A GEOGRAPHICALLY WEIGHTED REGRESSION BASED ANALYSIS OF RAIL COMMUTING AROUND CARDIFF, SOUTH WALES

Simon Blainey, University of Southampton, S.P.Blainey@soton.ac.uk

John Preston, University of Southampton

ABSTRACT

Journey to work data from the 2001 census at ward level is used together with data on a range of socio-demographic variables, rail service levels and bus stop locations to develop and calibrate conventional regression models of the propensity to travel to work by rail in the area around Cardiff, South Wales. The best-performing models are then recalibrated using Geographically Weighted Regression (GWR), allowing local variations in the effect of the independent variables on rail demand to be mapped using GIS, investigated and incorporated in the modelling process. Flow level models of rail demand within the case study are then calibrated, and different methods for incorporating spatial parameter variations in these models are tested. Some conclusions are drawn about possible reasons for the spatial variations in rail use that have been identified, and the implications for demand forecasting for new stations and lines are discussed.

Keywords: Railway station, Geographically Weighted Regression, demand model

1) BACKGROUND

Rail use will often vary widely between different areas of a city region which are demographically very similar and which have comparable rail service levels. This variation is difficult to account for in conventional aggregate rail demand models, as it may result from differing local ‘cultures’ of rail use, which cannot be explained using easily quantifiable independent variables. Such variations may therefore have a negative impact on the goodness of fit of these models.

UK passenger rail use is currently at record levels and, despite the current recession, concerns over congestion and the environment mean that many new railway stations are proposed. For example, a recent report by ATOC (2009) outlined the potential for a large number of new lines serving areas currently isolated from the rail network. However, the construction of any such new lines or stations is dependent on the production of a positive business case, which will in turn require accurate forecasting of passenger numbers at the new stations. The main source of information on rail demand modelling in the UK is the Passenger Demand Forecasting Handbook (PDFH) (ATOC, 2005), produced by the Association of Train Operating Companies (ATOC). The methods recommended in the PDFH for demand forecasting at local stations and suburban stations are rather outdated, being largely based on Preston (1991) and related work, but a new integrated methodology for investigating the potential for constructing new local railway stations within a given area and assessing their likely performance has recently been developed (Blainey & Preston, 2009). While this methodology performs well at identifying station sites and forecasting total demand from new stations, the procedures for forecasting trip destinations are still in need of some refinement. Part of the reason for the problems encountered in this area is likely to be that the loglinear regression models used for forecasting flow level demand are unable to account for the spatial variations described above in the propensity to travel by rail.

However, modelling techniques are available which can explicitly incorporate spatial variations in the effect of independent variables on rail demand in the model form, and therefore improve the explanatory power of the model. One such technique which seems to offer particular promise for enhancing rail demand models is Geographically Weighted Regression (GWR), described in full by Fotheringham et al. (2002). GWR has previously been applied in several areas of transport analysis (see, for example, Du and Mulley (2006), and Clark (2007)), and in this particular field has been used to enhance national trip end models of rail demand at local stations in England and Wales, which can forecast the total number of trips made from particular stations to a high degree of accuracy (Blainey, 2010). However, GWR has not so far been used to investigate spatial variations in rail demand at the city region scale, or to model demand at the flow level, and this paper aims to address both these issues. It describes the development of regression models of the propensity to travel to work by rail in the area around Cardiff and the South Wales valleys, and of flow level rail demand from the local stations within this area.

2) WARD LEVEL MODELS

2.1. Data sources and case study area

The first stage of this work involved the development of models of the level of rail commuting to work. Attention focused on travel to work because an extensive dataset on commute mode choice was available from the 2001 Census, allowing analysis of spatial variations in rail use at a more detailed level than would be possible using ticket sales data (the usual basis for rail demand models in the UK). The chosen case study area comprises 6 unitary authorities in South-East Wales. The main reason for choosing this area was because it is one of very few large urban areas in the UK where no travelcard ticketing scheme operates. Such schemes have a detrimental impact on the accuracy with which ticket sales data can reflect actual travel patterns. Such data had been essential for previous work on local rail demand modelling (Blainey & Preston, 2009) and would be required for the later stages of this work (see Section 4), so this case study was an obvious choice on the grounds of both model comparability and reliability.

The first set of models developed aimed to forecast the number of rail travellers to work from each ward as recorded in the 2001 census. This data was downloaded for all wards in Cardiff, Caerphilly, Rhondda Cynon Taff, Merthyr Tydfil, Blaenau Gwent, and the Vale of Glamorgan from the CASWEB online interface for census data, along with data on a range of socio-demographic variables which might influence the level of rail use. This gave a total of 164 wards, and the number of rail commuters in each ward is mapped in Figure 1 along with the rail network in the area. This shows that, unsurprisingly, the highest levels of rail commuting tend to be found in wards containing railway stations, although there are some local variations in this pattern.

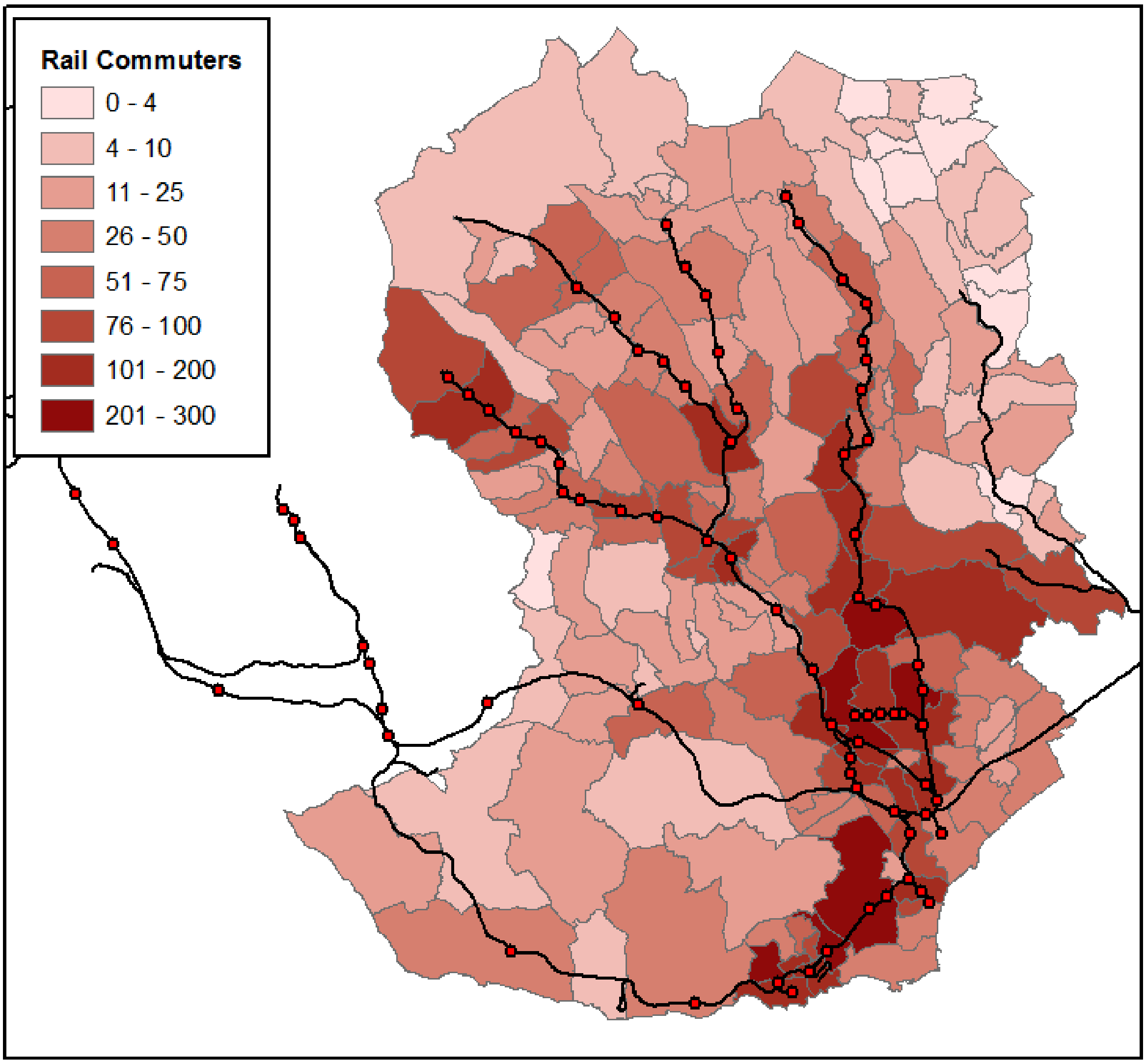

Figure 1: Wards and rail network within case study area

2.2. Global regression models

Conventionally, mode choice tends to be modelled using some form of logit model. However, such models are unsuitable for use with GWR and a number of problems had been encountered in previous attempts to calibrate aggregate logit mode split models on Census journey to work data (Blainey, 2009). It was therefore decided to test several regression model forms, as these should in theory be capable of forecasting rail's mode share and would also allow the later implementation of GWR once reliable models had been produced.

Two forms of the dependent variable were tested with all model types, firstly the total number of rail commuters from the ward, and secondly the probability of a commuter from a particular ward travelling by rail. The first models to be calibrated predicted rail commuting based on ward population, distance to the nearest station, and train frequency at the nearest station. Because some wards are quite sizeable, and some contain several stations, measuring access distance and train frequency from a single point could be misleading. Data were available allowing the wards to be spatially disaggregated into smaller units (census output areas), and these were used to calculate the average access distance from the ward to the nearest railway station and the average train frequency at the nearest station using Equations (1) and (2).

$$A_{is} = \sum_{a}^{n} \left(D_{as} \frac{P_a}{P_i} \right) \quad (1)$$

$$F_{is} = \sum_{a}^{n} \left(F_{as} \frac{P_a}{P_i} \right) \quad (2)$$

Where:
A_{is} is the average access distance (in km) for an individual living in ward *i* to their nearest station *s*
D_{as} is the road distance (in km) from the population-weighted centroid of output area *a* to its nearest station *s*
F_{is} is the average train frequency at the nearest station to an individual living in ward *i*
F_{as} is the train frequency at the nearest station (in distance) to the population-weighted centroid of output area *a*
P_a is the working population resident in output area *a*
P_i is the working population resident in ward *i*
a…n are the output areas making up ward *i*

Linear, semi-log and double-log model forms were tested for predicting the total number of rail commuters to work, with the best fit (R_{adj}^2=0.533) given by the semi-log Model 3. In addition to these three forms, logistic regression was also tested for predicting the probability of commuters using the train, as it is questionable whether the other forms are suitable for use with a probability-based dependent variable. The logistic regression model (Model 4) was found to give a better fit (R_{adj}^2=0.680) than the other model forms, and this was therefore adopted as the preferred probabilistic model after this initial stage of the research. The use of access time rather than access distance as the measure of proximity to railway stations was tested with both the preferred model forms, but was found to give inferior results.

$$\hat{P}_{it} = \propto + \beta lnP_i + \gamma lnA_{is} + \delta lnF_{is} \quad (3)$$

$$ln\left(\frac{Pr_{it}}{1 - Pr_{it}}\right) = \propto + \beta lnP_i + \gamma lnA_{is} + \delta lnF_{is} \quad (4)$$

Where:
P_{it} is the number of people resident in ward *i* travelling to work by train
Pr_{it} is the proportion of the working population in ward *i* travelling to work by train

In addition to population, data on a number of other socioeconomic variables were obtained from the 2001 census via CASWEB. These included the number of people within each National Statistics SocioEconomic Class (14 variables; see Office for National Statistics, 2005), household car ownership as the number of households owning certain numbers of vehicles and as a mean value per ward (6 variables), the number of residents of each ward within different age bands (15 variables), the number of residents of each ward attaining certain levels of educational achievement (5 variables), and the number of households containing dependent children (1 variable). All of these variables could be included in the models either as absolute values or as proportions of the population of each ward.

A further variable likely to affect the likelihood of rail commuting is the level of bus service provided in each ward. It proved extremely difficult to incorporate this factor in previous rail demand models (Blainey, 2009; Blainey & Preston, 2010), largely because of deficiencies in the available data. For these models a variable representing the number of bus stops located in each ward was tested, as this should give some indication of the bus service density present in each case. Data on bus stop locations was obtained via Google Earth and plotted in ArcGIS, allowing them to be allocated to particular wards. Such a variable is not a perfect solution to the problem, as it does not account for differences in service frequencies between stops, but no better option was available for this work.

In total the data described above provided 83 variables with potential for inclusion in the models. While a significant parameter could be obtained for most of the variables by adding them individually to Models 3 and 4, it was obvious that including them all in a single model would not give sensible results. The SPSS backward stepwise calibration procedure was therefore used with each of the preferred regression model types (semi-log number of rail commuters and logistic probability of commuting by rail) to determine the combination of variables which gave the best results. This procedure works by entering all possible variables into the model and then sequentially removing them. The variable with the smallest partial correlation with the dependent variable is considered for removal first. If the F probability is greater than 0.1 then the variable is removed, and the model is rerun. This procedure is then repeated with the independent variable with the next smallest partial correlation, and this continues until there are no variables in the model which satisfy the removal criteria. Implementing this procedure showed that the optimal combination of variables was given by Models 5 and 6 which had R_{adj}^2 values of 0.598 and 0.736 respectively.

$$\hat{P}_{it} = \alpha + \beta lnP_i + \gamma lnA_{is} + \delta lnF_{is} + \kappa_1 lnCP_{i0} + \kappa_2 lnCP_{i2} + \lambda_1 lnAgP_{i16-17} + \lambda_2 lnAgP_{i18-24} + \nu_1 lnQ_{io} + \nu_2 lnQ_{i45} \quad (5)$$

$$Ln\left(\frac{Pr_{it}}{1 - Pr_{it}}\right) = \alpha + \gamma lnA_{is} + \delta lnF_{is} + \kappa_1 lnCP_{io} + \kappa_2 lnCP_{i2} + \lambda_1 lnAgP_{i0-4} + \nu_1 lnQ_{io} + \nu_2 lnQ_{i45} + \zeta lnB_i \quad (6)$$

Where:
CP_{i0} is the probability of a household in ward *i* owning no cars
CP_{i2} is the probability of a household in ward *i* owning two cars
AgP_{i0-4} is the probability of a resident of ward *i* being aged between 0 and 4
AgP_{i16-17} is the probability of a resident of ward *i* being aged between 16 and 17
AgP_{i18-24} is the probability of a resident of ward *i* being aged between 18 and 24
Q_{i0} is the number of people resident in ward *i* with no formal qualifications
Q_{i45} is the number of people resident in ward *i* qualified to level 4/5
B_i is the number of bus stops in ward *i*

While these models were not predicting travel to a particular destination, given that central Cardiff is the primary centre of employment in the area it seemed likely that proximity to this city would have an impact on the propensity to commute by rail. The addition of variables representing this proximity to Models 5 and 6 was therefore tested, with proximity measured by mean rail distance, mean rail journey time and mean rail speed (calculated using equations (7), (8) and (9) respectively).

$$D_{ic} = \sum_{a}^{n}\left(D_{aC}\frac{P_a}{P_i}\right) \quad (7) \qquad T_{ic} = \sum_{a}^{n}\left(T_{aC}\frac{P_a}{P_i}\right) \quad (8) \qquad Rs_{ic} = \sum_{a}^{n}\left(Rs_{aC}\frac{P_a}{P_i}\right) \quad (9)$$

Where:
D_{iC} is the average access distance (in km) for an individual living in ward *i* to the closest central Cardiff station via their nearest origin station
D_{aC} is the distance (in km) from the population-weighted centroid of output area *a* to its nearest station by road and then by rail to the closest central Cardiff station
T_{iC} is the average rail journey time (in minutes) to a central Cardiff station from the nearest station (by access time) to ward *i*
T_{aC} is the rail journey time (in minutes) to a central Cardiff station from the nearest station (by access time) to the population-weighted centroid of output area *a*
Rs_{iC} is the average rail speed (in km/h) to a central Cardiff station from ward *i* via the nearest station (by access time) to ward *i*
Rs_{aC} is the rail speed (in km/h) to a central Cardiff station from the population-weighted centroid of output area *a* via its nearest station (by access time)

Using rail journey time as the measure of proximity gave the best model fit, and the removal of some insignificant variables gave the preferred ward level models, 10 and 11. The results from calibrating these models are summarised in Table 1. Some parameter values were unexpected, with for example Model 11 suggesting that increasing bus stop density corresponds to increased rail use, implying that rail and bus are complimentary rather than competing modes in this area.

$$\hat{P}_{it} = \alpha + \beta lnP_i + \gamma lnA_{is} + \delta lnF_{is} + \kappa_1 lnCP_{i0} + \kappa_2 lnCP_{i2} + \lambda_1 lnAgP_{i16-17} + \lambda_2 lnAgP_{i18-24} + \nu_1 lnQP_{io} + \nu_2 lnQP_{i45} + \psi \ln T_{iC} \quad (10)$$

$$Ln\left(\frac{Pr_{it}}{1-Pr_{it}}\right) = \alpha + \gamma lnA_{is} + \delta lnF_{is} + \kappa_1 lnCP_{io} + \kappa_2 lnCP_{i2} + \lambda_1 lnAgP_{i0-4} + \nu_1 lnQ_{io} + \zeta lnB_i + \psi \ln T_{iC} \quad (11)$$

Table 1 – Summarised results from calibration of Models 10-11

Model	10		11	
Parameter	Value	t stat	Value	t stat
Intercept	-642.353	-4.285	-0.251	-0.132
Population	113.986	3.556	n/a	n/a
Access distance	-35.102	-9.734	-0.993	-20.820
Train frequency	31.497	4.144	0.630	6.075
Prob. 0 cars	41.500	2.024	0.515	1.872
Prob. 2 cars	66.170	2.773	0.834	2.617
Prob. age 0-4	n/a	n/a	0.754	3.432
Prob. age 16-17	-38.166	-1.892	n/a	n/a
Prob. age 18-24	-27.311	-2.330	n/a	n/a
Qualified to level 0	-48.943	-2.107	-0.400	-2.889
Qualified to level 4/5	-18.579	-1.383	n/a	n/a
Bus stops	n/a	n/a	0.181	1.756
Time to central Cardiff	14.599	3.254	0.287	5.118
R_{adj}^2		0.621		0.757

The prediction errors from Model 10 are mapped in Figure 2, to illustrate any spatial patterns in model accuracy which may exist. There do not appear to be any major variations of this kind, although consistent over/underprediction can be seen along some of the valleys, and the model seems to be particularly inaccurate in some of Cardiff's outer suburbs. Mapping of the prediction errors from Model 11 showed a very similar pattern. However, calculation of average deviation values for the number of rail commuters predicted by both models (with Model 11 predictions obtained by multiplying the predicted probability of using rail in a ward by the ward's working population) showed that Model 11 gave a far better representation of reality (AD=0.506) than Model 10 (AD=1.532) and Model 11 should therefore be preferred.

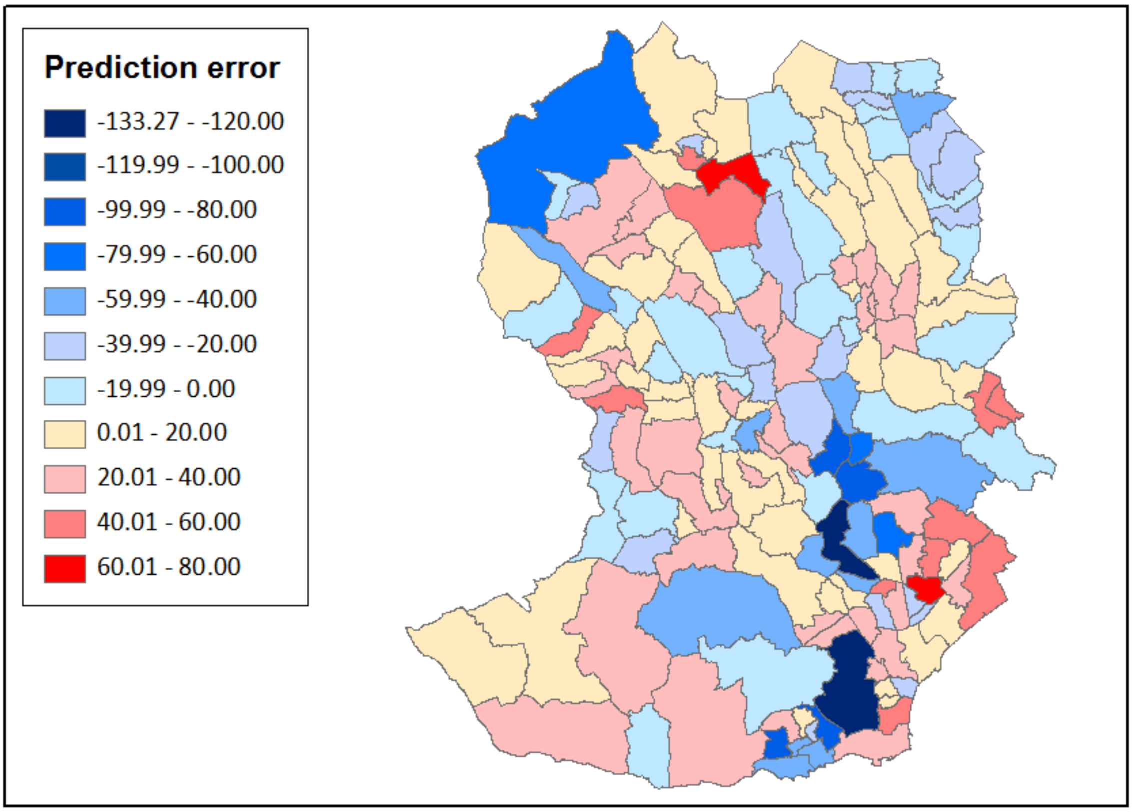

Figure 2 – Prediction errors from Model 97

2.3. GWR models

While no major spatial trends in accuracy could be identified from the mapped residuals, it still seemed possible that recalibrating the models using GWR might improve model fit. The GWR methodology is integrated into the regression process, which makes it easy to integrate spatial characteristics into existing models, such as those described in Section 2.2. For a 'traditional' multivariate global regression model like (12), the corresponding GWR model is given by (13) (Fotheringham et al, 2002).

$$y_i = \alpha + \sum_k \beta_k x_{ik} + \varepsilon_i \quad (12)$$

$$y_i = \alpha(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (13)$$

Where:
(u_i,v_i) denotes the coordinates of the *i*th point in space
$\beta_k(u_i,v_i)$ is a realisation of the continuous function $\beta_k(u,v)$ at point *i*

In GWR each data point is weighted by distance from a local regression point by fitting a spatial kernel in the form of a distance decay function to the data. As the regression point is moved across the region this gives unique parameter estimates for each location based on the varying data point weightings (Fotheringham et al, 2002). In other words, when forecasting the number of trips made by rail from a particular ward based on the values at that point of the model independent variables, the unique parameter estimates used in this forecast will be influenced more by the observed values of the independent variables in nearby wards than the values in distant wards. Software for GWR is available from the National Centre for Geocomputation at the National University of Ireland and was used to calibrate the models described here.

The best global regression models (10 and 11) were recalibrated using GWR to establish whether there was significant spatial variation in the variables influencing rail demand. A Gaussian model form was used, with adaptive kernel bandwidths determined by Akaike Information Criterion (AIC) minimisation. Using adaptive bandwidths means that the number of wards considered during local parameter estimation is not affected by variations in ward size across the case study area. Consideration was given to omitting the 'journey time to central Cardiff' variable, but while this could be seen as being a 'spatial' variable and therefore one whose effects should be captured by GWR as spatial variation in other variables, it seemed reasonable to assume that the relative effects of journey time to central Cardiff might not be constant over space, and therefore the variable was retained. The results of the GWR calibration of these models are summarised in Table 2.

Table 2 – Summarised results from GWR calibration of Models 10 and 11

Model 10						
Parameter	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Monte Carlo P-value
Intercept	-1856.242	-995.303	-370.666	-107.844	382.369	0.040
Population	-86.024	29.151	57.943	85.199	202.022	0.980
Access distance	-87.902	-59.783	-30.078	-17.174	-11.579	0.000
Train frequency	-44.320	10.365	26.631	50.884	105.138	0.030
Prob. 0 cars	-76.196	-14.561	8.028	49.130	107.099	0.670
Prob. 2 cars	-73.240	-8.671	13.807	76.422	174.956	0.590
Prob. age 16-17	-136.831	-81.727	-11.621	26.241	58.436	0.190
Prob. age 18-24	-91.034	-34.637	-5.822	14.632	54.891	0.750
Qualified to level 0	-155.928	-28.395	-14.247	8.696	70.824	0.980
Qualified to level 4/5	-54.630	-13.386	1.023	22.783	68.407	0.870
Time to central Cardiff	-85.706	-27.196	-0.369	21.284	95.750	0.000
R_{adj}^2	0.833		F stat		4.461	
Model 11						
Intercept	-11.818	-3.001	-1.532	5.140	18.594	0.030
Access distance	-1.164	-0.995	-0.934	-0.842	-0.506	0.020
Train frequency	-0.308	0.343	0.620	0.770	1.458	0.010
Prob. 0 cars	-0.515	0.205	0.482	1.091	2.430	0.260
Prob. 2 cars	-0.387	0.409	0.636	1.031	2.048	0.820
Prob. age 0-4	-0.296	0.331	0.495	0.846	2.066	0.220
Qualified to level 0	-1.174	-0.578	-0.268	-0.241	0.291	0.400
Bus stops	-0.255	-0.078	0.105	0.417	0.810	0.110
Time to central Cardiff	-1.875	-0.513	0.016	0.377	1.107	0.000
R_{adj}^2	0.839		F stat		2.954	

Comparison of the R_{adj}^2 values in Tables 1 and 2 show that calibration using GWR gives a major improvement in fit over the conventional global regression models. This is confirmed by the significant F statistics in Table 2, which indicate that the GWR models fit the data better than the global models. An indication of the spatial variation in the GWR parameters can be obtained by comparing the values given in Table 2, but not all this variation is necessarily significant. The Monte Carlo P-values in the final column show whether there is significant spatial variation in the parameters, with a value of 0.050 or less indicating that the variation is significant at the 5% level. This shows several parameters in each model exhibited significant spatial variation, specifically the intercept, access distance, train frequency and time to central Cardiff parameters in both models. The spatial variation in the access distance and frequency parameters is mapped in Figures 3 and 4, along with the local t statistics which show where the parameters are significant.

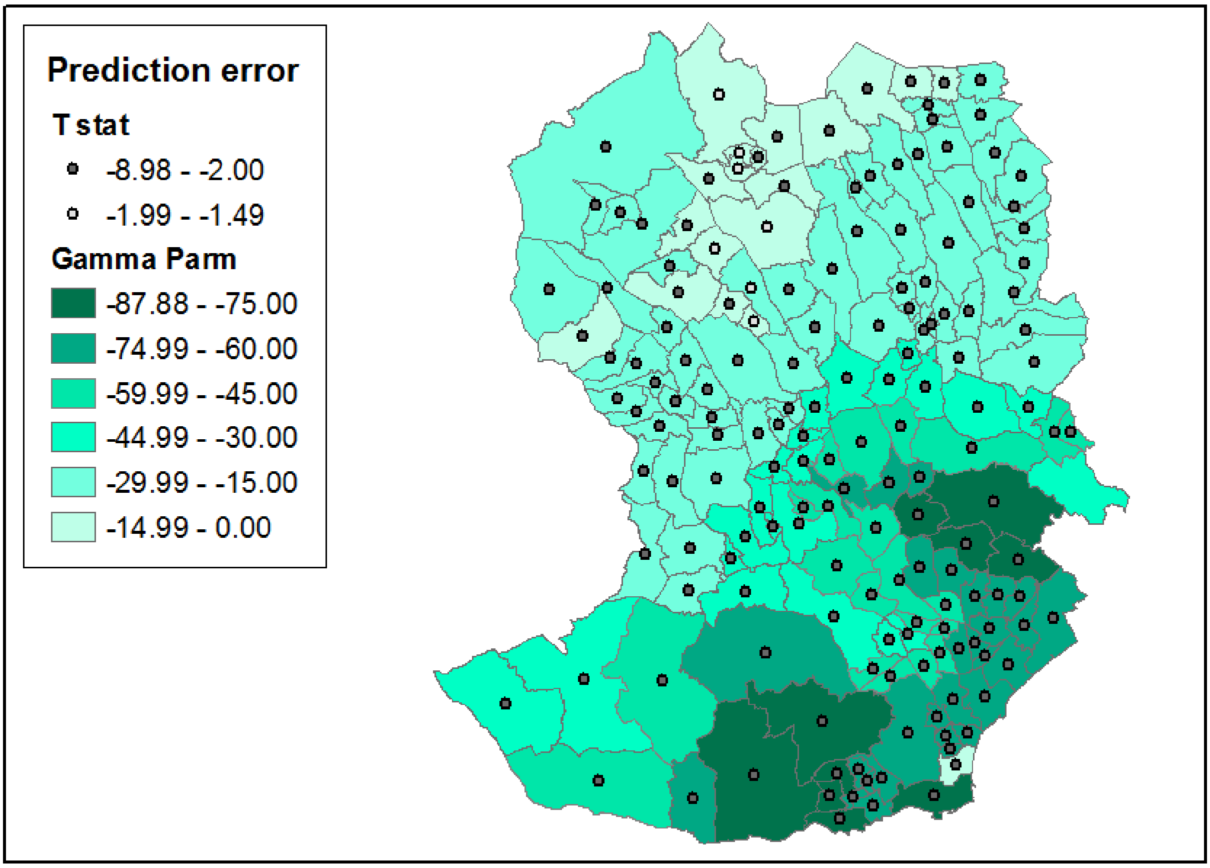

Figure 3 – Spatial variation in access distance parameter from GWR calibration of Model 10

Figure 3 indicates that station access distance has the greatest (most negative) impact on rail travel to work in the south of the case study area. This may be because this area includes some wards containing stations with an extremely good rail service to Cardiff, while adjacent wards contain no station at all, meaning that the contrast between them is emphasised. The parameter is significant in almost the entire case study area, indicating that it is one of the major determinants of rail use for travel to work.

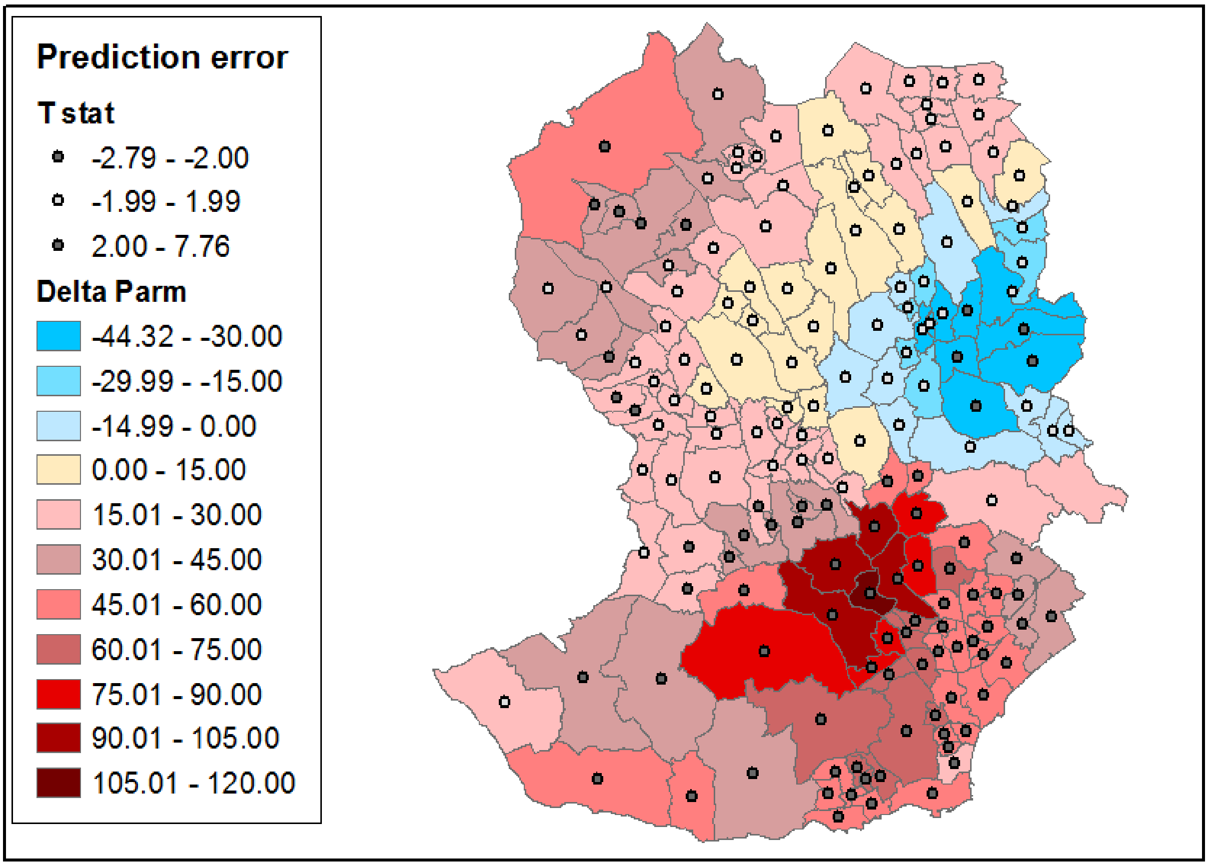

Figure 4 – Spatial variation in frequency parameter from GWR calibration of Model 10

Figure 4 shows that there is a strong contrast between the effect of train frequency in the area immediately to the west and north-west of central Cardiff (where it has a strongly positive effect) and in the north-east of the case study area, where it appears to have a negative effect. This seems rather unlikely, and may mean that there is a problem with the model, but may also result from the particular characteristics of the rail services in this area. These wards are served by but not adjacent to the Rhymney branch line, and a number of trains on this line turn back before reaching the north end of the valley, meaning that stations further south have a higher service frequency. However, peak trains on the line can be extremely busy, and it seems likely that commuters from these wards may travel via more northerly stations with a lower service frequency because this means they are more likely to get a seat. This might explain the apparently illogical parameter values, and illustrates the capability of GWR to highlight local anomalies and variations.

The prediction errors from the GWR calibration of Models 10 and 11 are mapped in Figures 5 and 6. Overall the prediction errors are much smaller than for the global model, with no obvious spatial patterns remaining. However, Figure 5 shows extremely large prediction errors (one positive and one negative) for two wards close to central Cardiff, suggesting that these possess some unusual features which are reducing forecasting accuracy. The overprediction in Grangetown ward may occur because while the stations within it provide a high quality rail service, most commuting will be over short distances to Cardiff city centre and Cardiff Bay where the bus provides a more convenient option. Figure 6 shows that the error pattern from Model 11 is extremely similar to that from Model 10, and the same was found to be the case for the spatially varying parameter values.

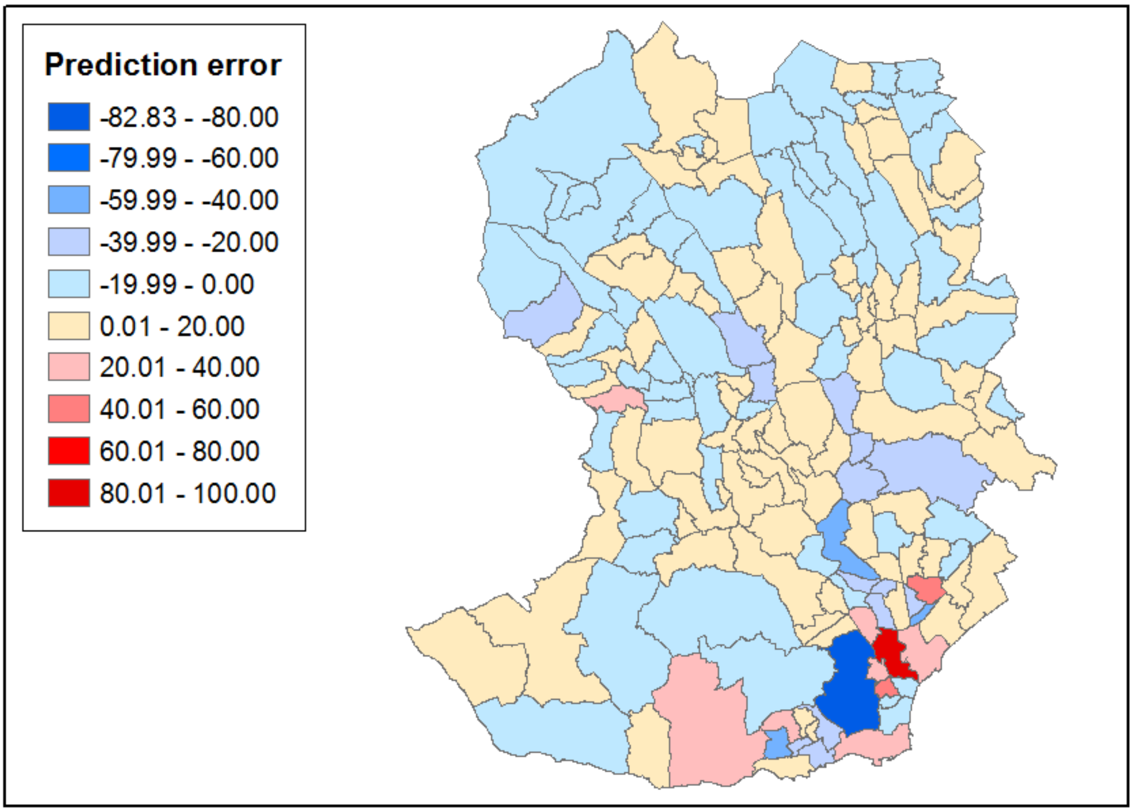

Figure 5 – Prediction errors from GWR calibration of Model 10

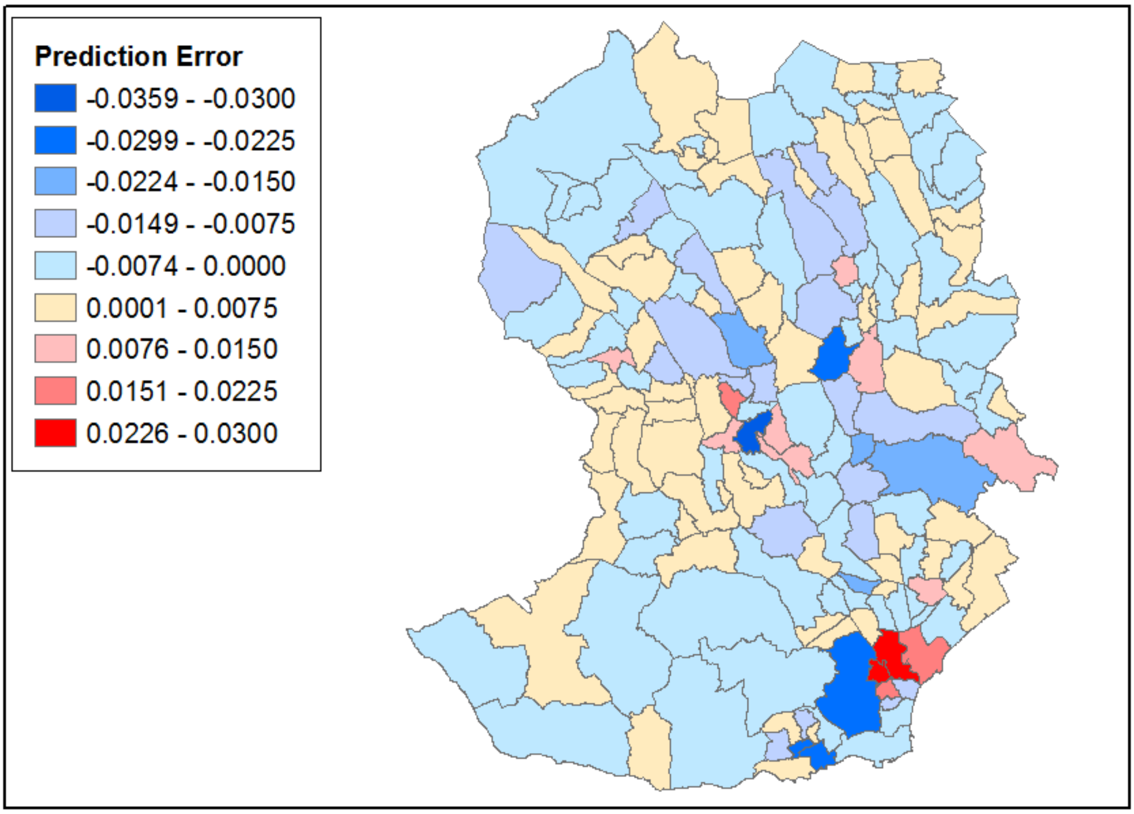

Figure 6 – Prediction errors from GWR calibration of Model 11

3) OUTPUT AREA LEVEL MODELS

Geographical variations in the effect of particular independent variables on the likelihood of travelling to work by rail may exist at a more local scale than can be captured by ward level data. The calibration procedure described above was therefore repeated using census data at the output area level. However, previous work on aggregate logit mode split models using this data (Blainey, 2009) encountered problems resulting from the use of the 'Small Cell Adjustment Methodology' (SCAM) to maintain data confidentiality. This means that if three or less people travel by a particular mode from a particular output area, the figure reported will be randomly adjusted to equal zero or three. This causes particular problems where rail is the minor mode, as is the case for most of this case study area. However, the existence of accurate ward level totals for rail use allowed SCAM-related errors to be identified by checking if the output area totals summed to give the ward totals. If a discrepancy was found then all output area totals of three or zero were adjusted by the same amount so that the OA sum would equal the ward sum, allowing models to be calibrated as before.

As with the ward level models, linear, semi log, loglinear and logit regression forms were tested, with double log models found to give the best fit when forecasting both the total number of rail commuters from the output area and the probability of a commuter from a particular output area travelling by rail. Backward stepwise calibration was used to determine the optimal combination of variables, and this showed that the best fit with the observed data was given by Models 14 and 15, although the fit of these models was very poor, as shown by Table 3, which summarises the results from calibrating these models. Comparison of AD values confirmed that Model 14 (AD=1.569) was more accurate at forecasting rail use than Model 15 (AD=1.928). The poor fit of these models may result from data deficiencies or from micro-scale variations in the effects of bus competition, and means the models cannot be used with any confidence for forecasting rail use.

$$\hat{P}_{it} = \alpha A_{is}^{\gamma} F_{is}^{\delta} SP_2^{\eta_1} SP_3^{\eta_2} Ag_{i0-4}^{\lambda_1} Ag_{i18-24}^{\lambda_2} Ag_{i45-59}^{\lambda_3} Ag_{i65-74}^{\lambda_4} QP_{i0}^{\nu_1} QP_{i3}^{\nu_2} T_{iC}^{\psi} \quad (14)$$

$$\hat{P}r_{it} = \alpha P_i^{\beta} A_{is}^{\gamma} F_{is}^{\delta} SP_2^{\eta_1} SP_3^{\eta_2} SP_4^{\eta_3} Ag_{i0-4}^{\lambda_1} Ag_{i18-24}^{\lambda_2} Ag_{i45-59}^{\lambda_3} Ag_{i65-74}^{\lambda_4} QP_{i0}^{\nu_1} QP_{i3}^{\nu_2} T_{iC}^{\psi} \quad (15)$$

Where:
SP_2 is the proportion of the working population resident in output area *i* which falls within NSSEC class 2
SP_3 is the proportion of the working population resident in output area *i* which falls within NSSEC class 3
SP_4 is the proportion of the working population resident in output area *i* which falls within NSSEC class 4
Ag_{i0-4} is the number of residents in output area *i* aged between 0 and 4
Ag_{i16-17} is the number of residents in output area *i* aged between 16 and 17
Ag_{i18-24} is the number of residents in output area *i* aged between 18 and 24
Ag_{i45-59} is the number of residents in output area *i* aged between 45 and 59
Ag_{i65-74} is the number of residents in output area *i* aged between 65 and 74

QP_{i0} is the probability of a resident in output area *i* having no formal qualifications
QP_{i3} is the probability of a resident in output area *i* being qualified to level 3

Table 3 – Summarised results from calibration of Models 14-15

Model	14		15	
Parameter	Value	t stat	Value	t stat
Intercept	-4.450	-9.997	-4.962	-8.878
Population	n/a	n/a	-0.738	-6.555
Access distance	-0.734	-35.196	-0.612	-29.897
Train frequency	0.665	14.581	0.593	13.420
Prob. NSSEC class 2	0.286	4.048	0.244	3.366
Prob. NSSEC class 3	0.117	2.067	0.100	1.740
Prob. NSSEC class 4	n/a	n/a	0.069	1.687
Age 0-4	0.102	2.602	0.075	1.832
Age 18-24	0.207	4.112	0.245	4.309
Age 45-59	0.316	4.645	0.296	3.812
Age 65-74	-0.161	-3.704	-0.167	-3.943
Prob. qualified to level 0	-0.443	-6.884	-0.355	-5.341
Prob. qualified to level 3	0.101	2.102	0.101	2.154
Time to central Cardiff	0.533	19.543	0.475	17.971
R_{adj}^2	0.402		0.310	

It seemed possible that the poor fit might result from major spatial variations in the effect of the independent variables on rail use. The models were therefore recalibrated using GWR with a Gaussian model form used as before and adaptive kernel bandwidths determined AIC minimisation. The results of this recalibration are summarised in Table 4.

Table 4 – Summarised results from GWR calibration of Models 14 and 15

Model 14						
Parameter	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Monte Carlo P-value
Intercept	-70.472	-7.358	-2.540	0.615	42.587	0.000
Access distance	-1.327	-0.812	-0.602	-0.437	0.269	0.000
Train frequency	-4.306	0.068	0.536	1.327	30.900	0.000
Prob. NSSEC class 2	-0.876	0.192	0.352	0.566	1.392	0.730
Prob. NSSEC class 3	-0.938	-0.112	0.103	0.423	1.155	0.040
Age 0-4	-0.647	-0.039	0.079	0.170	1.035	0.020
Age 18-24	-0.495	0.053	0.239	0.470	1.038	0.080
Age 45-59	-1.455	-0.066	0.206	0.421	1.842	0.000
Age 65-74	-0.777	-0.260	-0.098	0.086	0.982	0.040
Prob. qualified to level 0	-1.675	-0.502	-0.239	-0.071	1.063	0.020
Prob. qualified to level 3	-0.569	-0.094	0.083	0.222	1.038	0.220
Time to central Cardiff	-14.894	-0.784	0.111	0.840	9.179	0.000
R_{adj}^2	0.558		F stat		4.402	

Table continued on next page

Model 15						
Intercept	-60.726	-7.618	-3.464	-0.323	59.355	0.000
Population	-2.716	-0.874	-0.516	-0.167	1.539	0.210
Access distance	-1.504	-0.739	-0.486	-0.341	0.744	0.000
Train frequency	-6.168	-0.088	0.286	1.026	24.763	0.000
Prob. NSSEC class 2	-1.007	0.125	0.304	0.494	1.356	0.570
Prob. NSSEC class 3	-0.975	-0.091	0.072	0.279	1.324	0.190
Prob. NSSEC class 4	-0.527	-0.176	-0.050	0.066	0.638	0.630
Age 0-4	-0.522	-0.081	0.027	0.123	1.160	0.010
Age 18-24	-0.526	0.022	0.196	0.371	1.116	0.500
Age 45-59	-1.477	-0.181	0.145	0.345	1.276	0.180
Age 65-74	-1.023	-0.223	-0.102	0.063	1.130	0.000
Prob. qualified to level 0	-2.179	-0.449	-0.150	0.108	1.560	0.000
Prob. qualified to level 3	-0.524	-0.076	0.071	0.206	1.096	0.190
Time to central Cardiff	-14.288	-0.621	0.037	0.759	10.411	0.000
R_{adj}^2	0.512		F stat		4.395	

As expected, the use of GWR gave a significant improvement in fit over the global calibration for both model forms, although even the fit of the GWR calibration was only moderate. A number of parameters exhibited significant spatial variation, specifically the intercept, access distance, train frequency, age 0-4, age 65-74, level 0 qualifications and time to central Cardiff parameters in both models, and the NSSEC class 3 and age 45-59 parameters in Model 14. As before, this significant spatial variation was mapped to allow any patterns to be identified. There is insufficient space to show all these maps here, but Figures 7 and 8 show the local variation in the access distance and train frequency parameters to allow comparison with the variation from the ward level models. These figures show some major differences in the spatial variation in the equivalent parameters from the ward and output area models. The access distance parameter from the output area models is strongly negative in the north-east of the case study area, whereas in the ward level models it had a relatively minor effect in this area (see Figure 3). Similarly, the train frequency parameter in the output area level model exhibits relatively little variation across most of the case study area, but is strongly positive in a small area in the north-west. This contrasts with the ward level model (see Figure 4) where there is much more variation and where the parameter was most strongly positive towards the centre of the case study area. More trust should be placed in the variation from the ward level models given the better fit of this model and the problems with the output area data, but the results from the output area models should not be discarded entirely. The prediction errors from the GWR calibration of Model 14 are mapped in Figure 9. This map is rather misleading, as the colouring is skewed by some extremely large prediction errors at a few small output areas. The general pattern of errors is in fact quite similar to that shown in Figure 5, although there is obviously much more variation in Figure 9 because of the smaller size of the individual units.

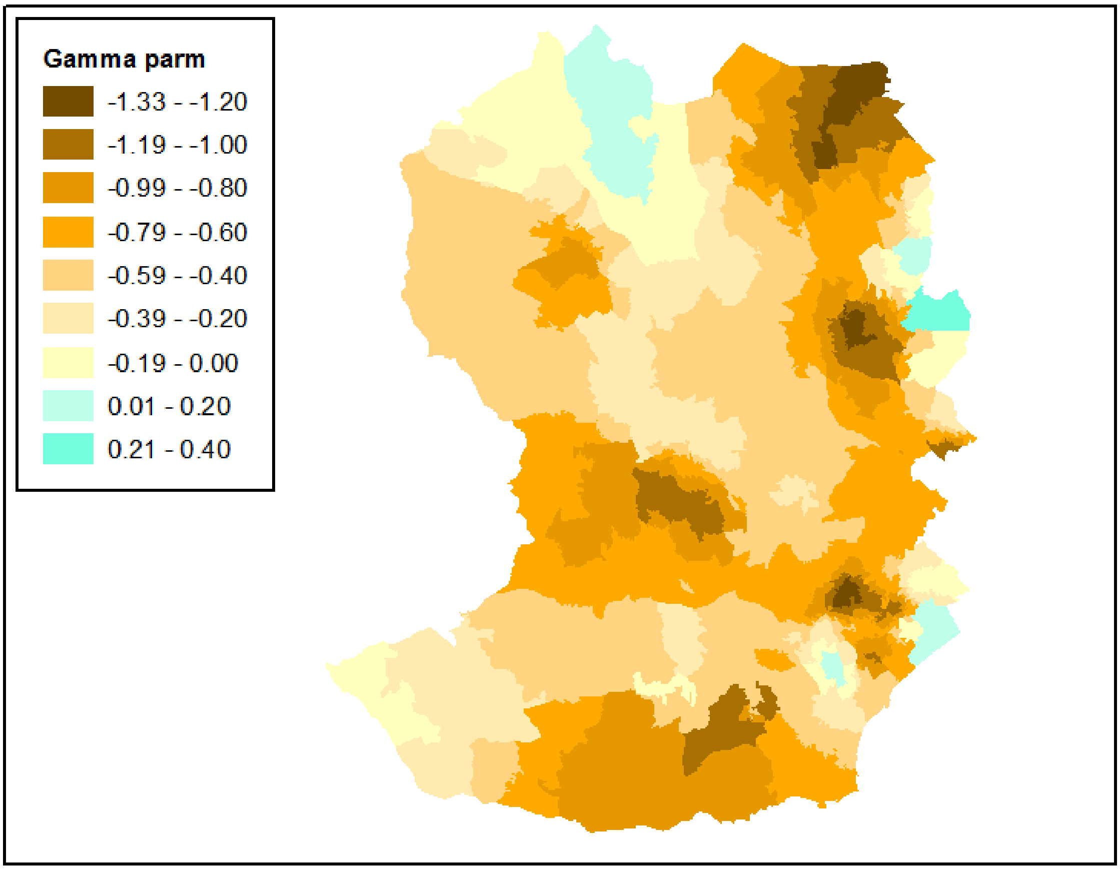

Figure 7 – Spatial variation in access distance parameter from GWR calibration of Model 14

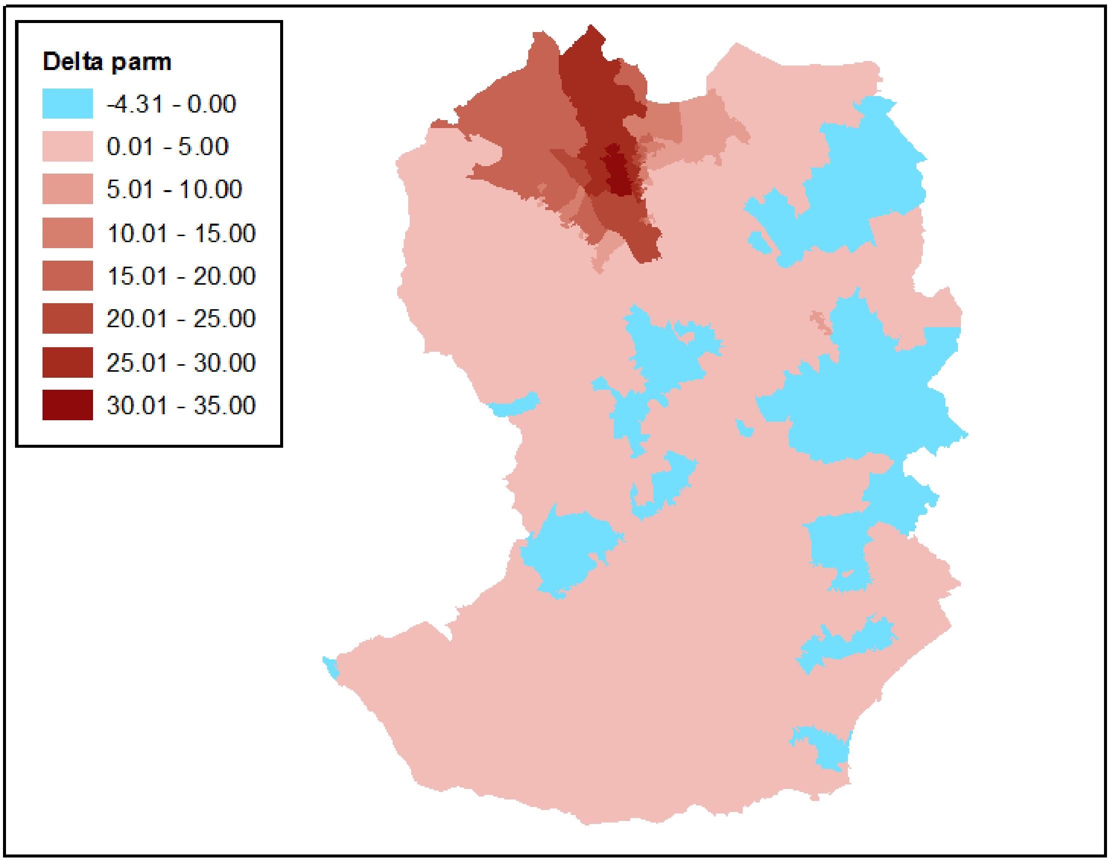

Figure 8 – Spatial variation in frequency parameter from GWR calibration of Model 14

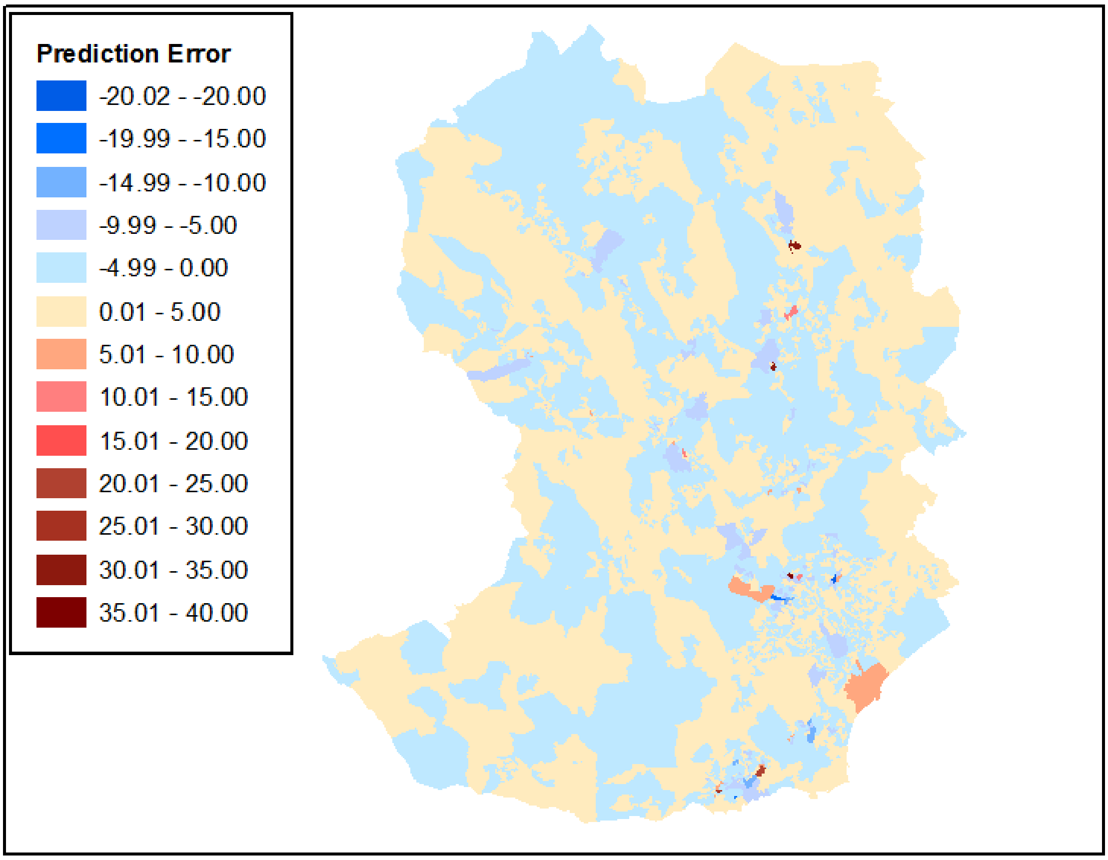

Figure 9 – Prediction errors from GWR calibration of Model 14

4) FLOW LEVEL DEMAND MODELS

4.1 Conventional direct demand models

While the models described above should be reasonably effective at forecasting rail's mode share of the travel to work market, and at identifying spatial variations in the parameters determining this share, these models are of limited use in forecasting the demand at new stations as they will only predict commuting trips. There is however potential for incorporating the spatial variations identified during this work in direct demand models, which can forecast total usage on specified flows from a new station rather than just work trips.

The best direct demand model with generalised origin variables developed during previous work by Blainey (2009) is given by (16). This model was recalibrated on the largest flows from all 68 stations within the case study area for the ward and output area level travel to work models. These flows were selected by ranking the flows from each station in descending order of size, and then selecting progressively smaller flows until 95% of the total trip origins at from each station were included in the dataset, giving a total of 1289 flows. Calibration of Model 16 on this dataset gave the results summarised in Table 5. While the model only had a moderate fit, all the parameters except the road speed parameter were of the correct sign and strongly significant, meaning that it could form the basis for further investigations.

$$\hat{T}_{ij} = \alpha\left(\sum_{a} P_a W_a\right)^{\beta} J_{i4}^{\tau} Pk_i^{\rho}\left(\prod_{j}^{n} D_j^{\gamma}\right) D_{ij}^{\omega} Rs_{ij}^{\delta} Cs_{ij}^{\kappa} H_{ij}^{\eta} Rfkm_{ij}^{\lambda} \quad (16)$$

Where:

$\hat{T}_{ij}$ is the predicted number of trips made from station *i* to station *j*

P_a is the population in output area *a*, for which station *i* is the closest station

$w_a = (t+1)^{-3.25}$

t is the travel time by road from output area *a* to station *i*

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

Pk_i is the number of parking spaces at station *i*

D_j is a dummy variable which takes the value 1 if *j* is station *j*, and 0 otherwise

Di_{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*

Table 5 – Summarised results from calibration of Model 16

Parameter	Value	t stat
Intercept	7.244	12.807
Weighted population	0.315	8.778
Jobs	0.179	5.586
Parking spaces	0.177	8.805
Distance	-1.698	-12.439
Rail speed	0.827	10.140
Car speed	0.148	0.479
Headway	-0.719	8.742
Rail fare per km	-2.042	-10.432
R_{adj}^2	0.537	

4.2 Incorporation of spatial parameter variation

Using GWR directly in the calibration of direct demand models is not straightforward, because flows do not have point-based geographical locations, making it difficult to assign them to a single set of coordinates. However, it may be possible for geographical variations identified in other models to be transferred to direct demand models.

Looking first at the results from Model 10, no significant spatial variation was found in the population parameter. However, there was significant spatial variation in the access distance parameter, and this could be used to weight the population variable in the direct demand model. This process is complicated by the fact that the catchment population term in Model 16 is a composite of the populations in the output areas closest to the origin station in question, weighted by access distance. It is therefore necessary to further weight these access distances using the spatial variation in the parameter estimates. However, it is not sufficient to simply multiply the access distance by the parameter value from Model 10 for the relevant ward, because this would make some of the access distances strongly negative. The spatially varying parameter values were therefore standardised using (17).

$$\beta_{Si} = 1 + \frac{(\beta_i - \mu)}{\mu} \quad (17)$$

Where:
β_{Si} is the standardised parameter value for ward i
β_i is the unstandardised parameter value for ward i
μ is the mean of the β_i parameter values across all wards

The access distances for each output area were then weighted by multiplying them by the standardised parameter value for the corresponding ward. The weighted catchment population for each station could then be calculated using the weighting function (18), and this in turn allowed Model 19 to be calibrated giving the results summarised in Table 6. Similar spatial variation was found in the access distance parameter from the GWR calibration of Model 11, and this was used to weight the population terms in the same way. The resulting recalibration of Model 19 is also summarised in Table 6.

$$P_{iw} = \sum_a (P_a (Aw_a + 1)^{-3.25}) \quad (18)$$

$$\hat{T}_{ij} = \alpha P_{iw}^{\beta} J_{i4}^{\tau} Pk_i^{\rho} \left(\prod_j^n D_j^{\gamma} \right) D_{ij}^{\omega} Rs_{ij}^{\delta} Cs_{ij}^{\kappa} H_{ij}^{\eta} Rfkm_{ij}^{\lambda} \quad (19)$$

Where:
P_{iw} is the weighted catchment population for station i
P_a is the resident population in output area a, for which station i is the nearest station
Aw_a is the access distance from output area a to station i, weighted using (17)

Table 6 – Summarised results from calibration of Model 19

Distance weighting	GWR Model 10		GWR Model 11	
Parameter	Value	t stat	Value	t stat
Intercept	7.904	13.986	6.393	10.296
Weighted population	0.263	6.267	0.484	8.156
Jobs	0.191	5.868	0.114	3.252
Parking spaces	0.205	10.216	0.216	10.905
Distance	-1.749	-12.623	-1.764	-12.869
Rail speed	0.773	9.297	0.809	9.884
Car speed	0.460	1.490	0.663	2.208
Headway	-0.886	-10.373	-0.686	-8.266
Rail fare per km	-2.002	-10.080	-2.024	-10.300
R_{adj}^2	0.523		0.534	

The fit of both calibrations of Model 19 is inferior to that of Model 16, although the second calibration of Model 19 (with access distance weighted using parameter values from Model 11) should perhaps be preferred over Model 16 because all parameters are significant. There was also spatial variation in the train frequency parameter in Models 10 and 11 but while its use to weight the headway variable in the direct demand model was considered, this did not prove possible because the variables are not directly comparable. Similarly, there was no variable in Model 16 which corresponded with the only other parameter in Models 10 and 11 that exhibited significant spatial variation, the distance to Cardiff parameter. Overall, therefore, the incorporation of the spatially-weighted population parameter in direct demand models using this methodology does not give a significant improvement in model fit.

4.3 GWR calibration of direct demand models

The main problem with applying GWR to direct demand models is that, by definition, a rail flow does not have a single point location. However, because Model 16 accounts for origins in much more detail than destinations, it can be argued that each flow should be allocated the coordinates of the origin station. Ideally the model should be calibrated as a mixed GWR model, with some parameters stationary over space and some allowed to vary, but the current version of the GWR software cannot calibrate such models. A normal GWR model therefore had to be used, even though this form is of dubious validity for this type of application. Model 16 could not be calibrated using GWR, as the number of destination dummy variables required would exceed the total variable limit, and these were therefore replaced by a total exit variable giving Model 20. Such a variable was shown to be a reasonable substitute for destination dummy variables in previous work (Blainey & Preston, 2010). The results from a global calibration of Model 20 are shown in Table 7.

$$\hat{T}_{ij} = \alpha \left(\sum_a P_a W_a \right)^{\beta} J_{i4}^{\tau} Pk_i^{\rho} Ex_j^{\gamma} D_{ij}^{\omega} Rs_{ij}^{\delta} Cs_{ij}^{\kappa} H_{ij}^{\eta} Rfkm_{ij}^{\lambda} \quad (20)$$

Where:
Ex_j is the total number of trips ending at station j in the year

Table 7 – Summarised results from calibration of Model 20

Parameter	Value	t stat
Intercept	1.031	1.798
Weighted population	0.302	8.067
Jobs	0.146	4.464
Parking spaces	0.192	9.099
Destination exits	0.579	19.539
Distance	-1.324	-18.940
Rail speed	0.855	12.091
Car speed	0.572	2.263
Headway	-0.498	-6.525
Rail fare per km	-1.160	-12.400
R_{adj}^2	0.457	

The fit of this model is noticeably worse than that of Model 16, although all parameters are significant (with the exception of the intercept). However, Model 20 is suitable for GWR calibration, and it may therefore be possible to improve the fit of this model. The results of such a calibration are summarised in Table 8.

Table 8 – Summarised results from GWR calibration of Model 20

Model 20						
Parameter	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Monte Carlo P-value
Intercept	-10.626	-2.727	2.946	4.974	20.209	0.000
Weighted population	-0.987	-0.095	0.240	0.567	0.983	0.000
Jobs	-1.316	-0.158	0.139	0.350	0.847	0.000
Parking spaces	-0.187	0.000	0.143	0.281	0.847	0.000
Destination exits	0.377	0.558	0.636	0.712	0.850	0.100
Distance	-2.141	-1.738	-1.521	-1.386	-1.013	0.150
Rail speed	-0.010	0.426	0.743	0.947	1.767	0.000
Car speed	-0.907	1.219	1.583	2.562	4.609	0.000
Headway	-1.262	-0.722	-0.508	-0.140	2.660	0.000
Rail fare per km	-2.481	-1.269	-1.027	-0.725	-0.394	0.010
R_{adj}^2	0.668		F stat		9.263	

Table 8 shows that calibrating Model 20 using GWR gave a major improvement in model fit over the global calibration, and also over the global calibration of the destination dummy variable Model 19. Significant spatial variation was found in all parameters except for those representing the destination exits and distance variables. This is perhaps surprising, as it includes some variables relating to particular flows rather than to particular origins, such as the speed of different modes and train headway. However, this may indicate that these factors are more important in determining demand for all flows at some origins than they are at others.

The spatial variation was mapped for all parameters where it was significant, with inverse distance weighting (IDW) interpolation used in ArcGIS to create a parameter surface from the values at each station. There is insufficient space to include all these maps here, so only two with features of particular interest are included. Figure 10 shows the spatial variation in the employment parameter, which appears to show that the level of employment around the origin station has a positive effect on rail demand in the centre of the case study and around Cardiff, but has a negative effect in the Vale of Glamorgan (the south-western part of the case study area), in the north of the area, and around Cardiff Queen Street station. The latter observation is unlikely to mirror reality as it almost certainly results from problems encountered in defining catchments for the two main stations in central Cardiff, Queen Street and Central (Blainey & Preston, 2010). Census output areas (the basic unit used to define catchments) are geographically defined based on population density, which is low in the central area of Cardiff even though a large number of people are employed there. This, combined with the close proximity of the two stations, means that a single output area covers the whole of the area between them. In catchment definition output areas must be allocated to a single station, meaning that the catchment employment figures for both stations will be extremely skewed. One way to overcome this problem would be to allow stations to 'compete' for the employment and population in particular output areas, but this would make model calibration much more complicated. An alternative would be to use census data which has been reassigned into a raster (cell-based) format to define catchments (see for example Martin, 1989), as this could overcome the problems caused by arbitrary output area boundaries and therefore improve catchment accuracy.

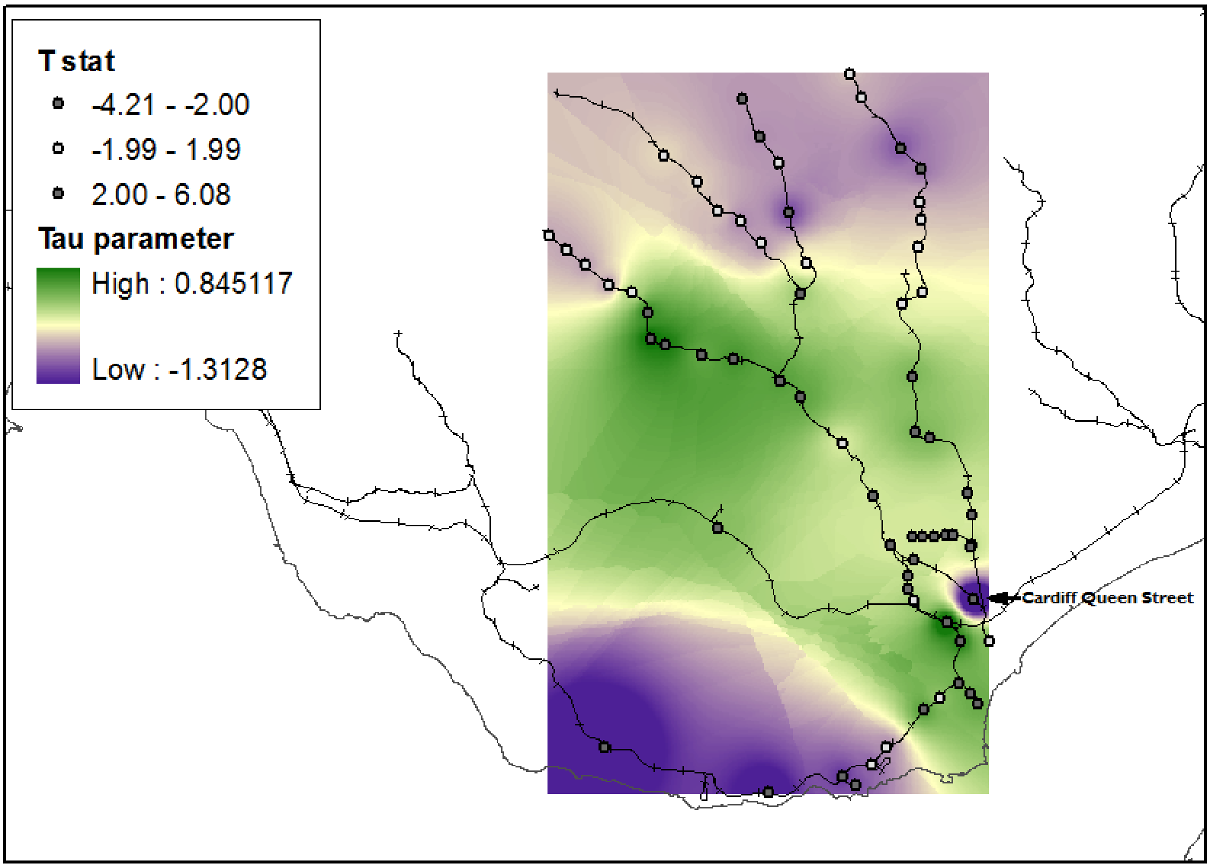

Figure 10 – Spatial variation in employment parameter from GWR calibration of Model 20

Figure 11 shows the spatial variation in the train headway parameter across the case study area. This approximately corresponds to the train frequency parameters in the ward and output level models but effectively has the opposite effect, as a low headway is equivalent to a high train frequency. This means that a negative value in Figure 11 corresponds to a positive value in Figures 4 and 8. As described above the output area model parameter values are rather suspect and there is little similarity between Figures 8 and 11. However, there is some consistency between Figures 4 and 11, with train frequency having a particularly positive impact on rail demand in the north-western suburbs of Cardiff in both models. However, the apparent negative impact of train frequency shown in Figure 11 in the north-west and south of the case study area (although the parameter is insignificant in the latter area) is not reflected in Figure 4, and seems unlikely to mirror reality.

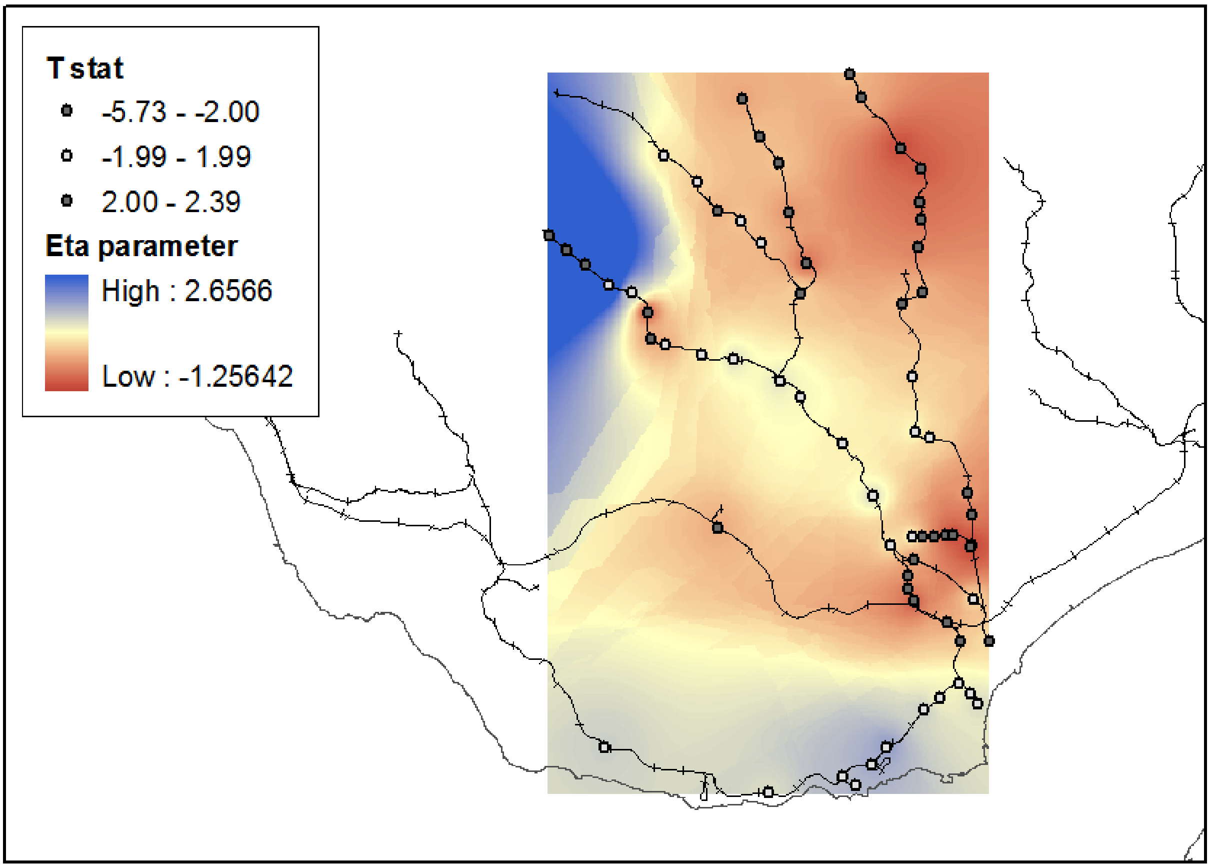

Figure 11 – Spatial variation in headway parameter from GWR calibration of Model 20

5) CONCLUSIONS AND FUTURE WORK

5.1 Main findings

The first major finding from this research was that models could be calibrated based on census journey to work data which indicate that a range of socio-economic variables can have a significant impact on levels of rail use. Although a few previous models have incorporated such variables (Preston, 1991), in general it has proved difficult to establish the relationships between these variables and rail demand. Correlations between such variables are likely to be complex, and hence the best models developed here only contain a small number of those tested. However, while this work was not aiming to directly analyse the effect of socio-economic variables on rail demand, the findings from it indicate that there is potential for such analysis of these relationships using the same data sources.

A related finding was that the density of bus stops in a given area may have a positive impact on rail demand, suggesting that rail and bus are complementary rather than competing modes, at least in this area. It is possible that bus stop density may be acting as a proxy for other variables omitted from the models, and it would therefore be desirable to incorporate a more detailed representation of bus services in the models. Collating a suitable dataset for this may be difficult, but this area again appears worthy of further investigation and could lead to improvements in the accuracy of demand forecasts.

It proved possible to calibrate regression models which give an accurate prediction of the level of rail commuting at ward level within the case study area, and the application of GWR to the best model forms gave a major improvement in model fit. Significant spatial variation was found in several parameters, indicating that their effect on rail demand varies across the case study area. The calibration of similar models at census output area level was rather less successful. The use of GWR did allow a reasonable level of fit to be obtained, but mapping the residuals and the prediction errors highlighted anomalies in the results, which suggests that data problems may limit the potential for modelling at this spatial scale.

Conventional direct demand (rail flow level) models developed in previous work were recalibrated for this case study area (so far the largest-scale application of these particular models). Attempts were made to weight some of the variables in these models using spatially varying parameter values from the ward level models, but this did not give any improvement in model fit. However, somewhat unexpectedly the application of GWR based on origin locations to the direct demand models gave a large improvement in fit and also allowed significant spatial variations to be identified in the majority of the model parameters.

5.2 Application to demand forecasting

This work has two main applications to rail demand forecasting. Firstly, the ward level models could be used to forecast the demand impacts of changes to rail services or of new station opening by estimating the current proportion of rail use made up of work trips, using the models to predict ward level travel to work after the changes, calculating the change in rail travel to work, and then using the current proportions to scale up this change to give an estimated total number of new trips. This procedure would not allocate trips to any particular station, and would therefore need to be used in conjunction with other models, but might for example help provide a means of accounting for the level of abstraction of trips by new stations from existing ones.

More immediately, the use of GWR in direct demand models appears to have great potential for enhancing the flow level forecasting of rail demand, by enabling local variations in the effect of parameters on rail demand to be taken into account. The models described here are still in need of some refinement, but no previous flow level model has been able to account for such spatial variations in the factors influencing demand. While the models described here would require recalibration before being used in areas other than South-East Wales, there is no fundamental reason why the general model forms should not be equally valid in other geographic contexts.

5.3 Future work

There are three particular areas in which this work should be extended. The first would involve testing the use of the ward level models in demand forecasting as described above, to establish how well they perform at predicting the demand impacts of opening particular stations or making alterations to services. The second would aim to exploit the initial success of using GWR to enhance rail demand models, which took place very late on in the work described here. Further analysis is needed to establish an optimal model form, and to extend the models so that they can account for the effects of factors not currently considered, such as intervening opportunities and bus competition. Finally, the use of raster-based station catchments should be investigated, with census data reallocated to a regular grid-based zoning system, as this might enable some of the problems with the existing catchment definition methods to be overcome, and remove obvious anomalies (such as the value of the employment parameter at Cardiff Queen Street) from the model results.

REFERENCES

Association of Train Operating Companies (ATOC) (2005). Passenger Demand Forecasting Handbook. ATOC, London.

Association of Train Operating Companies (ATOC) (2009). Creating Communities: Expanding Access to the Rail Network. ATOC, London.

Blainey, S.P. (2009). Forecasting the use of new local railway stations and services using GIS. PhD thesis, University of Southampton.

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

Blainey, S.P. and Preston, J.M., (2009). Assessing the potential performance of new local railway stations. European Transport Conference, Leeuwenhorst Conference Centre, Noordwijkerhout, The Netherlands, October 5th-7th.

Blainey, S.P. and Preston, J.M., (2010). Modelling Local Rail Demand in South Wales. Transportation Planning and Technology, 33(1), 55-73.

Clark, S. D., 2007. Estimating local car ownership models. Journal Of Transport Geography, 15, 184-197.

Du, H., Mulley, C., 2006. Relationship between transport accessibility and land value: local model approach with geographically weighted regression. Transportation Research Record, 1977, 197-205.

Fotheringham, A. S., Brunsdon, C., Charlton, M. E. (2002). Geographically Weighted Regression. Wiley, Chichester.

Martin, D., (1989). Mapping population data from zone centroid locations. Transactions of the Institute of British Geographers, 14, 90-97.

Office for National Statistics (2005). The National Statistics Socio-Economic Classification: User Manual. Palgrave Macmillan, Basingstoke.

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

LENNON ticket sales data supplied by Arriva Trains Wales

Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland. Source: 2001 Census; Key Statistics

All map data ©Crown Copyright/database right 2010. An Ordnance Survey/EDINA supplied service.

This work is based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown.