Australasian Transport Research Forum 2013 Proceedings
2-4 October,Brisbane Australia
Publication website: http://www.patrec.org/atrf.aspx

Using Geographically Weighted Regression to forecast rail demand in the Sydney Region

Dr Simon Blainey[1], Professor Corinne Mulley[2]

[1] Transportation Research Group, University of Southampton, UK

[2] Institute of Transport and Logistics Studies, Business School, The University of Sydney

Email for correspondence: corinne.mulley@sydney.edu.au

Abstract

This paper is centred on using Geographically Weighted Regression (GWR) to investigate the spatial variability in the factors which influence rail demand patterns in the Sydney region of New South Wales, and to explicitly incorporate these variations in the demand forecasting process. It explores whether allowing spatial parameter variability to be included in demand models enhances their explanatory power and improves their forecasting capability. The paper firstly reviews the different methodologies in use for estimating and forecasting demand, both in the Sydney region and elsewhere around the world. Against this background the methodology of GWR is explained, identifying the key differences in both inputs and outputs of a GWR based model. The paper then describes the development of 'conventional' global regression models of rail demand in the Sydney region, and reports on the recalibration of the most successful model using GWR. A large number of explanatory variables are tested in the models, including catchment population and employment, household size, income and age profile, car ownership levels, train frequency, bus interchange potential, bicycle storage provision and distance to the city centre. The results from the global and GWR models are then compared, with the significance of the spatial variation in the GWR models tested using an AICc-based criterion. This comparison informs a discussion of the extent to which modelling spatial parameter variation contributes to a better understanding of rail demand and its forecasting. The discussion concludes by identifying the relative merits of the developed GWR models as compared to other models for forecasting rail travel.

1. Introduction

This paper is centred on using Geographically Weighted Regression (GWR) to investigate the spatial variability in the factors which influence rail demand patterns in the Sydney region of New South Wales (NSW), and to explicitly incorporate these variations in the rail demand forecasting process. It explores whether allowing spatial parameter variability to be included in demand models enhances their explanatory power and improves their forecasting capability. GWR has previously been used in forecasting demand for new railway stations in the UK (Blainey, 2010) and in the transport context in NSW, Australia in explaining vehicle kilometres travelled at the household level, as a prelude to forecasting.

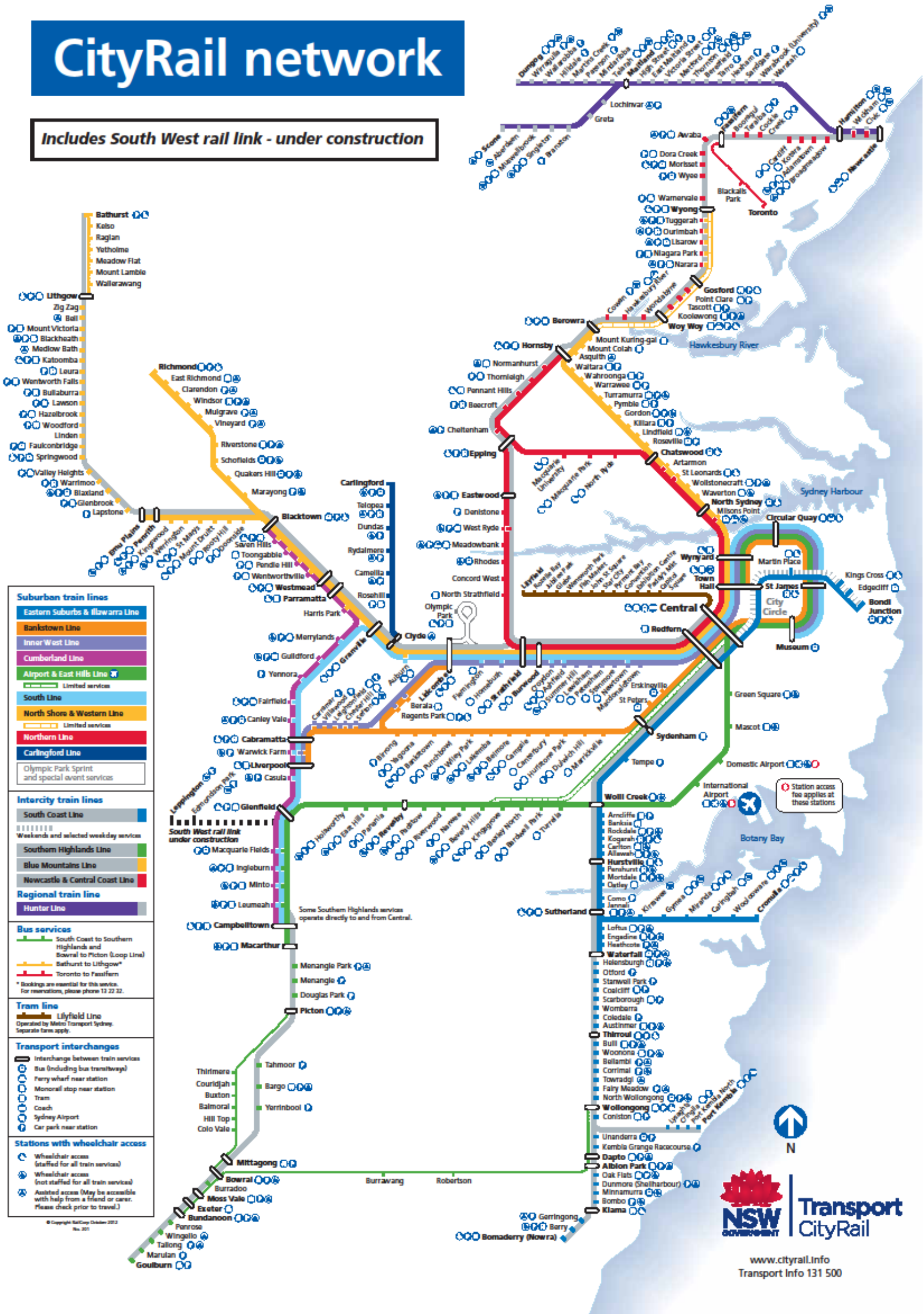
The paper is structured as follows. The next section outlines the background to the study through a short review of the different methodologies in use for estimating and forecasting rail demand, both for the Sydney region and elsewhere around the world. Against this background the GWR methodology is explained, identifying the key differences in both inputs and outputs of a GWR based model. The paper then describes the development of 'conventional' global regression models of rail demand in the Sydney region, before reporting on the recalibration of the most successful model using GWR. The patterns and implications of the spatial variation displayed by the parameters from the GWR model are discussed in the penultimate section. The final section concludes with a discussion of the relative merits of using GWR in this context

2. Background

2.1 The Sydney rail network

Sydney is the capital city of New South Wales state and is the largest city in Australia, as measured by population. In 2010, the population of the Sydney Greater Metropolitan area was around 5.56 million. The primary modes of public transport in Sydney are train and bus. The train system serving the metropolitan area (and beyond) is operated by Rail Corporation New South Wales (RailCorp) as two segments: the CityRail suburban service and the Countrylink inter-city services linking the major settlements of NSW and beyond to other States in Australia. CityRail's suburban operations extend to 2,242km of track reaching north to the Hunter and Central Coast, west to the Blue Mountains and Southern Highlands and to South Coast regions in the south and the train network within this area is shown in Figure 1.

Figure 1 The CityRail Network of the Sydney Metropolitan area
Source: www.cityrail.info/stations/network_map



The first link on this network between Sydney and Paramatta was opened in 1885. Electrification was implemented progressively from 1926 but was not fully completed on the network until 2002. The shape of the network was established by the late 1930s with minor (but important) changes coming in 1956 with the completion of the City Circle with Circular Quay station, the completion of the Eastern Suburbs line in 1979, and the extension of the Cumberland line to Blacktown in 1996. The Sydney Olympics 2000 prompted the building of the 5km Olympic Park Line, opened in 1999, and access to Sydney airport with a 7.3km airport line was created with a PPP project around the same time. The most recent network extension is the 12.5km line linking Epping and Chatswood stations which opened in 2009. Two further rail extensions are under construction at the time of writing, the southwest and northwest rail links.

2.2 Overview of existing rail modelling methods used in NSW

In NSW, the Sydney Strategic Transport Model (STM) is utilised by the Bureau of Transport Statistics (BTS) to provide and predict travel patterns in the Greater Metropolitan Area (GMA) of Sydney under alternative land use, transport network, service and policy scenarios. The STM has two main parts, the population model which segments the population into groups, based on socio-demographic variables which affect travel choices, and the travel model which is primarily a four step model encompassing trip generation or travel frequency from origins, trip distribution identifying the destination of trips from each origin (ODs), mode choice, and trip assignment or route chosen for travel between each OD pair.

The STM uses 2,690 travel zones as its basic geography. Demographic forecasts (of population and workforce), travel survey data, and digital representations of the transport network are combined with census demographics to construct current and future synthetic populations. The synthetic population is broken down into segments using variables known to be key predictors of travel behaviour, so that the STM models seven different travel purposes separately (work (commute from home to work and back), business, primary education, secondary education, tertiary education, shopping and other). The travel behaviour modelled in the STM is underpinned by the long-standing annual rolling Household Travel Survey (HTS) which gives the STM a distinctive advantage in its ability to use high quality data in modelling.

Rail is one of several modes explicitly included in the mode choice model which is applied to each travel purpose. This model is applied to different population segments, and is driven by socio-demographic variables as well as car availability and proximity to rail and competing public transport services, with service choices being based on travellers seeking to minimise generalised cost. The STM is an iterative model and the road (and bus) speeds from one iteration are fed back into the accessibility and generalised cost calculations for the next.

The BTS also has a Public Transport Project Model (PTPM) which is used in project evaluation to model more detailed aspects of individual projects. This model uses current origin destination matrices by mode (collected through public transport OD surveys) which are projected into the future based on STM growth projections. The PTPM has a more detailed mode choice model than the STM, explicitly modelling access mode choice, station choice and service crowding in the case of rail projects.

2.3 Modelling methods used elsewhere

Similar four-stage strategic transport models are used widely throughout the developed world to forecast future demand for both public and private transport, with examples being the PRISM model used in the West Midlands region of the UK (Mott Macdonald, 2010) and Transport Scotland’s TMfS model (MVA, 2011). However, such models may not be the most suitable tool for forecasting usage at individual railway stations, or in particular for predicting the likely level of passenger demand for new railway stations or lines. They tend to be based on a ‘conventional’ four stage form, with a logit-based mode choice model containing a utility function which may not be specified in a form which can adequately replicate the ‘step change' in rail service quality generated by the opening of a new station, particularly when rail has previously been a very minor mode. As a result, rail-specific elasticity-based simulation models are commonly used to forecast changes in demand at existing stations (with a comprehensive guide to such models provided by ATOC, 2011), and 'trip end' or 'sketch' models of rail demand have been successfully applied in both the USA and the UK (Lane et al., 2006; Preston, 1991) to forecast the demand for new railway stations or lines. In the latter context, the use of a technique called 'Geographically Weighted Regression' (GWR) during model calibration has been found to improve the fit and predictive power of such models (Blainey, 2010). This paper describes the application of such a modelling approach to the Sydney area, along with some modifications and enhancements to this modelling framework.

3. Methodology

The modelling of rail demand in this paper involves using a trip end modelling framework to explain the observed patronage for all stations in the Sydney GMA (as shown in Figure 1 above) based on the levels of a number of explanatory variables.. Trip end models are conventionally calibrated using multiple linear regression methods. Traditional multiple regression OLS modelling assumes that the relationship to be modelled is uniform over the study area, but when geographical data, such as distances to stations, are used as explanatory variables this can give rise to spatial effects. These effects may occur in two different forms: one is concerned with spatial dependency, or its weaker expression, spatial autocorrelation (they are not identical though they are often used interchangeably in the literature) and the other form is spatial heterogeneity, namely spatial non-stationarity (Anselin 1999). Spatial autocorrelation is a form of spatial interaction whilst spatial heterogeneity (spatial non-stationarity) refers to spatial structure (Anselin 1999). The two issues of spatial dependency and spatial non-stationarity have been the major challenges facing spatial data analysis (Fotheringham et al. 2002).

It was in order to tackle these challenges that GWR was developed by Fotheringham et al (2002). GWR is a relatively new technique for spatial data analysis and has been applied in the transport sector for the analysis of land value uplift (Du and Mulley (2007, 2012) and in trip end analysis of transport demand for rail (Blainey 2010). It has the ability to take account of spatial non-stationarity by accounting for coordinates in parameter estimates and spatial dependency by explicitly considering geographical location in calculation of the intercept values. The ,methodological enhancement provided by GWR is shown below in relation to a traditional cross-sectional regression model (as described by Fotheringham et al (2002)), with this model written as

$$Y_i = \beta_0 + \sum_k \beta_k \beta_{ik} + \varepsilon_i \qquad (1)$$

GWR involves expanding this model to a form which allows for local variations in the parameter values which take account of the coordinates of individual regression points. If the dependent variable has the coordinates (ui,vi), the model expressed in (1) above can be rewritten as the following GWR local model:

$$Y_i (u_i v_i) = \beta_0 \; (u_i v_i) + \sum_k \beta_k \; (u_i v_i) \beta_{ik} + \varepsilon_i \qquad (2)$$

The parameters are estimated at the location (u_i, v_i) using a weighted least squares method and a predicted value of y. Estimation is a trade-off between efficiency and bias in the estimators with a weighting process using spatial kernels which capture the data points to be regressed by moving the regression point across the region. The weights are chosen so that observations near (u_i, v_i), have more influence on the result than observations further away, with the weight being a function of the bandwidth of the distance from data point j to regression point i. While the results are sensitive to the choice of bandwidth, the GWR software allows adaptive spatial kernels to be used so that the bandwidth is narrow when data are dense but wider where data are sparse. The model provides a set of parameter estimates for each data point (or alternatively for a set of prespecified regression points), which then can be mapped, allowing visual inspection of spatial parameter variation. The GWR process provides a unique advantage over other spatial methods in that each observation is treated as a separate observation, as opposed to observations lying within a particular boundary being grouped (for example within a particular political boundary) as is required by for example, multi-level modelling. The Akaike Information Criterion (AIC) is used to evaluate the goodness of fit in GWR modelling as it explicitly accounts for the complexity of the model. A rule of thumb is if the AIC of two models differ by more than 3 then they are statistically significantly different with the lower AIC suggesting a better fit (Fotheringham et al (2002)). If adaptive kernels are used in the estimation process, the GWR software chooses bandwidths so as to minimise AIC(Fotheringham et al 2002).

This paper is the first in the transport domain to use the recently released GWR4 software (version 4.0.72) developed at the National Centre for Geocomputation (NUIM), Ireland and Ritsumeikan University, Japan, and this software was used to undertake all GWR calibrations described here. This version provides some important improvements over earlier releases of the software, and particularly relevant is that it now allows the flexibility to use a semi-parametric (partial linear) GWR model form incorporating both fixed and geographically varying explanatory variables. Fixed variables may be chosen before calibration because of an a priori view that such variables do not vary spatially or through an empirical testing procedure that allows some variables to be fixed and some to vary over space. The GtoF/FtoG routine within GWR4 allows two different approaches for determining whether a semi-parametric approach is preferred. The GtoF (geographically varying to fixed) begins with a full GWR model and then compares models to find the optimal combination of varying and fixed explanatory variables in much the same way as stepwise regression proceeds for a traditional multiple regression. The FtoG (fixed to geographically varying) routine is the reverse where the yardstick is the traditional regression or global model with all parameters being fixed, and optimality is sought by varying explanatory variables in turn and in combination. Both these routines use AIC as the basis for choosing the optimal model. The advantage of using a semiparametric approach is that including explanatory variables as fixed when they do not vary significantly over space can both improve the overall model fit and provide a model which is more conceptually satisfactory.

The GWR methodology requires a spatial location in terms of a Cartesian or Geographic coordinate. The BTS uses a Cartesian system based on The Geocentric Datum of Australia (GDA) and using the map grid of Australia Zone 56 – a Universal Transverse Mercator projection, using the GRS80 ellipsoid – and this location system was used for this paper.

4. Results

4.1 Global Model calibration

The first stage in this analysis was the calibration of the best global trip end rail demand model given the data available for the region centred on Sydney, in New South Wales. Initially, all 307 stations served by the CityRail Network were selected for inclusion in the model, and these are mapped in Figure 1. The dependent variable in the model is the number of passengers using each station per week, based on barrier count data undertaken by Transport for New South Wales (TfNSW). A range of additional data on factors which might help to explain these rail usage levels were supplied by TNSW, and these can be summarised briefly as follows:

- Daily train frequency at station, disaggregated into am peak, inter-peak, evening peak and off-peak.
- Average headway of buses serving the station during the morning peak.
- Number of commuter car parking spaces provided at each railway station.
- Distance in km from station to Sydney CBD (defined as distance to Town Hall station).
- Distance in km from station to nearest Metropolitan Strategic Centre.

- Interchange category as defined by TNSW (classifies stations by the extent to which interchange opportunities are available).
- Number of bicycle racks and bicycle lockers provided at each station.
- Total population within a 1200m buffer around the station (with catchments allowed to overlap between stations). Null population figures are given for a small number of stations.
- Total employment within a 1200m buffer around the station using overlapping catchments.
- Number of people within the SA1 census zone in which the station is located falling within four age groupings, specifically 0-19, 20-34, 35-64 and 65+ years.
- Total number of dwellings within the SA1 census zone in which the station is located
- Car ownership variable, approximated by the number of dwellings within the SA1 census zone in which the station is located with the following numbers of motor vehicles registered: 0, 1, 2, 3 and 4+.
- Average household size, approximated by the number of households with the following numbers of residents within the SA1 census zone in which the station is located: 1, 2-3, 4-5, and 6+. .
- Average income variable, approximated by the number of households with the following income levels within the SA1 census zone in which the station is located: 1-399, 400-999, 1000-1499 and 1500+AUD per week.
- Assorted raw census data covering the majority of the demographic variables described above, permitting reallocation to non-overlapping road network based catchments.

Ten of the stations (Kembla Grange Racecourse, Penrose, Wingello, Bell, Zig Zag, Mindaribba, Hilldale, Wallarobba, Wirragulla and Lochinvar) were found to have no trips recorded in the barrier counts, and these were therefore removed from the calibration dataset, as when using a double-log model form (which has previously been found to give superior results in such modelling (Blainey, 2010)) the dependent variable cannot take a value of zero. Similarly, no train frequency data were available for three stations (Domestic Airport, International Airport and Mascot) which are operated by a private company. Ideally the model would also include a measure of the average fare charged at each station, but no fare data was available for this study, and it is in any case difficult to estimate an average fare measure which is not strongly correlated with distance to a particular destination (such as central Sydney) and/or actual destination choice (which will be influenced by fare levels, and therefore introduce an element of circularity to the model).

An initial model was calibrated based on the data supplied by TNSW using the effective catchments inferred by this data. The results showed that the model had a very good fit with the observed data, explaining over 87% of the logged variation in the dataset. However, a number of parameters expected to be important determinants of rail demand such as catchment population and employment were insignificant. The model was recalibrated using both the stepwise and backward stepwise calibration methods, since the latter is sometimes a better method of ensuring the most important explanatory variables are included in the model. Both these methods gave largely the same results, generating a model with comparable fit (adjusted R^2 = 0.87), but unlike the initial model they included catchment population as a significant variable although catchment employment was still excluded. Nearly all the model parameters were of the expected sign, with rail demand expected to increase with an increase in train frequency, car parking provision, provision of cycle storage facilities (as two variables, number of cycle racks and number of cycle lockers), catchment population, and average catchment income, but to decline with increasing distances to Sydney and to metropolitan strategic centres, and with higher levels of car ownership. Rail demand is also predicted to decline as the proportion of children in the catchment population increases, perhaps because the ‘door to door’ convenience of the car is more important when travelling with young children. While the inclusion of two separate significant cycle storage variables in the model might appear unexpected, combining these into a single variable was found to reduce model fit.

The absence of a significant employment parameter could result from correlations between the explanatory variables giving rise to multicollinearity. Pearson’s correlation coefficients were therefore calculated for each pair of explanatory variables to investigate this possibility, and this process identified a large number of significant correlations between the variables, as is to be expected with such a large dataset. Many of the correlations are weak, and in many cases even where stronger correlations exist these do not appear to have had a major impact on the significance of parameter estimates in the model results, such as those between the distance to Sydney CBD and levels of catchment population and employment, or between the number of dwellings and average income. For other correlations, such as those between service frequency and population, or employment and distance to central Sydney, there is no obvious way in which the variables could be adjusted to control for the correlations. These strong correlations may indicate the presence of multicollinearity with the outcome of important variables losing significance but with the maintenance of a high R^2. However, the strongest significant correlation is between catchment population and catchment employment (Pearson's $\rho = 0.886$), which may explain why both catchment population and employment were not included in the models produced by stepwise calibration. An attempt was made to control for this correlation by replacing the catchment employment variable with a variable giving the number of jobs per head of resident population (which was much less strongly correlated with the population variable) but further analysis showed that this and other similar approaches did not provide a significant parameter, and therefore employment was not included in the preferred model..

Another modification tested was to combine the four age variables into a single average age variable, which again made little difference to model fit, but did highlight that the data on age distribution was missing or incomplete for some predefined catchments. A further modification was therefore made to the dataset with the predefined catchment data supplied by TfNSW being replaced by new catchments defined using ArcGIS. All SA1 census zones within NSW were allocated to their nearest railway station by road distance. A catchment boundary of 1200m was defined, to give consistency with the predefined catchments, and all zones which were further than 1200m from a station were then removed from the dataset. The remaining zones were then aggregated by nearest station to give total catchment population and dwelling counts, and the mean age, car ownership level, household size and income for each catchment. The trip end demand model was then recalibrated with these new values replacing those in the calibration dataset taken from the predefined catchments, and stepwise calibration again used to select model variables. No data on employment levels within SA1 zones were available, and it was therefore necessary to use the employment figures from the predefined catchments. This model gave a small improvement in fit over previous models (adjusted $R^2 = 0.889$), suggesting that the non-overlapping catchments produced a more accurate demand model. However, the only catchment-related variables included were dwelling count and vehicle ownership, and not population. Investigation of the correlations between the new catchment-related variables showed that by far the strongest significant correlation was between population and dwelling count. The final and best model was therefore achieved by by using stepwise calibration with the dwelling count variable excluded from the variable list.

This final preferred model is defined by:

$$T_i = \alpha F_i^{\beta} B_i^{\gamma} C_i^{\rho} Ds_i^{\tau} Dm_i^{\zeta} Br_i^{\eta} Bl_i^{\kappa} P_{1.2i}^{\lambda} V_{1.2i}^{\chi} \quad (3)$$

Where:
T_i is the number of passengers per week at station *i*
F_i is the daily train frequency at station *i*
B_i is the average bus service headway at station *i*
C_i is the number of car parking spaces provided at station *i*
Ds_i is the distance in km from station *i* to the Sydney CBD
Dm_i is the distance in km from station *i* to the nearest Metropolitan Strategic Centre
Br_i is the number of bicycle racks provided at station *i*
Bl_i is the number of bicycle lockers provided at station *i*
$P_{1.2i}$ is the total resident population within all SA1 census zones within 1.2 km of station *i* for which station *i* is the closest station

$V_{1.2i}$ is the average number of vehicles registered at dwellings within all SA1 census zones within 1.2 km of station *i* for which station *i* is the closest station

Descriptive statistics for the model variables are given in Table 1 and the results from the global calibration of this model are summarised in Table 2.

Table 1: Summary of descriptive statistics for variables used in modelling

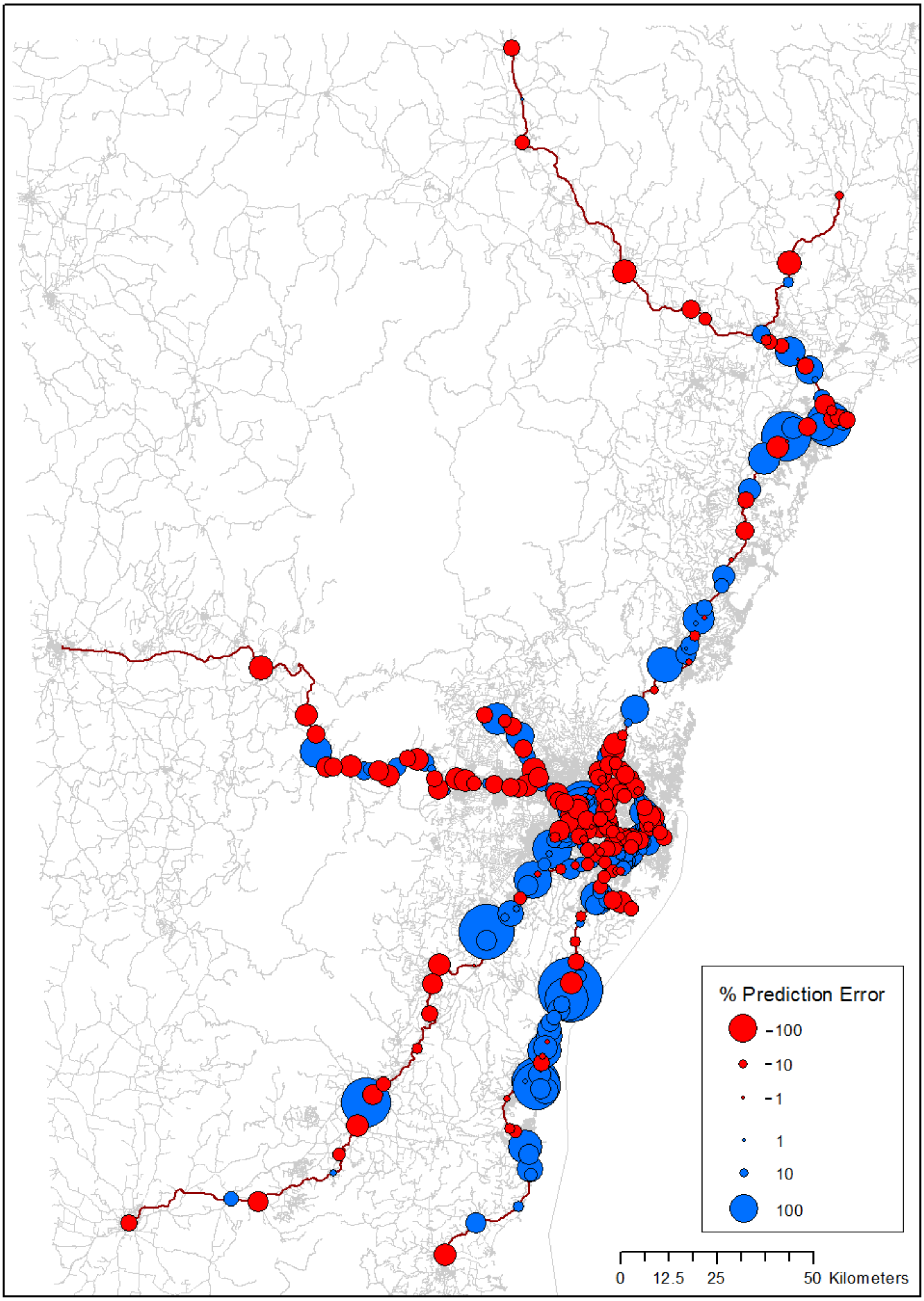
Variable	Unit of measurement	Mean	Standard deviation
Train frequency	Trains per day	156.29	192.35
Bus headway	Peak minutes between services	32.36	49.69
Car parking	Spaces provided	118.6	228.4
Distance to Sydney CBD	Km	62.04	58.98
Distance to nearest Metropolitan Strategic Centre	Km	16.07	2.65
Bike racks	Spaces provided	10.4	16.85
Bike lockers	Spaces provided	3.87	7.76
Catchment population	Number of residents	3551.28	3478.09
Average household vehicle ownership	Number of vehicles	1.42	0.31

Table 2: Summarised Results From Calibration Of the preferred global model

Variable	Value	T stat
α (intercept)	3.371	5.855
β (train frequency)	1.084	13.329
γ (bus headway)	0.053	2.203
ρ (car parking)	0.108	5.017
τ (distance to Sydney CBD)	-0.266	-4.431
ζ (Distance to nearest Metropolitan Strategic Centre)	-0.064	-3.827
η (bike racks)	0.172	4.817
κ (bike lockers)	0.160	3.536
λ (population)	0.136	7.684
χ (vehicles)	-0.704	-3.490
R_{adj}^2	0.889	
AD	0.068	

The prediction errors from the preferred model are mapped in Figure 2, with the red circles (and their sizes) showing stations where the model underpredicted the number of trips, and the blue circles (and their sizes) stations where the model overpredicted the number of trips. This shows some apparent spatial patterns in the model prediction errors, with clusters of stations where demand is either underpredicted or overpredicted. This suggests that spatial autocorrelation may exist in the dataset, and this justifies testing a spatial regression methodology in the form of GWR, as reported in the next section.

Figure 2: Prediction Errors From Model



4.2 Local Model results

The next step is to recalibrate the preferred global model using GWR. The GWR4 software was employed with adaptive spatial kernels determined using a bi-square function, as defined by equation (4) (Nakaya et al., 2012). The use of an adaptive kernel means that the same number of observations are used in the estimation of the model at each regression point. The golden section search method (a technique which successively narrows the range of values within which the optimum is known to exist) was used to determine optimal bandwidths based on small sample bias corrected AIC minimisation. The software was set to test the geographical variability of local coefficients, in order to determine which parameters exhibited significant spatial variations in their values. The software provides both global regression results to allow a consistency check with previous results (the results were identical to those produced by SPSS and reported above) and results from the specified GWR calibration, and the latter results from this calibration are summarised in Table 3.

$$w_{ij} = \begin{cases} \left(1 - d_{ij}^2/\theta_{i(k)}\right)^2 & d_{ij} < \theta \\ 0 & d_{ij} > \theta \end{cases} \qquad (4)$$

Where:

w_{ij} is the weight value of the observation at location *j* for estimating the coefficient at location *i*

d_{ij} is the Euclidean distance between *i* and *j*

$\theta_{i(k)}$ is an adaptive bandwidth size defined as the *k*th nearest neighbour distance

Table 3: Summarised Results From Local GWR Calibration Of the preferred model

Variable	Mean	Standard Deviation	Difference of Criterion
α (intercept)	3.049	1.898	-111.495
β (train frequency)	1.068	0.270	-50.357
γ (bus headway)	0.050	0.078	-4.766
ρ (car parking)	0.092	0.129	-32.291
τ (distance to Sydney CBDe)	-0.132	0.229	-4.348
ζ(distance to nearest Metropolitan Strategic Centre)	-0.060	0.032	2.745
η (bike racks)	0.157	0.051	4.949
κ (bike lockers)	0.141	0.071	7.066
λ (population)	0.150	0.078	-9.671
χ (vehicles)	-0.927	0.707	-5.800
R_{adj}^2	0.925		
F stat	3.287		

This table shows the mean value for each of the model coefficients, along with the standard deviations of these coefficients. Two measures of model fit are provided, allowing the GWR model to be compared to the previous global model. The adjusted R^2 value shows a clear improvement in model fit with the GWR model, and the significant F statistic (which tests the results of an ANOVA to establish whether the GWR model gives an improvement in fit over the global regression model) backs up this conclusion. The table also gives the ‘Difference of Criterion’ value for each variable, which is the result of a test of spatial variability in that variable’s coefficient (based on an AIC criterion). If this value is positive, it indicates that there is no significant spatial variability in the associated coefficient, and that the variable would be better represented as a global or spatially fixed term in the model. Table 3 suggests that there is no significant spatial variation in the effects of the two bike storage variables or of the distance to the nearest Metropolitan Strategic Centre on rail demand, and a partial GWR model may therefore give better results. Such models have not previously been tested in the rail demand context, as no suitable software for calibration was available. However, this problem has now been overcome with the availability of GWR4, and therefore a partial model was calibrated with the distance to nearest Metropolitan Strategic Centre, bike racks and bike lockers coefficients held constant over space, and the other coefficients allowed to vary. The results from this calibration are summarised in Table .

Table 4: Summarised Results From Partial GWR Calibration of the preferred Model

Variable		Mean	Standard Deviation	Difference of - Criterion
Local	α (intercept)	3.444	2.228	-70.245
	β (train frequency)	1.009	0.299	-23.994
	γ (bus headway)	0.046	0.082	-7.276
	ρ (car parking)	0.101	0.137	-42.819
	τ (distance to Sydney CBD)	-0.151	0.262	-3.959
	λ (population)	0.150	0.088	-8.955
	χ (vehicles)	-0.909	0.646	-5.268
		Value	**t stat**	
Global	ζ (distance to nearest Metropolitan Strategic Centre)	-0.073	-4.221	
	η (bike racks)	0.163	5.150	
	κ (bike lockers)	0.163	3.976	
R_{adj}^2		0.927		
F stat		4.108		

There are a number of points to note about these results. Firstly, the model fit shows a further slight improvement over the full GWR calibration, and the F statistic indicating an improvement in fit over the global model is more strongly significant. The three global parameters are all still significant, and their values are slightly different to those given by the global calibration. This is an important point, as a previous option tested for using GWR models to forecast demand (in the absence of partial GWR models) was to use the global parameter value for variables which exhibited no significant spatial variation, as a proxy for a partial GWR model calibration. The differing parameter estimates obtained from the partial GWR calibration show that this previous option is not directly substitutable for the use of a partial GWR model. The mean values of most of the spatially-varying parameters are also slightly different in the partial calibration from those obtained from the full GWR calibration, although all still exhibit significant spatial variations.

A further option tested was to allow the GWR4 software to select which variables in the GWR model should in fact be fixed over space. As described above (Section 3), the software does this in the GtoF routine by conducting a series of model comparison tests between the full GWR model and an alternative model in which one of the varying parameters has been changed to a fixed parameter. If the best of these alternative models has a better fit than the original then the relevant parameter is fixed, and further comparisons are then carried out for each of the remaining varying terms, until no improvement in model fit is gained by changing any further terms from varying to fixed. This procedure only altered the two bike storage parameters from local to fixed, leaving the MSA distance parameter to vary over space. The results of this model calibration are summarised in Table 1.

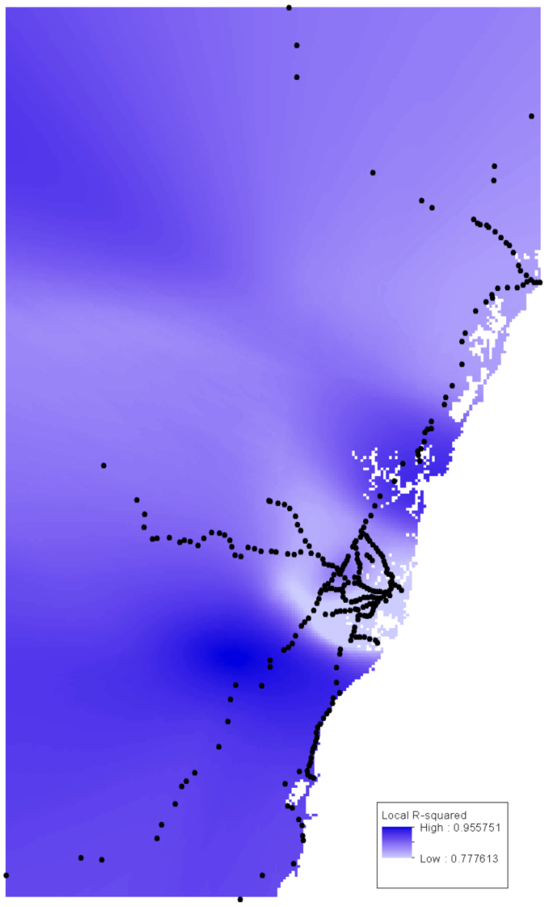
Table 5 shows that while this calibration gives the same adjusted R^2 value as the partial calibration with manual selection of global and local variables, the latter calibration gives a greater improvement over the global model as measured by the F statistic. The main difference between the two calibrations is that the distance to Metropolitan Strategic Centre variable was selected as being ‘local’ by the GtoF routine even though the difference criterion value indicates that it should be a global variable. The partial calibration with manual selection of global and local variables was therefore taken forward as the preferred option.

Table 1: Summarised Results from Partial GWR Calibration of the Preferred Model With Local To Global Variable Selection in GWR 4

Variable		Mean	Standard Deviation	Diff-Criterion
Local	α (intercept)	3.359	2.250	-108.568
	β (train frequency)	1.013	0.308	-7.305
	γ (bus headway)	0.044	0.082	-5.674
	ρ (car parking)	0.100	0.134	-39.936
	τ (distance to Sydney CBD)	-0.159	0.272	-4.759
	λ (population)	-0.060	0.031	-10.004
	ζ (distance to nearest Metropolitan Strategic Centre)	0.154	0.094	4.128
	χ (vehicles)	-0.983	0.788	-8.316
		Value	**t stat**	
Global	η (bike racks)	0.174	4.142	
	κ (bike lockers)	0.163	5.219	
R_{adj}^{2}		0.927		
F stat		3.724		

The next stage in the investigative process was to map the local model fit and parameter estimates from this calibration of preferred model. The local R^2 values produced by the model are mapped in Figure 3, which shows that while in general model fit is reasonably good throughout the study area (with a minimum value of 0.778), it is best on the northern and southern outskirts of Sydney, and is relatively poor around the south of the city centre. The reasons for this are not immediately clear, although it is possible that it results from a greater degree of variability in station usage levels in this area.

Figure 3: Local model fit from partial GWR calibration of the preferred model



The following figures show the spatial variation in the various model parameters which were set as being 'local' in the model calibration. The local parameter estimates are contained in raster grids (having been calculated by the software for all 1 km by 1 km grid cells across the study region), and are displayed with the t statistics for the relevant variables as calculated at each of the observation points. These are shown as circles on the maps, which are black where the parameter is statistically significant and white where it is not. When interpreting the maps it is important to note that less confidence can be placed in parameter observations which are further away from the railway stations where the observed data were collected.

Figure 4 shows the spatial variation in the train frequency and bus headway parameters and Figure 5 presents this for the car parking, distance to Sydney CBD, population and vehicle variables (as defined above). Train frequency appears to have positive effect on rail demand everywhere, and this effect is particularly strong in the area to the north of Sydney around the Hawkesbury River estuary, but makes relatively little difference to the level of rail demand around the periphery of the study area. The relationship between bus headway and rail demand appears rather more complex. In the area to the south of Sydney rail demand appears to reduce as bus headways increase, indicating that bus and rail services may compete for patronage in this area. In contrast, in the north-western suburbs of Sydney increased bus headways are positively correlated with rail demand, suggesting that bus services may act as feeders for the railway stations in this area. The impact of bus headway is less marked throughout the rest of the study area, and the parameter is not significant in these areas.

Figure 4: Spatial variation in train frequency and bus headway parameters

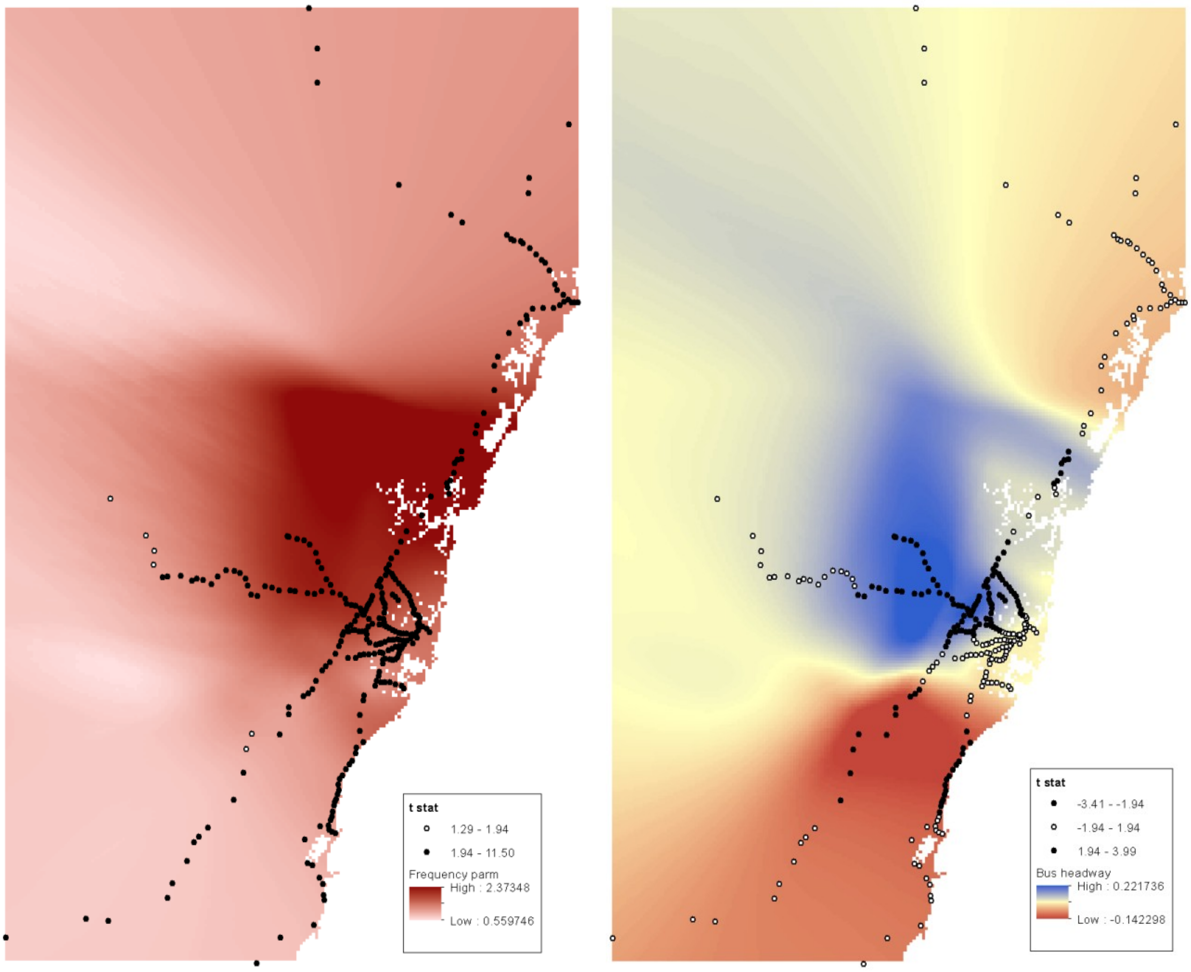


Figure 5: Spatial variation in car parking, Sydney distance, population and vehicle parameters

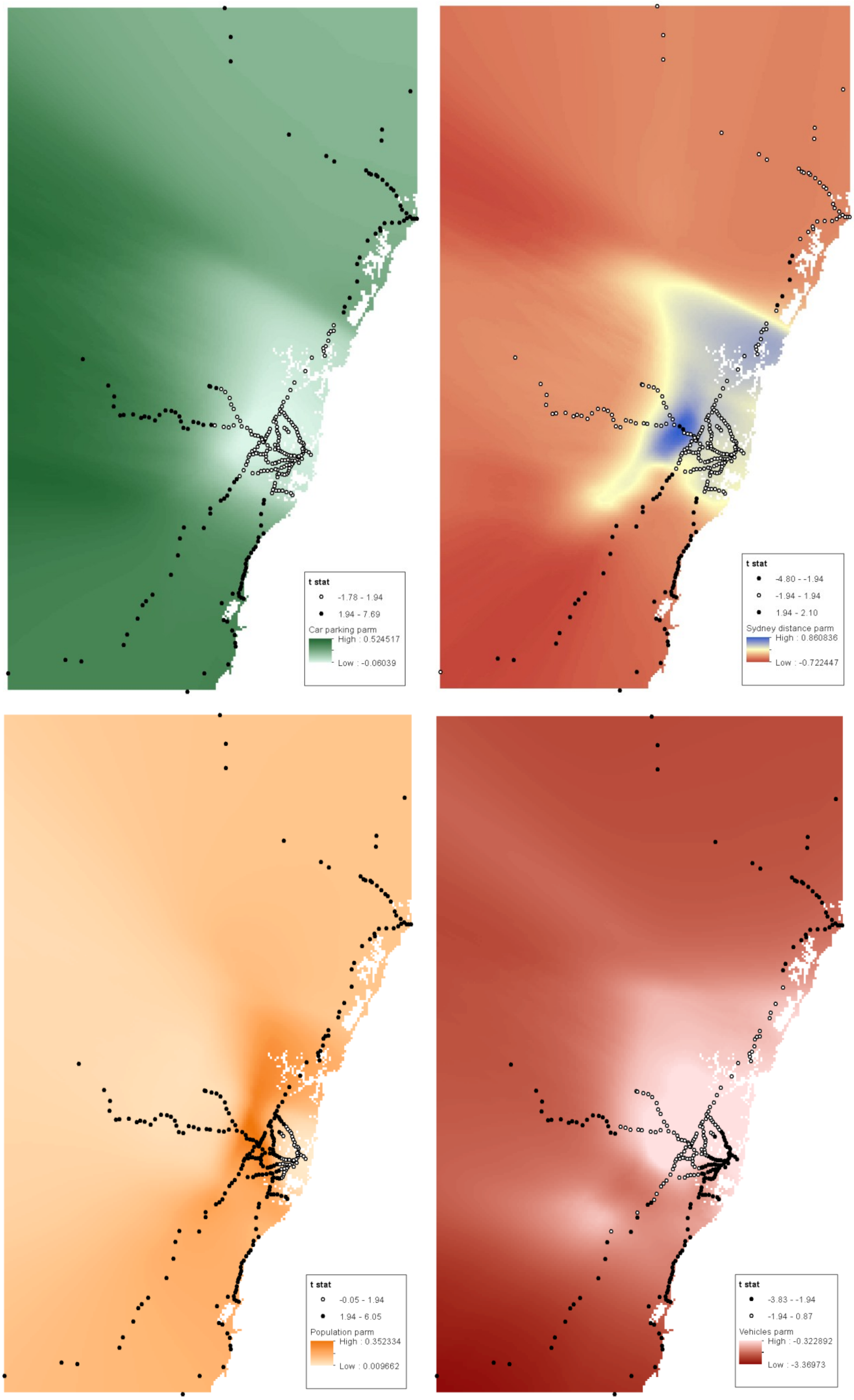


Figure 5 shows that the availability of car parking is in general positively correlated with rail demand, and that in general this relationship grows in strength with distance from Sydney. This may indicate that people making use of stations further away from Sydney CBD are more reliant on the car for accessing railway services, perhaps as a result of lower population densities and a lack of feeder bus services. At first sight, the variation in the distance to Sydney CBD parameter appears extremely complex. However, this parameter is in the main only significant in the southern part of the area, where it appears to indicate that increasing distance from Sydney leads to a reduction in rail trip making, as would intuitively be expected. The population parameter is significant in the majority of the study area, with the exception of Sydney CBD and the area immediately to its north (perhaps unsurprisingly, as population is unlikely to be the key determinant of demand in the city centre). To the west of the study area, Figure 5 shows a region where catchment population has a particularly strong positive impact on rail demand, which indicates that these are areas where rail commuting is particularly prevalent and the catchment population is more directly linked to trip rates than in other areas. Finally, the car ownership parameter is always negatively correlated with rail demand (increasing car ownership is linked to lower rail usage levels), but its significance is somewhat patchy. It appears to be strongly significant in the city centre, even though the implied elasticities here are quite small, but is also strongly significant around the periphery of the study area, where it also has a proportionally much larger impact on rail demand levels.

5) Discussion and conclusions

The analysis reported in this paper has shown that the use of GWR to calibrate rail demand models for New South Wales can give a clear improvement in model fit over equivalent ‘conventional’ multiple regression models. It has also demonstrated, for the first time in this field, that the use of partial GWR models gives superior results to models in which all parameters are allowed to vary over space, which has clear implications for this type of modelling more generally. The GWR models have also allowed a number of apparent spatial variations in the impact of different explanatory variables on rail demand in the Sydney area to be identified. Some suggestions for these variations have been provided in Section 4 but the results merit further investigation, in particular to identify whether the observed spatial variations result from the influence of additional variables which are not currently included in the demand modelling framework. This is important because otherwise there is a risk of rail demand patterns being incorrectly linked to particular causal factors with consequences for the accuracy of any future demand forecasts made using these models. It should also be noted that the model intercept parameter is still large and significant, indicating that ‘missing’ variables account for a non-trivial proportion of the variation in the observed data, and again attempts should be made to identify what these missing variables are through mapping to identify any significant patterns. These modifications notwithstanding, the GWR model could still be used in its current form to forecast the likely demand levels at new railway stations within the case study area, with the spatially varying parameter estimates providing place-specific patronage estimates which would take into account particular local circumstances. The next stage in this research will involve following both these steps, with the key aim being to use the partial GWR rail demand models to forecast passenger usage of the new Sydney North West Rail Link.

References

Anselin, L (1999) The future of spatial analysis in the social sciences, *Geographic Information Sciences,* 5, 67-76

Association of Train Operating Companies (2009) Passenger Demand Forecasting Handbook v5. London: ATOC.

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

Du H and Mulley C (2007), Transport accessibility and land value: a case study of Tyne and Wear, *RICS research paper* (Volume 7, Number 3), London, United Kingdom

Du H and Mulley C (2012) Understanding spatial variations in the impact accessibility on land value using geographically weighted regression, *Journal of Transport and Land Use,*.5(2), 46-59

Fotheringham AS, Brunsdon C, Charlton ME (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: Wiley.

Lane, C., DiCarlantonio, M. U., Usvyat, L., (2006). Sketch models to forecast commuter and light rail ridership: update to TCRP report 16. *Transportation Research Record* 1986, 198-210.

Mott Macdonald (2010) Prism: An Introductory Guide. Birmingham: Mott Macdonald

Mulley C and Tanner M (2009) The Vehicle Kilometres Travelled (VKT) by Private Car: A Spatial Analysis Using Geographically Weighted Regression, *32nd Australasian Transport Research Forum ATRF 2009*, Auckland, New Zealand,

MVA Consultancy (2011) *Land-Use and Transport Integration in Scotland: TMfS 07 Developer's Guide*. Glasgow: Transport Scotland.

NSW (2012) *Sydney's Rail Future, Modernising Sydney's Trains*

Preston, J. M. (1991) Demand forecasting for new local rail stations and services. *Journal of Transport Economics And Policy* 25(2), 183-202.