881	6.820	0.331
	4.11	1.25	0.587	2178.300	5.686	0.848	6.916	0.337
	4.12	0.25	0.470	1814.656	3.755	0.262	5.075	0.212
	4.13	1.00	0.581	1874.028	4.591	1.615	6.812	0.330
	4.14	0.75	0.590	2187.497	5.737	0.651	6.964	0.340

4.1.3. Method 3: two-stage catchment definition

Catchment area definition is an important procedure to find a real potential of the station. In this study, the first method captured only the potential users within walking distance as thought that people living in walkable distance are more willing to use station, but out of this area and motorised access modes were ignored. Then, the second method covered all potential users within Greater Manchester boundaries on basis of access-distance and access-time. It also considered that users' willingness to use station gradually decrease as distance or time increase. Thus, this behaviour was tried to represent by weighting functions in the model However, the search of Young, (2017) indicated that the willingness of people living within walkable distance does not decrease by distance change, so he defended that two-stage catchment definition would give more appropriate result. In this perspective, he redesigned the weighting functions, with no decay within specified distance and time, see functions (4.15-4.20) below.

Power function: $W_a = \begin{cases} (d+1)^{-w} & \text{if } d > x \\ 1 & \text{otherwise} \end{cases}$ (4.15)

Inverse exponential function 1: $W_a = \begin{cases} e^{-wd} & if\ \ d > x \\ 1 & otherwise \end{cases}$ (4.16)

Linear function: $W_a = \begin{cases} 1 - \frac{d}{w} & if\ \ d > x \\ 1 & otherwise \end{cases}$ where $\frac{d}{w} \leq 1 \ \ or \ \ W_a = 0$ otherwise (4.17)

Inverse exponential function 2: $W_a = \begin{cases} exp[-0.5\ (\frac{d}{w})^2] & \text{if } d > x \\ 1 & \text{otherwise} \end{cases}$ (4.18)

Logit function: $W_a = \begin{cases} \frac{1/\,exp[w\,(d+1)]}{1-(1/\,exp[w\,(d+1)])} & if\ \ d > x \\ 1 & otherwise \end{cases}$ (4.19)

Quadratic function: $W_a = \begin{cases} 1 - wd^2\ if\ \ d > x \\ 1 & otherwise \end{cases}$ where $wd^2 \leq 1\ or\ W_a = 0$ otherwise (4.20)

Where: W_a is the two stage distance decay weighting function attached to the output area *a; d* is the distance (km) from the output area *a* to its nearest station; *w* is a parameter defined by trying the number by 0.25 increase between 0.25 and 4.25; *x* is the walkable distance (km) defined by 0.1 raise between 0.3 and 0.8

It is important to note that this third method was applied for all variables as same as previously described in the Section 4.1.2. The only difference is the design of distance-decay weighting functions. After the allocation of output area centroids to stations, the redesigned weighting functions were processed to find demand at station, see model 4.21. The calibration of a model for each variable is summarised in Table 4.5. Two stage distance-decay weighting function did not improve the R^2 adj value of resident population, but the workplace population with the weighting function 4.17 with a *w* value of 4 gave the highest R^2 value of 0.434 compared to method 1 and 2.

$$V_i = \alpha + \beta \sum_{a}^{n} P_a\, W_a \tag{4.21}$$

Where: W_a is two stage distance decay weighting function attached to output area *a;* P_a is the resident population in output area *a,* or P_a is the non-car owning people in output area *a, or* P_a is the workplace population in output area *a;* α, β are parameters produced by calibration

Table 4.5 The calibration result of Model 4.21 with the distance-decay weighting function

Explanatory variables	Weighting function	Best w value	distance (km)	correlation with the dependent value	a parameter		β parameter		R^2 adj
					value	t statistic	value	t statistic	
Resident population	4.17	0.25	0.3	0.256	2493.094	4.643	2.784	2.529	0.055
	4.18	4.25	0.4	0.234	2977.877	6.687	0.222	2.300	0.045
	4.18	4.25	0.5	0.226	2911.012	6.308	0.214	2.218	0.041
	4.18	4.25	0.6	0.214	2835.652	5.821	0.202	2.081	0.035
	4.18	4.25	0.7	0.213	2747.145	5.376	0.196	2.083	0.035
	4.18	4.25	0.8	0.199	2719.078	5.083	0.173	1.936	0.029
Non-car owning people	4.17	0.25	0.3	0.298	2486.733	4.958	14.793	2.977	0.079
	4.17	0.25	0.4	0.238	2402.352	4.111	8.050	2.338	0.046
	4.18	4.25	0.5	0.224	3091.806	7.168	0.211	2.191	0.040
	4.19	0.25	0.6	0.224	3073.959	7.086	0.212	2.191	0.040
	4.18	4.25	0.7	0.227	3050.767	6.994	0.215	2.226	0.041
	4.18	4.25	0.8	0.225	3036.603	6.916	0.212	2.205	0.040
Workplace population	4.20	2.00	0.3	0.650	1826.065	4.902	0.887	8.170	0.417
	4.20	2.50	0.4	0.655	1849.277	5.016	0.865	8.263	0.422
	4.18	4.00	0.5	0.663	1717.913	4.621	0.709	8.456	0.434
	4.18	4.25	0.6	0.660	1646.145	4.348	0.635	8.372	0.429
	4.18	4.00	0.7	0.627	1609.004	4.009	0.554	7.678	0.386
	4.18	4.00	0.8	0.633	1432.994	3.490	0.555	7.809	0.395

Previous experience shows that the outcome of time-based allocation for the resident population is better than distance-based. Hence, the two-stage methodology was repeated on the basis of travel time. (See the weighting functions used on the process 4.22- 4.27.) Table 4.6 shows the outcome of model calibration. When compared the two stage distance decay and two stage time-decay function, it is seen that the R^2 adj value of resident population reached the highest value of 0.096. It is not surprising due to the experience gained by the second method, observed that the R^2 adj value decreased for the workplace population.

Power function: $W_a = \begin{cases} (t+1)^{-w} & \text{if } t > x \\ 1 & \text{otherwise} \end{cases}$ (4.22)

Inverse exponential function 1: $W_a = \begin{cases} e^{-wt} & if\ \ t > x \\ 1 & otherwise \end{cases}$ (4.23)

Linear function: $W_a = \begin{cases} 1 - \frac{t}{w} & if\ \ t > x \\ 1 & otherwise \end{cases}$ where $\frac{t}{w} \leq 1$ or $W_a = 0$ otherwise (4.24)

Inverse exponential function 2: $W_a = \begin{cases} exp[-0.5\,(\frac{t}{w})^2] & \text{if } t > x \\ 1 & \text{otherwise} \end{cases}$ (4.25)

Logit function: $W_a = \begin{cases} \frac{1/\,exp[w\,(t+1)]}{1-(1/\,exp[w\,(t+1)])} & if\ \ t > x \\ 1 & otherwise \end{cases}$ (4.26)

Quadratic function: $W_a = \begin{cases} 1 - wt^2 & if \ \ t > x \\ 1 & otherwise \end{cases}$ where $wt^2 \leq 1 \ \ or \ \ W_a = 0$ otherwise (4.27)

Where: W_a is the two stage time decay weighting function attached to the output area *a; t* is the time (minutes) from the output area *a* to its nearest station; *w* is a parameter defined by trying the number by 0.25 increase between 0.25 and 4.25; *x* is the time (minutes) defined by 0.5 raise between 0.5 and 2.

Table 4.6 The calibration result of Model 4.21 with the time-decay weighting function

Explanatory variables	Weighting function	Best *w* value	minute (*x*)	correlation with the dependent value	a parameter		β parameter		R^2 adj
					value	t statistic	value	t statistic	
Resident population	4.27	2.50	0.5	0.325	2223.263	4.180	2.406	3.279	0.096
	4.25	4.25	1.0	0.241	2722.534	5.512	0.163	2.367	0.048
	4.25	4.25	1.5	0.230	2575.847	4.746	0.139	2.255	0.043
	4.25	4.00	2.0	0.186	2691.504	4.791	0.083	1.805	0.024
Non-car owning people	4.27	2.50	0.5	0.340	2300.201	4.559	4.080	3.453	0.106
	4.27	0.75	1.0	0.297	2098.681	3.560	4.016	2.969	0.078
	4.26	3.50	1.5	0.247	2168.583	3.373	2.249	2.433	0.051
	4.26	0.25	2.0	0.222	2866.327	6.048	0.149	2.168	0.039
Workplace population	4.27	0.50	0.5	0.589	2073.570	5.326	0.566	6.948	0.339
	4.25	4.25	1.0	0.617	1835.948	4.696	0.463	7.473	0.373
	4.25	4.25	1.5	0.554	1722.205	3.940	0.372	6.356	0.300
	4.22	2.00	2.0	0.543	1596.380	3.496	0.389	6.178	0.288

As a conclusion, three methods were tried to be able to define the best-fit catchment for different variables. The highest R^2 adj value obtained from different definitions for each variable was presented to compare in Table 4.7. The number of bus stops just around the tram station can reasonably evaluate the connectivity between transport modes, so only buffer zone approach was used and reached the highest R^2 adj value with 550meter. The influence of the proportion of car owners was observed with a buffer zone and received the very poor result, so it was neglected the further process because of the possibility that the low influence could continue. This decision is also based on impression gained by previous research, see (Blainey, 2010; Gutiérrez et al., 2011). All these methods helped to improve the impact of the resident population on the passenger demand, but the result retained at the low level of 0.096, with the two stage time decay weighting function 4.25 with a *w* value of 2.5 and an *x* value of 1 minute. In addition, the highest influence of the workplace population on (Blainey, 2010) demand was gained by two stage distance decay weighting function 4.18 with a *w* value of 4 and an *x* value of 0.5kilometer, but the result of this complicated method is not much better than that of a buffer zone, so the necessity is questionable.

Table 4.7 The summary of highest R^2 adj value obtained by all catchment definitions

Explanatory variables	Method 1 Buffer zone	Method 2 Distance decay function	Method 2 Time decay function	Method 3 Two stage distance decay function	Method 3 Two stage time decay function
Resident population	0.014	0.036	0.093	0.055	0.096
Non-car owning people	0.174	0.047	0.133	0.079	0.106
Workplace population	0.421	0.421	0.340	0.434	0.373
The proportion of car owners	0	-	-	-	-
The number of bus stop	0.196	-	-	-	-

4.2. Multiple regression model

The individual contribution of some independent variables which need to have a service area definition was tested and presented in Section 4.1. Now, with the multiple regression model, the inclusion of multiple independent variables in one model will analyse. Therefore, to explore the impact of every explanatory variables on rail demand, the first step computed the correlation coefficient between them and revealed that tram in an hour and workplace population provided the strongest correlations, followed by bus stop number, bus or rail feeder, non-car owning people and resident population (see Table 4.8). Furthermore, car parking spaces, bicycle stands and distance from heavy rail to light rail station showed unexpected coefficient signs. Thus, they will be ignored in the modelling process.

While the correlation with the dependent variable is highly desirable, the strong intercorrelation among explanatory variables is undesirable. In this case, the number of bus stop has a high correlation value of 0.726 with workplace population and of 0.579 with bus or rail feeder. It was thought that the inclusion of the bus stop number is likely to affect the model negatively. Moreover, there is a correlation at a critical level between the resident population and non-owning people, their presence in a model will be tested.

Initial calibration established on the two variables with highest correlation, tram in an hour and workplace population within the defined service area (Model 4.28), then the number of the bus stop was added (Model 4.29). The calibration results were summarised in Table 4.9 for both. However, the bus stop variable became insignificant, as estimated, so it will be eliminated for the next step.

Table 4.8 The correlation coefficients among all variables

Explanatory variables	Boarding alighting	Trams in hour	Workplace population within catchment	Terminus (1=yes, 0=not)	Bus or rail feeder (1=yes, 0=not)	Bicycle stands	Car parking spaces	Distance to city centre	Distance to heavy rail station	Non-car owning people within catchment	Resident population within catchment	Number of bus stop within catchment
Boarding alighting	1.000	0.805	0.663	0.127	0.430	-0.308	-0.044	-0.343	-0.180	0.430	0.344	0.451
Trams in hour	0.805	1.000	0.449	-0.083	0.199	-0.450	-0.138	-0.532	-0.205	0.420	0.337	0.221
Workplace population*	0.663	0.449	1.000	0.274	0.317	-0.218	-0.079	-0.197	-0.087	0.407	0.216	0.726
Terminus	0.127	-0.083	0.274	1.000	0.571	0.066	0.055	0.208	-0.155	-0.115	-0.023	0.444
Bus or rail feeder	0.430	0.199	0.317	0.571	1.000	-0.004	0.139	0.110	-0.341	0.226	0.177	0.579
Bicycle stands	-0.308	-0.450	-0.218	0.066	-0.004	1.000	0.201	0.382	0.201	-0.282	-0.168	-0.106
Car parking spaces	-0.044	-0.138	-0.079	0.055	0.139	0.201	1.000	0.239	0.206	-0.208	-0.176	-0.009
Distance to city centre	-0.343	-0.532	-0.197	0.208	0.110	0.382	0.239	1.000	0.306	-0.384	-0.321	0.014
Distance to rail station	-0.180	-0.205	-0.087	-0.155	-0.341	0.201	0.206	0.306	1.000	-0.341	-0.235	-0.131
Non-car owning people**	0.430	0.420	0.407	-0.115	0.226	-0.282	-0.208	-0.384	-0.341	1.000	0.513	0.486
Resident population***	0.344	0.337	0.216	-0.023	0.177	-0.168	-0.176	-0.321	-0.235	0.513	1.000	0.269
Number of bus stop****	0.451	0.221	0.726	0.444	0.579	-0.106	-0.009	0.014	-0.131	0.486	0.269	1.000

* within catchment defined by method 3

** within catchment defined by method 1

*** within catchment defined by method 3

**** within catchment defined by method 1

$$V_i = \alpha + \delta F_i + \beta \sum_a^n P_a W_a \quad (4.28)$$

Where: F_i is the number of tram in an hour; P_a is the workplace population in output area *a*; W_a is two stage distance decay weighting function attached to the output area *a*

$$V_i = \alpha + \delta F_i + \beta \sum_a^n P_a W_a + \lambda T_i^{550} \quad (4.29)$$

Where: T_i^{550} is the number of bus stop located within 550meter walking distance of station i

Table 4.9 Summarised calibration result of Models 4.28 - 4.29

Model		4.28		4.29
Parameter	Value	t statistic	Value	t statistic
Constant (α)	-1556.664	-4.057	-1887.824	-3.785
Tram in an hour (δ)	481.700	11.046	489.325	11.069
Workplace population (β)	0.404	6.565	0.338	3.824
Bus stop number (λ)	n/a	n/a	15.816	1.039
R^2 adj		0.757		0.757

The connection of tram stations with other public transport modes might have important influence on the passenger demand. However, the catchment definition did not include the passengers provided by feeder mode. Therefore, the next variable was bus or rail feeder to evaluate whether the feeder improves the model calibration. If there was a bus or heavy rail connection, a dummy variable was inserted as 1, otherwise 0. Table 4.10 summarises the calibration result of Model 4.30 and shows that it has a major improvement expectedly. Additionally, the inclusion of the non-car owning people and resident population were respectively tested, but both did not give a significant result.

$$V_i = \alpha + \delta F_i + \beta \sum_a^n P_a W_a + \rho B_i \quad (4.30)$$

Where: B_i is the dummy variable if station *i* has a bus or rail feeder it is1, otherwise 0

$$V_i = \alpha + \delta F_i + \beta \sum_a^n P_a W_a + \rho B_i + \tau J_i^{700} \quad (4.31)$$

Where: J_i^{700} is the non-car owning people located within 700meter walking distance of station i

$$V_i = \alpha + \delta F_i + \beta \sum_a^n P_a W_a + \rho B_i + \beta \sum_a^n L_a W_{2a} \quad (4.32)$$

Where: L_a is resident population; W_{2a} is two stage time decay weighting function

Table 4.10 Summarised calibration result of Models 4.30 - 4.31- 4.32

Model		4.30		4.31		4.32
Parameter	Value	t statistic	Value	t statistic	Value	t statistic
Constant (α)	-1692.728	-4.760	-1669.159	-4.289	-1747.270	-4.788
Tram in an hour (δ)	470.316	11.659	472.143	11.169	462.995	11.084
Workplace population (β)	0.341	5.810	0.343	5.647	0.339	5.752
Bus or rail feeder (ρ)	2186.051	4.087	2193.876	4.060	2151.920	3.995
Non-car owning people (τ)	n/a	n/a	-0.081	-0.154	n/a	n/a
Resident population (β)	n/a	n/a	n/a	n/a	0.263	0.703
R^2 adj		0.793		0.791		0.792

Other explanatory variables (distance to city centre and terminus) were separately added on Model 4.33 - 4.34 and observed that there is no further improvement as displayed in Table 4.11. To examine the underlying reason, the residuals from Model 4.30 are displayed in Figure 4.5, indicating of which station has higher outlier value affecting the model fit. While passenger demand at some stations (Victoria, Piccadilly, Cornbrook) are underestimated, the value of St Peter's Square tram station is highly overestimated. Smaller observed demand might be related to spatial patterns and characteristic of stations, for example, Piccadilly and Victoria stations are also major heavy rail interchanges and close to important attraction points and, so might attract people for different aims. Also, Cornbrook is the intersection point for all Metrolink, so its demand is at a high level due to the not only pedestrians but also transfer between the lines. The smaller demand observation for St Peter's Square is likely to relate to its location. Other very close stations to the St Peter's Square station probably abstract its potential users. Therefore, it was decided to eliminate these stations to have high predictive power. A total of 89 stations were used for model calibration.

$$V_i = \alpha + \delta\, F_i \; + \beta \sum_{a}^{n} P_a\, W_a + \rho B_i + \eta D_i \tag{4.33}$$

Where: D_i is the distance from station *i* to the city centre (Manchester centre library)

$$V_i = \alpha + \delta\, F_i \; + \beta \sum_{a}^{n} P_a\, W_a + \rho B_i + \gamma H_i \tag{4.34}$$

Where: H_i is the dummy variable if station *i* is the end of the line, it is 1, otherwise 0

Table 4.11 Summarised calibration result of Models 4.33- 4.34

Model		4.33		4.34
Parameter	Value	t statistic	Value	t statistic
Constant (α)	-2104.745	-3.085	-1637.923	-4.392
Tram in an hour (δ)	488.279	10.228	461.183	10.704
Workplace population (β)	0.342	5.802	0.352	5.748
Bus or rail feeder (ρ)	2086.845	3.764	2446.233	3.741
Distance to city centre (η)	0.035	0.708	n/a	n/a
Terminus (ɣ)	n/a	n/a	-616.393	-0.691
R^2 adj		0.792		0.791

The exclusion of these 4 stations improved the Model 4.30 from 0.793 to 0.834, giving summary result in Table 4.12. This increased result encourages to continue modelling without 4 stations because the high predictive model is essential for proposed new stations.

Table 4.12 Summarised calibration result of Models 4.30 without 4 stations

Model		4.30
Parameter	Value	t statistic
Constant (α)	-1059.433	-3.905
Tram in an hour (δ)	351.151	10.286
Workplace population (β)	0.478	11.332
Bus or rail feeder (ρ)	1027.820	2.724
R^2 adj		0.834

After removing some stations, resident population was again inserted into the equation, similar to model 4.32 and outcome still shows that resident population within the predefined catchment is insignificant. Thus, with a stepwise regression model, all two stage time decay weighting function was tried to include in the model, and the population with weighting function 4.27 with a *w* value of 0.5 and an *x* value of 1 minute gave the significant result, Table 4.13 allows comparing both models.

Table 4.13 Summarised calibration result of Models 4.32 without 4 stations

Model		4.32		4.32
Parameter	Value	t statistic	Value	t statistic
Constant (α)	-1102.596	-4.021	-1263.578	-4.411
Tram in an hour (δ)	340.294	9.552	336.936	9.808
Workplace population (β)	0.478	11.345	0.482	11.605
Bus or rail feeder (ρ)	1018.143	2.699	997.650	2.685
Resident population (ƥ)	0.278	1.057	n/a	n/a
Resident population within new catchment (ƥ)	n/a	n/a	0.185	1.961
R^2 adj		0.834		0.839

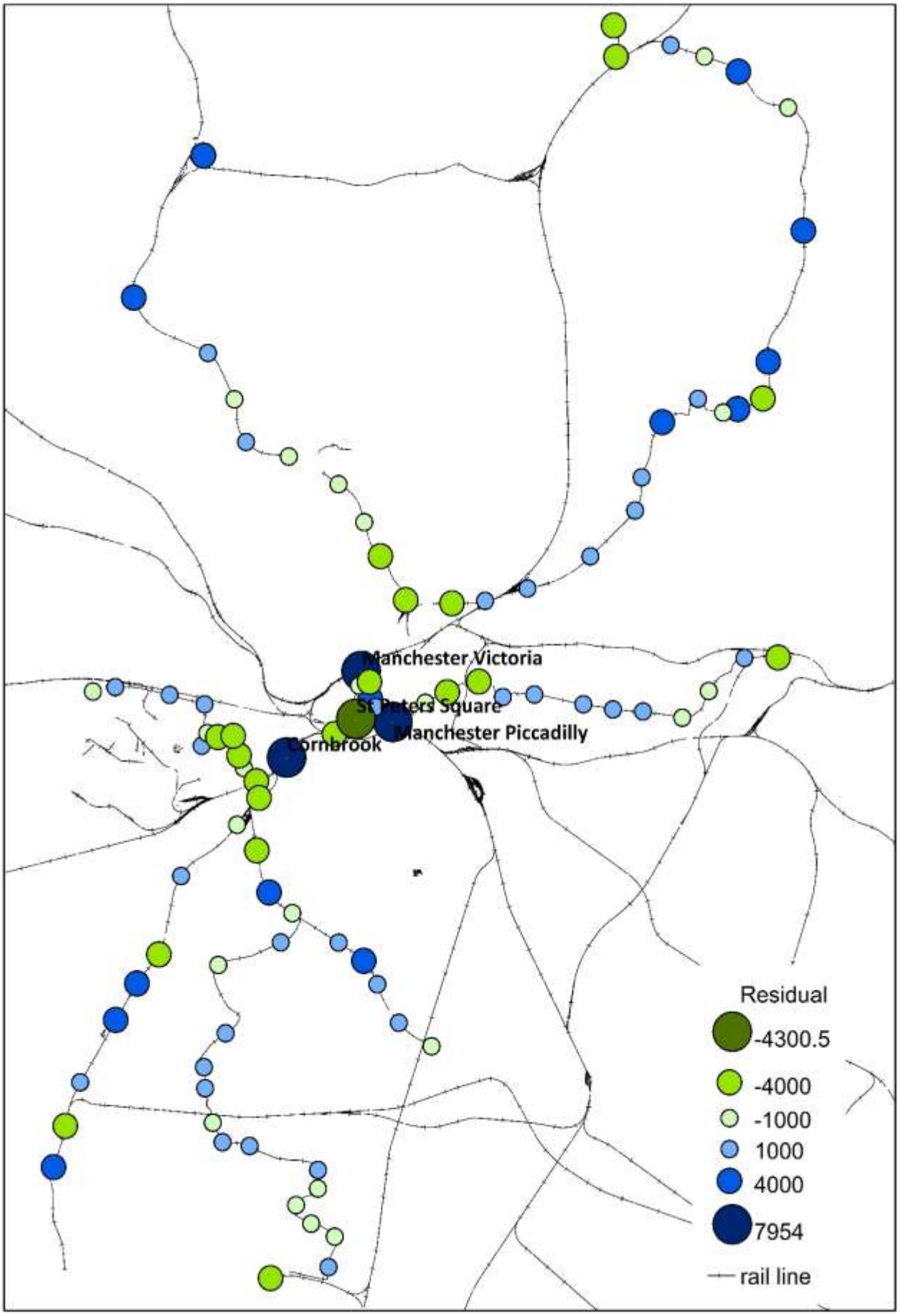

Figure 4.5 The residual based on Model 4.30

For further inclusion of the variables, Table 4.14 gives information of which variable might help to improve the model. Bicycle stands and distance to rail station still have wrong correlation sign, but car parking spaces are no longer wrong. Thus, the following steps will test the presence of terminus, distance to city centre and car parking spaces in the model.

Table 4.14 The correlation table for some variables after the exclusion of 4 stations

Explanatory variables	Boarding alighting	Distance to city centre	Terminus	Distance to the rail station	Car parking spaces	Bicycle stands
Boarding alighting	1.000	-0.201	0.243	-0.008	0.027	-0.182
Distance to city centre	-0.201	1.000	0.198	0.250	0.223	0.330
Terminus	0.243	0.198	1.000	-0.177	0.049	0.050
Distance to rail station	-0.008	0.250	-0.177	1.000	0.190	0.145
Car parking spaces	0.027	0.223	0.049	0.190	1.000	0.184
Bicycle stands	-0.182	0.330	0.050	0.145	0.184	1.000

Station characteristics is important factor affecting the station demand, interchange and terminus stations tend to have more passenger than others, the former can transfer passengers from same or other public transport modes, the latter has larger catchment areas because people are willing to go further to arrive the end of line (Gutiérrez et al., 2011). The assessment of the impact of the interchange stations has been done by transit feeder and the tram number on previous steps. The next variable to be tested was based on the terminus dummy variable, but the model did not improve, whereas it decreases the influence of other important factors. Thus, it will be ignored. This calibration is given in Table 4.15.

$$V_i = \alpha + \delta\, F_i \; + \beta \sum_{a}^{n} P_a\, W_a + \rho B_i + \text{ƥ} \sum_{a}^{n} L_a\, W_{2a} + \text{ɣ}\text{H}i \qquad (4.35)$$

Where: H_i is the dummy variable, if station *i* is the end of the line, it is 1, otherwise

Table 4.15 Summarised calibration result of Models 4.35 without 4 stations

Model		4.35
Parameter	Value	t statistic
Constant	-1274.308	298.157
Tram in an hour (δ)	338.663	36.854
Workplace population (β)	0.481	0.043
Bus or rail feeder (ρ)	945.106	471.370
Resident population within new catchment (ƥ)	0.186	0.097
Terminus (ɣ)	107.800	603.303
R^2 adj		0.837

The Pearson correlation coefficient Table 4.14 demonstrates that even though the correlation is weak, the distance from the station to the city centre might affect the passenger demand. The variable increased model fit from 0.839 to 0.848, but the significance of bus or rail feeder is no longer less than the critical value of 0.05. Hence, transit feeder was omitted and computed again. The R^2 adj value of model slightly declined, but all parameters are significant. Table 4.16 allows to the comparison.

$$V_i = \alpha + \delta\, F_i \; + \beta \sum_a^n P_a\, W_a + \rho B_i + \mathfrak{p} \sum_a^n L_a\, W_{2a} + \mathfrak{y} D_i \tag{4.36}$$

$$V_i = \alpha + \delta\, F_i \; + \beta \sum_a^n P_a\, W_a + \mathfrak{p} \sum_a^n L_a\, W_{2a} + \mathfrak{y} D_i \tag{4.37}$$

Where: D_i is the distance from station *i* to the city centre

Table 4.16 Summarised calibration result of Models 4.36 – 4.37 without four stations

Model			4.36			4.37
Parameter	Value	t statistic	sig.	Value	t statistic	sig.
Constant	-2243.559	-4.573	0.000	-2482.007	-5.185	0.000
Tram in an hour (δ)	384.950	9.916	0.000	401.779	10.523	0.000
Workplace population (β)	0.486	12.023	0.000	0.508	13.023	0.000
Bus or rail feeder (ρ)	689.945	1.803	0.075	n/a	n/a	n/a
Resident population within new catchment (ƥ)	0.192	2.089	0.040	0.200	2.151	0.034
Distance to city centre (ƞ)	0.077	2.426	0.017	0.096	3.162	0.002
R^2 adj			0.848			0.844

Also, although Pearson correlation of car parking space in Table 4.14 is very weak, the Spearman (nonparametric) correlation coefficient indicates the value of 0.25. Therefore, Model 4.38 was calibrated, and the results were summarised in Table 4.17. This provides an improvement in model fit from 0.844 to 0.8546.

$$V_i = \alpha + \delta\, F_i \; + \beta \sum_a^n P_a\, W_a + \mathfrak{p} \sum_a^n L_a\, W_{2a} + \mathfrak{y} D_i + \mathfrak{k} S_i \tag{4.38}$$

Where: S_i is the size of car parking space around tram station *i*

Table 4.17 Summarised calibration result of Models 4.38 without 4 stations

Model			4.38
Parameter	Value	t statistic	sig.
Constant	-2534.046	-5.450	0.000
Tram in an hour (δ)	397.320	10.695	0.000
Workplace population (β)	0.513	13.521	0.000
Resident population within new catchment (ϐ)	0.238	2.580	0.012
Distance to centre meter (ɳ)	0.081	2.693	0.000
Car parking spaces (ƙ)	3.904	2.653	0.010
R^2 adj			0.8546

Previous researches point out that using a natural logarithm for variables help to enhance the model. The log transformation makes highly skewed distributions less skewed thus, decreased variability among data might provide a better correlation coefficient. Also, a logarithmically transformed model is more interpretable because the coefficient obtained from calibration can be directly called as elasticity. Firstly, The Model 4.39 which is the transformed form of a dependent and all independent variables of the last best Model 4.38 were tested, but the model fit highly decreased see Table 4.18. Then, the partial application of natural logarithm for some or all the independent variables was also tested, and results given were summarised below Table 4.19 and Table 4.20.

$$lnV_i = \alpha + \delta lnF_i + \beta ln \sum_{a}^{n} P_a W_a + ϐ\ln \sum_{a}^{n} L_a W_{2a} + ɳ\ln D_i + ƙ\ln S_i \qquad (4.39)$$

Table 4.18 Summarised calibration result of Models 4.39 without 4 stations

Model		4.39
Parameter	Value	t statistic
Constant	3.982	2.823
Ln Tram in an hour (δ)	1.116	5.292
Ln Workplace population (β)	0.091	3.369
Ln Resident population within new catchment (ϐ)	0.038	1.142
Ln Distance to centre meter (ɳ)	0.058	0.457
Ln Car parking spaces (ƙ)	0.079	2.206
R^2 adj		0.442

As a result, while the previous work suggested to use of logarithmic form of model, there is no major improvement for trip end passenger demand model applied here. Only when the logarithmic transformation of car parking spaces made a slight contribution on the model development, see Model 4.40. Detailed calibration outcomes are displayed in Table 4.21. Workplace population is highly significant in our final model, following by tram in an hour. Moreover, the residual map (Figure 4.6) allows the interpretation of model validation for every station. The final model gave the highest error for Altrincham terminus station.

Table 4.19 The R^2 adj value of the 32 different models

boarding alighting				resident population	ln resident population
tram in an hour	workplace population	distance to centre	car parking	0.8548	0.8469
			ln car parking	0.8551	0.8482
		ln car parking	car parking	0.8524	0.8460
			ln car parking	0.8523	0.8467
	ln workplace population	distance to centre	car parking	0.5843	0.5844
			ln car parking	0.5858	0.5859
		ln distance to centre	car parking	0.5574	0.5585
			ln car parking	0.5624	0.5632
ln tram in an hour	workplace population	distance to centre	car parking	0.8161	0.8070
			ln car parking	0.8132	0.8053
		ln distance to centre	car parking	0.8082	0.8002
			ln car parking	0.8061	0.7991
	ln workplace population	distance to centre	car parking	0.5135	0.5146
			ln car parking	0.5106	0.5119
		ln distance to centre	car parking	0.4889	0.4913
			ln car parking	0.4911	0.4932

Table 4.20 The R^2 adj value of the 32 different models when the dependent value is

Ln boarding alighting				resident population	ln resident population
tram in an hour	workplace population	distance to centre	car parking	0.5542	0.5137
			ln car parking	0.5752	0.5387
		ln distance to the centre	car parking	0.5621	0.5251
			ln car parking	0.5752	0.5387
	ln workplace population	distance to centre	car parking	0.4526	0.4342
			ln car parking	0.4704	0.4534
		ln distance to centre	car parking	0.4433	0.4277
			ln car parking	0.4627	0.4474
ln tram in an hour	workplace population	distance to centre	car parking	0.5688	0.5292
			ln car parking	0.5782	0.5400
		ln distance to centre	car parking	0.5708	0.5340
			ln car parking	0.5787	0.5426
	ln workplace population	distance to centre	car parking	0.4617	0.4437
			ln car parking	0.4708	0.4538
		ln distance to centre	car parking	0.4441	0.4286
			ln car parking	0.4573	0.4419

$$V_i = \alpha + \delta F_i + \beta \sum_a^n P_a W_a + \beta \sum_a^n L_a W_{2a} + \eta D_i + \kappa lnS_i \quad (4.40)$$

Table 4.21 The calibration result of best fit Model 4.40

Model			4.40
Parameter	Value	t statistic	sig.
Constant	-2429.140	-5.258	0.000
Tram in an hour (δ)	391.603	10.582	0.000
Workplace population (β)	0.511	13.570	0.020
Resident population within new catchment (ϐ)	0.213	2.372	0.019
Distance to centre meter (ɳ)	0.073	2.396	0.008
Ln car parking spaces (ƙ)	166.590	2.712	0.000
R^2 adj			0.8551

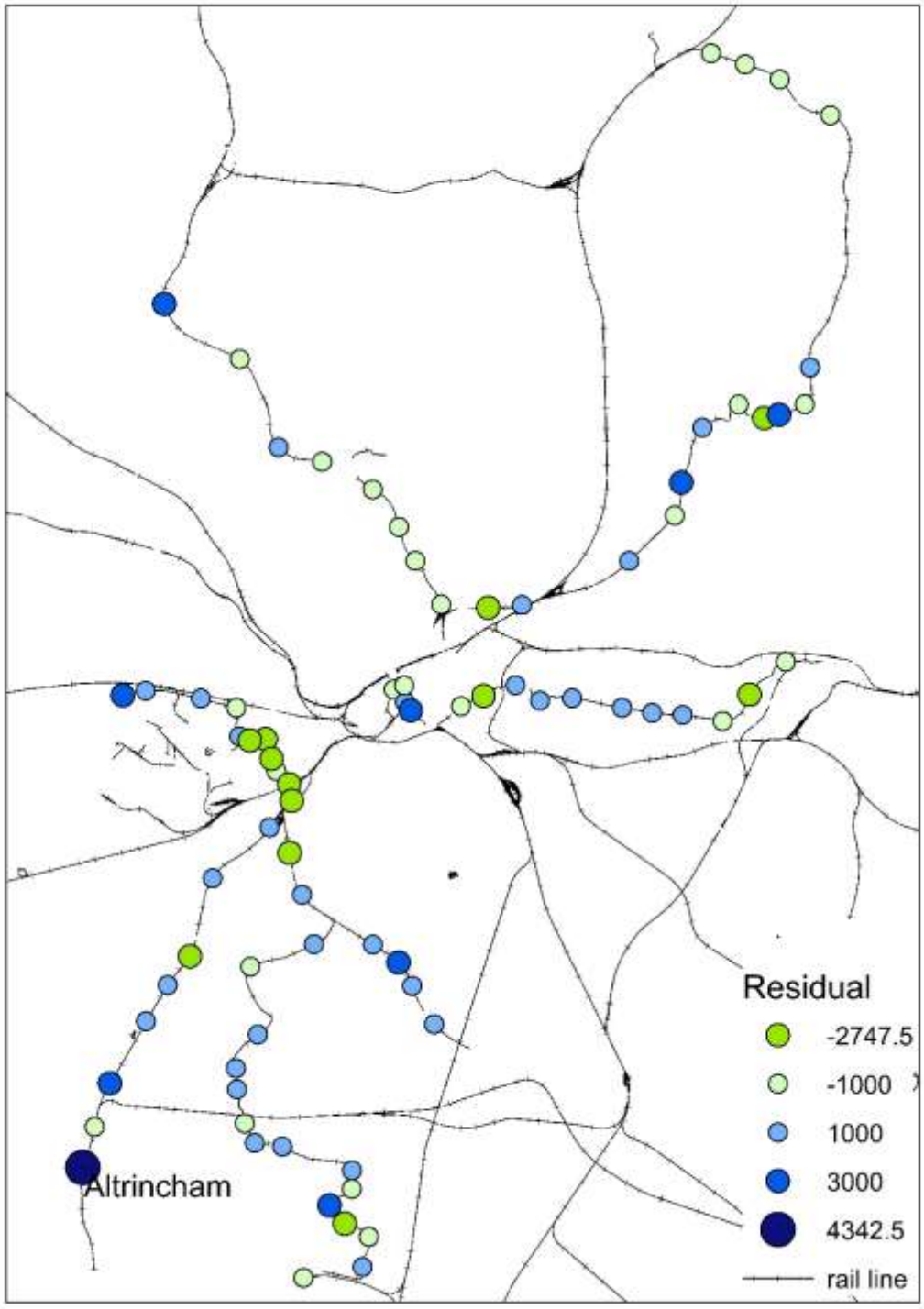

Figure 4.6 The residual based on model 4.40

4.3. Modelling results

Different catchment area definitions were tested to find the highest correlation value for each variable regarding station characteristic, socio-economic factor and service area feature. Then, they were used in the model calibration to find the best fit multiple regression model predicting the demand at Manchester Metrolink station. The final model contains the five variables which are resident population within catchment, workplace population within catchment, tram in an hour, distance to city centre and the size of car parking space at station. The resident population catchments were defined by the two stage time-based allocation, while the workplace population catchments were determined by two stage distance-based allocation method as described in Section 4.1.3.

As for eliminated variables, the number of bicycle stand and the distance to closest rail station were eliminated from the model due to the wrong correlation sign with independent value. The influence of car ownership was weak on passenger demand (see Table 4.7) whereas the non-car owning people within the buffer zone had a moderate correlation see (Table 4.8). However, there was a multicollinearity problem between non-car owning people and resident population. This forced to select either of them to incorporate into the model, so resident population was chosen. Furthermore, the number of bus stops within buffer zone provided almost an enough effect, but the variable has a very high correlation with the workplace population which has developed model significantly, so the bus stop value was removed in the process. Additionally, the feeder services were initially included, but the further process indicated that its exclusion allows variables to remain at significant level. The presence of the terminus was also tested, but it did not improve the model.

The mapping of interim model residuals has helped to examine the outlier stations which decrease the model fit. Thus, these stations were removed from model to provide the model with better estimation power. Also, linear and loglinear forms of all variables were compared, and the linear form provides superior results. As a result, the final model reached an adjusted R^2 of 0.855, hence, the model can estimate around 85% of the passenger demand at new stations (see Table 4.21). This model could be useful to assess stations viability and sustainability before being construction.

5. Conclusions

The project aim and the objectives listed in Section 1.3 will be compared to the modelling results mentioned above to provide conclusions about whether they have been fulfilled.

The primary aim of this study, which developing trip end model to forecast passenger demand at new light rail stations on the Manchester Metrolink was achieved in Section 4.2. This model can forecast around 85% of the passenger demand at new stations.

The first obtective of this study is to review literature to understand differences between trip end model and other demand estimation models. It was reached in Section 2.

The second objective is to test the influence of a wide range of factors. This analysis has been done by comparing the linear relationship between the dependent and explanatory factors. Also, the contribution of the variables was tested by putting incrementally them into the model. Then the combination of the variables giving the highest predictive power was chosen as the final model. This covered in Section 4.2.

The third objective is to examine the methodologies tried on previous and adapt them to the Manchester Metrolink. Thus, different catchment definitions obtained by this examination were tested and then the best one for each explanatory variable was chosen. This covered in Section 4.1.

The final objective is to use GIS throughout the study. the usage of ArcGIS has assisted to overcome the difficulties of the data manipulation and provided the satisfying visual presentation.

The aim and objectives established at the beginning of this study have been performed, and now there is a trip end model which can forecast the passenger demand at new light rail stations on the Manchester Metrolink.

- *Limitations*

This study has purposed to provide the model which forecasts passenger demand at light rail station in Greater Manchester. The main limitation of this study has been time, preventing the examination of more variables, and the evaluation of data used detailly. For example, the factor for bus and rail feeder was incorporated into the model as dummy variables. The inclusion of the number of bus line might have more significant influence. Additionally, the daily passenger count data obtained from Transport for Greater Manchester was used as dependent variable in model (see Section 3.2.1), however this dataset does not belong to a standardised day in a defined month. These seasonal differences might affect the accuracy of the model.

- *Recommendations for further research*

The trip end model developed here has been calibrated using the service area, station and socio-economic parameters. These all are aggregate data based on output zone. The usage of aggregate level data requires the deterministic catchment area definition for stations. However, this method ignores the competition between a new station and existing stations because does not reflect the station choice behaviour of passengers as being in real life. However, the probabilistic catchment definition applied by Young and Blainey (2018) for local rail stations in UK enables the realistic assessment of this behaviour. Therefore, the application of more realistic catchment definition can more improve the model accuracy.

6. References

All map data from Ordnance Survey (GB), Using: EDINA Digimap Ordnance Survey Service, <http://digimap.edina.ac.uk>

Association of Train Operating Companies (ATOC) (2013) *Passenger Demand Forecasting Handbook*, Version 5.1, London: ATOC.

Bowman, J. L., and Ben-Akiva, M. E. (2001) 'Activity-based disaggregate travel demand model system with activity schedules', *Transportation Research Part A*, 35(1), pp. 1–28.

Blainey, S.P (2010) 'Trip end models of local rail demand in England and Wales', *Journal of Transport Geography*, 18(1), pp. 153–165. doi: 10.1016/j.jtrangeo.2008.11.002.

Blainey, S.P and Mulley, C. (2013) 'Using Geographically Weighted Regression to forecast rail demand in the Sydney Region', Australian Transport Research Forum.

Blainey, S. P. (2009) 'Forecasting the Use of New Local Railway Stations and Services Using GIS', PhD Thesis. University of Southhampton.

Blainey, S. P. and Preston, J. M. (2013) 'A GIS-based appraisal framework for new local railway stations and services', *Transport Policy*, 25, pp. 41–51. doi: 10.1016/j.tranpol.2012.11.008.

Cardozo, O. D., García-Palomares, J. C. and Gutiérrez, J. (2012) 'Application of geographically weighted regression to the direct forecasting of transit ridership at station-level', *Applied Geography*. doi: 10.1016/j.apgeog.2012.01.005.

Choi, J. *et al.* (2012) 'An analysis of Metro ridership at the station-to-station level in Seoul', *Transportation*, 39(3), pp. 705–722. doi: 10.1007/s11116-011-9368-3.

Currie, G., Ahern, A. and Delbosc, A. (2011) 'Exploring the drivers of light rail ridership: An empirical route level analysis of selected Australian, North American and European systems', *Transportation*, 38(3), pp. 545–560. doi: 10.1007/s11116-010-9314-9.

Data.cdrc.ac.uk. (2018) *Welcome - CDRC Data*. [online] Available at: https://data.cdrc.ac.uk/ [Accessed 28 June. 2018].

Data.gov.uk. (2018) *Metrolink Stops and Rail Stations - data.gov.uk*. [online] Available at: https://data.gov.uk/dataset/8faea7ee-eb7d-43dd-b1d4-f01aac4c44d3/metrolink-stops-and-rail-stations [Accessed 2 July. 2018].

Data.gov.uk. (2018) *National Public Transport Access Nodes (NaPTAN) - data.gov.uk*.[online] Available at: https://data.gov.uk/dataset/ff93ffc1-6656-47d8-9155-85ea0b8f2251 /national-public-transport-access-nodes-naptan [Accessed 2 July. 2018].

De Andrade, G. T., Gonçalves, J. A. M. and da Silva Portugal, L. (2014) 'Analysis of Explanatory Variables of Rail Ridership: The Situation of Rio de Janeiro', *Procedia - Social and Behavioral Sciences*. doi: 10.1016/j.sbspro.2014.12.226.

Debrezion, G., Pels, E. and Rietveld, P. (2009) 'Modelling the joint access mode and railway station choice', *Transportation Research Part E: Logistics and Transportation Review*. doi: 10.1016/j.tre.2008.07.001.

Department for Transport, (2018) *Light Rail and Tram Statistics, England: 2017/18.* Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/ attachmentdata /file/720315/light-rail-and-tram-statistics-england-2018.pdf [Accessed 20 August. 2018].

El-Geneidy, A., Tétreault, P., Surprenant-Legault, J., 2010. Pedestrian access to transit: identifying redundancies and gaps using variable service area analysis. Paper Presented at the 89th, Transportation Research Board Annual Meeting, Washington, DC, USA.

Gmtu.gov.uk. (2018) *Highways Forecasting and Analytical Services* - Transport statistics 2011. [online] Available at: http://www.gmtu.gov.uk/reports/transport2016.htm [Accessed 20 Sep. 2018].

Gutiérrez, J., Cardozo, O. D. and García-Palomares, J. C. (2011) 'Transit ridership forecasting at station level: An approach based on distance-decay weighted regression', *Journal of Transport Geography*. doi: 10.1016/j.jtrangeo.2011.05.004.

Gutiérrez, J. and García-Palomares, J. C. (2008) 'Distance-measure impacts on the calculation of transport service areas using GIS', *Environment and Planning B: Planning and Design*, 35(3), pp. 480–503. doi: 10.1068/b33043.

GOV.UK. (2018) *Light rail and tram statistics.* [online] Available at: https://www.gov.uk/ government /collections/light-rail-and-tram-statistics [Accessed 20 August 2018].

Hildebrand, E., (2003) 'Dimensions in elderly travel behaviour: a simplified activitybased model using lifestyle clusters'. *Transportation,* 30 (3), pp. 285–306.

Huang, Z., Zhang, M. and Liu, X. (2017) 'Estimating light-rail transit peak-hour boarding based on accessibility at station and route levels in Wuhan, China', *Transportation Planning and Technology*, 40(5), pp. 624–639. doi: 10.1080/03081060.2017.1314497.

Kepaptsoglou, K., Stathopoulos, A. and Karlaftis, M. G. (2017) ‘Ridership estimation of a new LRT system: Direct demand model approach’, *Journal of Transport Geography*. Elsevier Ltd, 58, pp. 146–156. doi: 10.1016/j.jtrangeo.2016.12.004.

Kuby, M., Barranda, A. and Upchurch, C. (2004) ‘Factors influencing light-rail station boardings in the United States’, *Transportation Research Part A: Policy and Practice*. doi: 10.1016/j.tra.2003.10.006.

Lane, C., DiCarlantonio, M. and Usvyat, L. (2006) ‘Sketch Models to Forecast Commuter and Light Rail Ridership: Update to TCRP Report 16 ’, *Transportation Research Record*, 1986(1), pp. 198–210. doi: 10.3141/1986-26.

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

McNally, M. G. (2007) *The four step model. In D. A. Hensher & J. K. Button (Eds.), Handbook of transport modeling*. Oxford: Pergamon.

Miller, H. J. and Shaw, S.-L. (2001) *Geographic Information Systems for Transportation: Principles and Applications*. Oxford University Press, Oxford.

Miller, H. J. and Shaw, S.-L. (2015) ‘Geographic Information Systems for Transportation in the 21st Century’, *Geography Compass*. doi: 10.1111/gec3.12204.

Preston J. (1987) *The Evaluation of New Local Rail Stations in West Yorkshire*, School of Economic Studies (Institute for Transport Studies), University of Leeds, PhD thesis

Preston, J. (1991a) Demand Forecasting for New Local Rail Stations and Services, *Journal of Transport Economics and Policy* 25(2):183-202

Preston, J. (1991b) *Passenger Demand Forecasting for New Rail Services: Manual of Advice*, Institute for Transport Studies, University of Leeds, Working Paper 352

Preston, J. (2015) ‘Public transport demand’, in Nash, C. (ed.) *Handbook of research methods and applications in transport economics and policy*. Cheltenham, UK: Edward Elgar, pp.192-211.

Sun, L.-S. *et al.* (2016) ‘Estimation of transit ridership based on spatial analysis and precise land use data’, *Transportation Letters*. doi: 10.1179/1942787515Y.0000000017.

Transport for Greater Manchester. (2018a) *Metrolink Geographical Map*. [online] Available at: https://www.tfgm.com/public-transport/tram/geographical/network-map [Accessed 1 Sep. 2018].

Transport for Greater Manchester. (2018b) *The future of rapid transit*. [online] Available at: https://www.tfgm.com/future-travel/tram [Accessed 15 Sep. 2018].

Webarchive.nationalarchives.gov.uk. (2018) *Population-weighted centroids* - ONS. [online] Available at: http://webarchive.nationalarchives.gov.uk/20160110200251/http://www.ons.gov.uk/ons/guide-method/geography/products/census/spatial/centroids/index.html [Accessed 20 June. 2018].

Wyg, (2015), 'How far do people walk?' Presented at the PTRC Transport Practitioners' Meeting London, July 2015

Young, M. A (2017) 'Development of Integrated Demand and Station Choice Models for Local Railway Stations and Services', *European Transport Conference, 2017*

Young, M. A. and Blainey, S. P. (2018) 'Development of railway station choice models to improve the representation of station catchments in rail demand models', *Transportation Planning and Technology*. Taylor & Francis, 41(1), pp. 80–103. doi: 10.1080/03081060.2018.1403745.

Zhang, D. and Wang, X. C. (2014) 'Transit ridership estimation with network Kriging: A case study of Second Avenue Subway, NYC', *Journal of Transport Geography*. doi: 10.1016/j.jtrangeo.2014.08.021.

Zhao, J. *et al.* (2014) 'Analysis of Metro ridership at station level and station-to-station level in Nanjing: An approach based on direct demand models', *Transportation*, 41(1), pp. 133–155. doi: 10.1007/s11116-013-9492-3.