Extending geographically weighted regression from points to flows: a rail-based case study
Simon P Blainey and John M Preston
DOI: 10.1177/0954409713496987

The online version of this article can be found at:
http://pif.sagepub.com/content/227/6/724
Extending geographically weighted regression from points to flows: a rail-based case study

Simon P Blainey and John M Preston

Abstract
At present in the UK an elasticity-based approach is used to forecast changes in rail passenger demand resulting from changes in both the rail service offer and external conditions, with uplift factors calculated based on the proportional change in the level of explanatory variables over time. Changes in these explanatory variables may have differing effects on rail demand in different areas. This is currently controlled for via a limited segmentation of the market with different elasticities estimated for each segment, which inevitably limits the complexity of the variations which can be captured. This paper describes the use of geographically weighted regression (GWR) to enhance the modelling of such spatial variation. First, conventional cross-sectional demand models were calibrated covering major rail flows across the Great Britain. These models were then recalibrated using GWR to allow assessment of spatial variations in rail demand elasticities. Previous applications of GWR have almost exclusively focused on spatial data which have a single point location. This is not the case for rail flows, and this paper compares the results given by several different methods for defining point locations for flows. It also assesses different methods for approximating GWR results to simplify their application in real-life forecasting situations. The results show that the use of GWR can give a significant improvement in the fit of flow-based rail demand models, and that it is possible to spatially segment the UK passenger rail market based on the results from these models. In order to integrate such segmentations with the standard UK rail demand forecasting methodology it would, however, be necessary to extend the GWR methodology further to allow the calibration of GWR models on panel data.

Keywords
Geographically weighted regression, rail demand, railway station, elasticity

Introduction and background
Despite continuing economic stagnation, British passenger rail use has continued to grow, with the annual trip total increasing by 16% in the last 2 years and 50% in the last 10 years.1 Alongside the challenges posed by this growth, the spatial distribution of rail demand is not static, and service levels and patterns have to be continually adjusted by planners and operators in response to changes in the external variables which determine rail demand. This situation is not unique to the Great Britain, as many other developed countries (for example, France, Germany, Sweden and Switzerland) have seen similar growth in passenger numbers in recent years,2 and spatially dynamic demand variations will be experienced by virtually all rail operators worldwide. There are also significant cost challenges facing the British rail industry,3 which create pressures to increase fares where the market is most able to bear it and to make sure that available investment funds are spent where they will be of most use. In order to optimise service provision under these changing conditions, policy-makers and planners need to be able to accurately forecast how passenger numbers will respond to changes in the various demand drivers.

At present in the Great Britain, an elasticity-based approach forms the established basis for rail demand forecasting,4 as outlined in the Passenger Demand Forecasting Handbook (PDFH)5 and based on an
extensive body of research, with demand uplift factors calculated based on the proportional change in the level of explanatory variables over time. Because the impact of a change of a given magnitude in one of these explanatory variables may have a different effect on rail demand under different circumstances the PDFH segments the rail passenger market into nine flow types (London Travel Card Area (LTCA); rest of South East to LTCA; LTCA to rest of South East; other South East flows; rest of country to LTCA; LTCA to rest of country; non-London inter-urban; non-London short distance; to and from airports). However, this segmentation assumes that the characteristics of the rail market outside the London and South-East area are spatially homogeneous. Previous work using a technique called geographically weighted regression (GWR) to forecast total trip numbers from railway stations has shown that in reality there are complex spatial variations in the effect of explanatory variables on rail demand. This paper extends the application of this approach from point-based models to flow-based models, which forecast trip volumes between station pairs, to produce a new spatial segmentation which better reflects actual spatial variations in the effects of demand drivers.

**GWR – a brief overview**

GWR is a statistical modelling methodology first developed by Fotheringham et al. which allows spatial variations in the effect of independent variables on the dependent variable to be explicitly incorporated in the form of a regression model. Using the notation listed in Appendix 1, the standard regression model can be written as equation (1) and the equivalent GWR model as equation (2):

$$y_i = \alpha + \sum k \beta_k x_{ik} + \epsilon_i$$  \hspace{1cm} (1)

$$y_i = \alpha(u_i, v_i) + \sum k \beta_k(u_i, v_i)x_{ik} + \epsilon_i$$  \hspace{1cm} (2)

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. In this context this means that when forecasting the number of trips made from a particular railway station based on the values at that point of a number of independent variables, the unique parameter estimates used in this forecast will be influenced more by the observed values of the independent variables at nearby stations than the values at distant stations.

While GWR has previously been applied to projects in a wide range of research areas, these have almost exclusively involved the use of spatial data where each observation has a single point location, as was the case for the previous work on rail demand modelling. The application of GWR to flow-based observations during the research described here therefore represents a departure from the accepted norm, and the issues which arose in the extension of GWR from points to flows are discussed in more detail in the section ‘GWR models’.

All models tested were designed to be compatible with the PDFH elasticity-based modelling framework. An alternative way to forecast rail demand using GWR would be to adopt a more ‘conventional’ four-stage modelling framework using origin coordinates with GWR to forecast trip production, destination coordinates with GWR to forecast trip attraction, and a global distribution model to link productions and attractions. However, as such an approach would not be compatible with the PDFH, it was not tested here.

**Global cross-sectional demand models**

In order to provide a base level of model fit and allow effective comparisons with previous research, the first stage of this project involved the calibration of ‘conventional’ multiple regression models of rail demand based on data for 19,646 origin-destination pairs. Observed trip data came from The Oxera Arup Dataset (TOAD) which is based on ticket sales data, and contains counts of all trips made on all flows included in the database for the period 1991–2008, with 2007–2008 data used here. While a dataset of this size and comprehensive nature allowed a very robust model to be developed, the trade-off was that processing times for GWR model calibration were extremely lengthy (24–36 h per model using a desktop PC). TOAD also provided data on the various model explanatory variables, including population, employment, car ownership, rail generalised journey time and rail fares. Values for non-rail-related variables are based on the tempro zoning system used in the UK Department for Transport’s National Trip End Model.

A range of global multiple regression models were calibrated, with the results summarised in Table 1. Two different representations of origin and destination attractiveness were used, with the first involving the use of total passenger origins/destinations at the origin/destination station as a proxy for origin/destination trip generation/attraction potential (‘Trips’, equation (3)). While this introduces an element of circularity into the model, it is an extremely simple measure which can provide a ‘benchmark’ level of model fit, and is arguably valid for modelling flow-level demand where the likelihood of choosing a particular destination is unlikely to be directly affected by the total trips generated at the origin. The second option was to use a range of variables which described the characteristics of the origin/destination, including catchment population, employment and car
parameters determined during calibration best global regression models using GWR, to allow The next stage in the analysis was to recalibrate the GWR models (3), (4), (5) and (6), respectively. The form of models A, B, C and D is given by equations (3), (4), (5) and (6).

\[ T_{ij} = e^{a_i}O_{ij}^Dj_{ij}GJT_{ij}^A R_{ij}^S C_{ij} \]  
\[ T_{ij} = e^{a_i}H_{ij}^C Pa_{ij}^D j_{ij} H_{ij}^C Da_{ij} GJT_{ij}^A R_{ij}^S C_{ij} \]  
\[ T_{ij} = e^{a_i}O_{ij}^D j_{ij} H_{ij}^C Da_{ij} GJT_{ij}^A R_{ij}^S C_{ij} \]  
\[ T_{ij} = e^{a_i}H_{ij}^C Pa_{ij}^D j_{ij} H_{ij}^C Da_{ij} GJT_{ij}^A R_{ij}^S C_{ij} \]

where \( a, b, \gamma, \delta, \zeta, \kappa, \xi, \rho, t, \varphi, \chi, \psi, \omega \) and \( v \) are fitting parameters determined during calibration.

### GWR models

The next stage in the analysis was to recalibrate the best global regression models using GWR, to allow explicit incorporation of spatial variations in the parameter values into the model form (full details of the GWR model calibration process are provided by Fotheringham et al.\(^\text{12}\)). For this calibration it was necessary to assign a single spatial location to each of the flows, and three potential options were identified for this: the flow origin; the flow destination; and the flow midpoint. There was little difference in the overall spatial distribution of the flow origins and destinations (as most origin/destination flows had a ‘balancing’ return flow), but the flow midpoints exhibited a very different spatial pattern, as shown by Figure 1.

While the flow midpoints give a more diverse geographical spread of observations, they have two major problems when used in this context. First, some midpoints are located ‘offshore’, which is obviously not a realistic location to assign to a flow, and second (for example) a flow from London to Glasgow has a very similar midpoint to a flow from Manchester to Burnley or from Bangor to Scarborough. Assigning the same location to these three flows is conceptually unsatisfactory, because the same set of parameter estimates would represent trips with a wide range of origin and destination characteristics, in contrast with the origin and destination coordinate sets where all data points in a particular area would relate to trips originating or terminating in that area.

A Gaussian model form was used along with adaptive spatial kernels (meaning that the same number of data points were considered for each regression point), kernel bandwidths being determined by Akaike Information Criterion minimisation. The significance of the spatial variation in the model parameters was tested using Monte Carlo simulation. The results from the GWR calibration of models A to F and I are summarised in Table 2, which shows the absolute fit of the model as measured by the adjusted \( R^2 \) value, the \( F \) statistic from an analysis of variance (ANOVA) which tests the null hypothesis that the GWR model represents no improvement over an equivalent global model (if this statistic is greater than two then the GWR model gives a statistically significant improvement over the global model), and the number of parameters where the spatial variation was statistically significant.

The model fit from the global and GWR calibrations is compared in Figure 2. It is clear that the GWR calibration of model A provided a better model fit than the global calibration, and model fit is also slightly superior to that given by the global calibration of the MVA segmented model. The GWR recalibration of the season and non-season ticket models (E and F) again gave an improvement in fit over the global calibration of the same models, but the model fit was still far inferior to that of the models calibrated on the combined dataset. There is relatively little difference between the model fit provided by the three different sets of coordinates, with the midpoint coordinates being marginally superior. The variation in the range of parameter values given by the different sets of coordinates for model A is shown in Figure 3, which shows that in general the variations between coordinate sets are in the extreme

### Table 1. Summarised results from global regression models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Equation</th>
<th>OD representation</th>
<th>Flows</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(3)</td>
<td>Trips</td>
<td>All</td>
<td>0.720</td>
</tr>
<tr>
<td>B</td>
<td>(4)</td>
<td>Characteristics</td>
<td>All</td>
<td>0.729</td>
</tr>
<tr>
<td>C</td>
<td>(5)</td>
<td>O Trips, D Char</td>
<td>All</td>
<td>0.730</td>
</tr>
<tr>
<td>D</td>
<td>(6)</td>
<td>O Characteristics, D Trips</td>
<td>All</td>
<td>0.729</td>
</tr>
<tr>
<td>E</td>
<td>(3)</td>
<td>Trips</td>
<td>Season</td>
<td>0.571</td>
</tr>
<tr>
<td>F</td>
<td>(3)</td>
<td>Trips</td>
<td>Non-Season</td>
<td>0.625</td>
</tr>
<tr>
<td>G</td>
<td>(4)</td>
<td>Characteristics</td>
<td>Season</td>
<td>0.563</td>
</tr>
<tr>
<td>H</td>
<td>(4)</td>
<td>Characteristics</td>
<td>Non-Season</td>
<td>0.648</td>
</tr>
<tr>
<td>I</td>
<td>(3)</td>
<td>Trips</td>
<td>MVA</td>
<td>0.798</td>
</tr>
<tr>
<td>J</td>
<td>(4)</td>
<td>Characteristics</td>
<td>MVA</td>
<td>0.783</td>
</tr>
</tbody>
</table>

\( O = \) origin, \( D = \) destination.
values of the parameters with the central range of values being reasonably consistent across the coordinate sets. Because of the previously described problems relating to the midpoint coordinates, these results were discounted. There were no clear grounds for choosing between the origin and destination coordinates, and the spatial variations in parameter values given by the two coordinate sets appeared very similar in most cases, as shown by Figure 4 which maps the variation in the generalised journey time (GJT) parameter. There are a small number of ‘hot’ and ‘cold’ spots visible on both maps, indicating points where a particular observation or cluster of observations differ markedly from surrounding observations.

A single set of spatial locations was used for the GWR calibration of the models incorporating detailed origin/destination characteristics, because it was clear that where variables giving a detailed description of origin characteristics were included in the model the origin coordinates were most appropriate, and vice versa. Once again, the use of GWR gave a clear improvement in model fit over the global models, although model fit for models C and D was slightly inferior to that for model A. The parameter values for the origin/destination specific variables in general exhibited a much higher degree of spatial variability than the more general variables, as shown in Figure 5. While the GWR calibration of the segmented model I using origin coordinates gave an improvement in fit over the global calibration, when the spatial variations were mapped this appeared to be heavily influenced by the definition of the dummy variables (and thus by the prior segmentation of the dataset). This adds weight to a prior concern about whether it is valid to use GWR with a model

Table 2. Summarised results from GWR models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Flows</th>
<th>Coordinate set</th>
<th>$R_{adj}^2$</th>
<th>$F$ stat</th>
<th>Spatial parameter variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>All</td>
<td>Origin</td>
<td>0.798</td>
<td>27.481</td>
<td>6 of 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Destination</td>
<td>0.812</td>
<td>32.765</td>
<td>6 of 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Midpoint</td>
<td>0.816</td>
<td>33.269</td>
<td>6 of 6</td>
</tr>
<tr>
<td>C</td>
<td>All</td>
<td>Destination</td>
<td>0.797</td>
<td>17.973</td>
<td>9 of 9</td>
</tr>
<tr>
<td>D</td>
<td>All</td>
<td>Origin</td>
<td>0.787</td>
<td>14.656</td>
<td>9 of 9</td>
</tr>
<tr>
<td>E</td>
<td>Season</td>
<td>Origin</td>
<td>0.681</td>
<td>15.330</td>
<td>5 of 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Destination</td>
<td>0.677</td>
<td>15.571</td>
<td>6 of 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Midpoint</td>
<td>0.699</td>
<td>15.000</td>
<td>6 of 6</td>
</tr>
<tr>
<td>F</td>
<td>Non-Season</td>
<td>Origin</td>
<td>0.747</td>
<td>32.773</td>
<td>6 of 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Destination</td>
<td>0.768</td>
<td>40.864</td>
<td>6 of 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Midpoint</td>
<td>0.769</td>
<td>40.023</td>
<td>6 of 6</td>
</tr>
<tr>
<td>I</td>
<td>All</td>
<td>Origin</td>
<td>0.815</td>
<td>15.256</td>
<td>6 of 15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Destination</td>
<td>0.789</td>
<td>32.455</td>
<td>15 of 15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Midpoint</td>
<td>0.773</td>
<td>49.536</td>
<td>15 of 15</td>
</tr>
</tbody>
</table>

Figure 1. Spatial distribution of flow origins and flow midpoints.
containing such explicitly spatial dummy variables, as it seems likely that a form of multicollinearity may influence the model results.

While it is not possible to offer convincing explanations for many of the variations illustrated in these maps, this does not prevent the variations being used to develop possible new segmentations of the rail market. It is clear that the significant spatial variations in parameter values identified by the GWR models cast doubt upon the validity of the current spatial segmentation of demand in the PDFH. However, the lack of consistency in patterns of spatial variation across parameters makes the definition of a spatial segmentation of the rail demand market more complicated, as it suggests that different spatial segmentations may be required for different variables, and thus could preclude the current PDFH approach based on a single set of zonal dummy variables.

**Application to market segmentation and demand forecasting**

The GWR model can be used as calibrated to produce spatially varying demand forecasts by incorporating spatial matrices of parameter values in a bespoke spreadsheet (as in Blainey11). However, in order to incorporate the GWR model results into the UK industry standard forecasting methodology, the PDFH,5 it is necessary to generalise the GWR results into a spatial segmentation. Two potential approaches were identified for this procedure.

1. Parameter-specific segmentation: apply different spatial segmentations to different parameters, average the GWR parameter values across each segment, and then use these average values to forecast demand.

2. Single spatial segmentation: apply a single spatial segmentation across all parameters, and then recalibrate the global regression models either separately for each segment, or using dummy variables to represent all spatial segments within a single model.

The first stage in both cases was to derive a segmentation from the GWR parameter values. While this could in theory be generated automatically using spatial clustering, no readily implementable methodology was available. Visual differentiation was therefore used, based on individual point parameter values, rather than the interpolated values shown in Figures 3 and 4, as this allowed turning points in the parameter values to be more easily observed.
Using a parameter-specific segmentation has some obvious advantages, most notably that it allows different variables to have different spatial segmentations, and this was tested using the origin coordinate version of model A. All flow origins were assigned to the relevant segment for each parameter, the GWR parameter values for all flow origins within each segment were averaged, and these average parameter values were substituted into equation (3) to produce a fresh set of demand predictions. However, while several different segmentations were tested, these new predictions gave a best $R^2$ value of only 0.486, far inferior to the model fit from both the global model and the full GWR model, suggesting that this method of generating spatial segments from the GWR results is not appropriate. An alternative

**Figure 4.** Spatial variation in GJT parameter from GWR calibration of model A.

**Figure 5.** Spatial variation in households and car ownership parameters from model D.
methodology for incorporating parameter-specific segmentations in the model form is to assign a separate set of dummy variables to each independent variable, giving model K (shown here by equation (7)), and then recalibrate the model using conventional multiple regression methods. The obvious disadvantage of this approach is that it greatly increases the number of model parameters (to 57 with the test segmentations used here), and while calibration of this model gave an adjusted $R^2$ value of 0.751, 25 of the dummy variable parameter values were not statistically significant.

$$T_{ij} = e^\beta \left( \prod_{a} n S_{a}^{T} \right) \left( \prod_{a} S_{aa}^{T} \right) \left( \prod_{a} S_{aa}^{\text{read}} \right) \left( \prod_{a} S_{ad}^{T} \right) \left( \prod_{a} S_{ad}^{\text{read}} \right) \times GJT_{ij} \left( \prod_{a} n S_{ag}^{T} \right) \left( \prod_{a} S_{ag}^{\text{read}} \right) \left( \prod_{a} n S_{ag}^{\text{read}} \right)$$

(7)

The other approach tested was to apply a single spatial segmentation to all parameters, and then recalibrate the model based on this segmentation, either using dummy variables to represent the segments or calibrating a separate model for each segment. However, because the spatial variation in parameter values differs between the model parameters, in order to spatially segment the dataset it was first necessary to aggregate this variation. This was achieved by first normalising the local values for all parameters using equation (8), and then summing these normalised values to give a single value for each flow, representing the total magnitude of variation from the mean. Two approaches were tested for summing the normalised values, using first the absolute level of variation from the mean (i.e. with the sign of the individual normalised variations ignored), and second allowing positive and negative variations to cancel each other out. The latter approach was found to give slightly superior results as measured by the fit of the final model, and the spatial segmentation derived based on this aggregated variation is shown in Figure 6.

Another option for aggregating spatial variation might be to use a sum of squares-based measure, but time constraints meant that this was not tested during this project

$$z_i = \frac{x_i - \mu}{\sigma}$$

(8)

As previously stated, two approaches were tested for model recalibration using this single segmentation. The first added segment dummy variables to model A, giving model L (shown here by equation (9)). While the fit of this model when calibrated on the full dataset was slightly superior to that of the global calibration of model A ($R_{adj}^2 = 0.741$), a major disadvantage of this approach is that the spatial variation incorporated in the model relates only to segment dummy variables rather than to the key explanatory variables, and is therefore conceptually very different to the spatial variation captured by GWR

$$T_{ij} = e^\beta \left( \prod_{a} D_{ij}^{T} \right) \left( \prod_{a} n S_{ag}^{T} \right) \left( \prod_{a} S_{ag}^{\text{read}} \right) \left( \prod_{a} n S_{ag}^{\text{read}} \right)$$

(9)

Better results were given by recalibrating model A separately for each of the 12 segments. This has the advantage of allowing the values of all model parameters to vary between segments, as illustrated in Figure 7, which also shows that the segments derived from the GWR models are very different to those used either in the current PDFH segmentation or in other quasi-spatial segmentations produced by MVA and by Arup and Oxera. The model fit ($R_{adj}^2$) was still inferior to the full GWR model but was better than that from the global calibration of model A, from model F and from model L, indicating that this is the best methodology of those tested for generalising the results from the GWR model. This methodology was therefore also used to generalise the results from the GWR calibration of model D, giving an overall $R^2$ value of 0.787, the same as that given by the full GWR model, indicating that the segmentation explains an equivalent proportion of the variation in demand levels. For both models the patterns of spatial variation across the segments differed widely between

![Figure 6. Segmentation derived from summed variation in normalised parameter values from mean.](image)
parameters, both in terms of the magnitude of the variation and the patterns exhibited (as shown in Figure 7). This finding confirms that such a spatial segmentation of the UK passenger rail market is appropriate (although not necessarily optimal) given that models calibrated on neighbouring segments generate very different parameter values. While the extreme values given for the rail fare parameter seem somewhat implausible, these only apply to a very small number of stations.

In order to illustrate the potential change in the magnitude of demand forecasts given by the different segmentations, the demand impact of a 5% reduction in generalised journey time has been predicted for three different flows using the current PDFH elasticities, and elasticities derived in this work from Model I (expanded MVA segmentation, the full GWR calibration of model A, and the GWR-based segmented calibration of model A). The results of these forecasts are given in Table 3.

Table 3 shows that there is a great deal of variation in the elasticity values from the different segmentations, with the variation between flows much greater for the three new segmentations than for the PDFH segmentation. This is reflected in the values of the change index, which shows that the choice of spatial segmentation can have a noticeable impact on the resulting demand forecasts, with over twice as much additional demand predicted by the new segmentations for flows 2 and 3. The elasticity values for the new segmentations are in all cases much larger than those from the PDFH (for example, for flow 3 the GWR segmentation has a GJT elasticity of –1.878 compared with a PDFH elasticity of –0.7).

It is likely that this results at least partly from the

---

**Figure 7.** GJT and rail fare parameter values from segmented calibration of model A.

**Table 3.** Forecasts of demand changes resulting from GJT reduction.

<table>
<thead>
<tr>
<th>Flow number</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base trips</td>
<td>1045</td>
<td>28,768</td>
<td>13,915</td>
</tr>
<tr>
<td>Base GJT</td>
<td>680.1</td>
<td>60.5</td>
<td>51.6</td>
</tr>
<tr>
<td>New GJT</td>
<td>646.1</td>
<td>57.5</td>
<td>49.0</td>
</tr>
<tr>
<td>GJT elasticity PDFH</td>
<td>–1.245</td>
<td>–1.663</td>
<td>–1.797</td>
</tr>
<tr>
<td></td>
<td>Expanded MVA</td>
<td>–2.916</td>
<td>–1.648</td>
</tr>
<tr>
<td></td>
<td>GWR segmentation</td>
<td>–1.292</td>
<td>–1.692</td>
</tr>
<tr>
<td>Change index PDFH</td>
<td>1.037</td>
<td>1.047</td>
<td>1.037</td>
</tr>
<tr>
<td></td>
<td>Expanded MVA</td>
<td>1.066</td>
<td>1.089</td>
</tr>
<tr>
<td></td>
<td>Full GWR</td>
<td>1.161</td>
<td>1.088</td>
</tr>
<tr>
<td></td>
<td>GWR segmentation</td>
<td>1.069</td>
<td>1.091</td>
</tr>
<tr>
<td>Predicted trips PDFH</td>
<td>1,083</td>
<td>30,127</td>
<td>14,424</td>
</tr>
<tr>
<td></td>
<td>Expanded MVA</td>
<td>1,114</td>
<td>31,330</td>
</tr>
<tr>
<td></td>
<td>Full GWR</td>
<td>1,214</td>
<td>31,306</td>
</tr>
<tr>
<td></td>
<td>GWR segmentation</td>
<td>1,117</td>
<td>31,376</td>
</tr>
</tbody>
</table>
calibration of these models on cross-sectional demand data, in contrast with the panel and time series data used to derive the PDFH elasticities, and this means that the results are not directly comparable. Cross-sectional models are affected by problems of simultaneity\(^\text{15}\) where the direction of causation is not clear, as without time series data it is not possible to tell whether, for example, low train frequencies result in low usage levels, or whether low usage results in a low-frequency train service being provided.

Looking forward: GWR and panel data

From the results and discussion in the previous section it is clear that the priority in extending this research must be to test the calibration of GWR models on panel data, to give direct comparability between the results from the GWR modelling and the elasticities and segmentations currently contained in the PDFH. Standard GWR is not suitable for use with panel data, and the use of GWR with panel data is not mentioned in the main GWR reference text\(^\text{12}\), however, a literature search identified a small amount of previous research in this area. One option is to undertake a separate calibration of the model for each time step, and in each calibration to apply a temporal kernel in addition to the spatial kernel with observations closest in time to the regression point having the greatest influence on the model\(^\text{16}\). Alternatively, GWR can be combined with more conventional panel data techniques, by first using a spatial kernel to weight all data points around each regression point (as in cross-sectional GWR), and then applying fixed or random effects panel analysis models to the weighted subsample to obtain local parameter estimates at each location\(^\text{17}\). The latter technique appears more suitable for use in the rail demand modelling context, but neither is supported by the GWR 3.0 software used for model calibration in this project. The use of flow-specific dummies to act as proxies for fixed effects would also not be an option in this case, as the software has a limit of 30 explanatory variables. Bespoke model code would need to be written to implement any of these techniques, and the complexity of the GWR calibration process means this is not an easy task.

An alternative possibility would be to adjust the form of the model so that the fixed effects were accounted for prior to model calibration, meaning that the model was effectively predicting changes in normalised demand levels and using dummy variables to represent each year. Application of this methodology to model A would mean conducting a GWR calibration of a model of the form shown here by equation (10). Because the model is only predicting the change in demand rather than absolute demand there is no need to include the origin and destination total trips variables (which had previously formed a proxy for origin/destination attractiveness), as these would now merely add an element of circularity to the model. A restriction of this model is that it would only be suitable for a maximum of 26 time periods, due to the software limit on the number of explanatory variables, but as rail demand data is currently only available for 18 years this would not be an issue in practice in the Great Britain, and the calibration of this model would seem an obvious starting point for any future work in this area.

\[
\frac{T_{ijt}}{T_{ijt-1}} = e^\left(\frac{GJT_{ijt}}{GJT_{ijt-1}}\right) \left(\frac{Rf_{ijt}}{Rf_{ijt-1}}\right)^{\xi} \left(\frac{Ct_{ijt}}{Ct_{ijt-1}}\right) \left(\prod_t Y_t\right)
\]

(10)

If data became available it also would be desirable to add further independent variables to the models, for example, on station access and station catchment characteristics, although it should be noted that these areas are extremely complex\(^\text{18}\) and it can therefore be very difficult to produce generalised variables. An alternative might be to examine the correlations between parameter estimates and variables such as the density of population and of road and rail infrastructure. The use of alternative segmentations, for example, based on train operating companies (TOCs), or of flow-type dummy variables (such as intercity or local) could also potentially give interesting results.

Conclusions

The analysis undertaken during this project has shown that GWR has the potential to enhance the accuracy of flow level models of rail demand in the Great Britain. Using GWR to calibrate such models gives a clear improvement in model fit over that obtained by using conventional (global) multiple regression methods. Significant spatial variation was found in the effect of various explanatory variables on rail demand, but this does not mean that space (or place) is in itself a determinant of rail usage levels. It is more likely that it acts as a proxy for variables which cannot easily be incorporated in demand models, such as habit, inertia, the level of diffusion of the rail product and local variations in intermodal competition.

The output from GWR models can be used to directly produce demand forecasts if incorporated in a bespoke spreadsheet, but this may not be appropriate in all circumstances (and is not compatible with the current British demand forecasting framework\(^\text{5}\)). An alternative approach has therefore been tested here which approximates the results from the GWR modelling by spatially segmenting the study area based on the variations in parameter values from the GWR models. Conventional regression models are then recalibrated separately for each segment, giving a set of segment-specific parameter values, and these have been shown to give an improvement in model fit over the single original multiple regression model.
While this research has used Great Britain as a case study area, there seems no reason why the general techniques and modelling approach used would not be suitable for use in forecasting rail demand levels in other countries. For example, as an extension of the work described here GWR rail demand models are currently being developed and applied to the local rail network in New South Wales, Australia, and similar applications could be possible in many other locations around the world. Given the likelihood that future needs to balance expanding urban transport demand against environmental constraints will mean a further expansion in local rail networks, accurate demand forecasting using techniques such as GWR should continue to play an important role in transport planning.

Funding
The research described in this paper was funded by the Passenger Demand Forecasting Council (PDFC) of the Association of Train Operating Companies, UK.

Acknowledgments
The authors gratefully acknowledge the helpful comments received from members of the PDFC project steering committee and from attendees at PDFC meetings and at the 45th Universities Transport Study Group Conference. This paper contains Ordnance Survey data ©Crown copyright and database right 2013.

References

Appendix 1
Notation
\[ CO_i \ (CO_j) \] the total number of cars owned by households in the Tempro zone containing station \( i \) \((j)\) divided by the number of households in this zone
\[ CT_{ij} \] estimated car journey time (in minutes) from station \( i \) to station \( j \) in 2007–2008 calculated using ArcGIS based on Ordnance Survey data
\[ D_j \] the total number of trip destinations from station \( j \) in 2007–2008 as recorded in the Office of Rail Regulation (ORR) station usage data
\[ GJT_{ij} \] the generalised rail journey time (in minutes) from station \( i \) to station \( j \) in 2007–2008
\[ H_i \ (H_j) \] the total number of households in the Tempro zone containing station \( i \) \((j)\)
\[ J_i \ (J_j) \] the total number of jobs in the Tempro zone containing station \( i \) \((j)\)
\[ O_i \] the total number of trip origins from station \( i \) in 2007–2008 as recorded in the ORR station usage data
\[ PA_i \ (PA_j) \] the number of car parking spaces provided at station \( i \) \((j)\)
the average rail fare for the journey (in £) from station $i$ to station $j$ in 2007–2008

dummy variable which takes the value 1 if flow $ij$ originates in area $a$ and 0 otherwise

dummy variable which takes the value $e^1$ if flow $ij$ falls within car journey time segment $s$, and $e^0$ otherwise

dummy variable which takes the value $e^1$ if flow $ij$ falls within destination segment $s$, and $e^0$ otherwise

dummy variable which takes the value $e^1$ if flow $ij$ falls within rail fare segment $s$, and $e^0$ otherwise

dummy variable which takes the value $e^1$ if flow $ij$ falls within GJT segment $s$, and $e^0$ otherwise

dummy variable which takes the value $e^1$ if flow $ij$ falls within intercept segment $s$, and $e^0$ otherwise

dummy variable which takes the value $e^1$ if flow $ij$ falls within origin segment $s$, and $e^0$ otherwise

the current time period

the preceding time period

$T_{ij}$ the number of trips from station $i$ to station $j$ in 2007–2008

$u_i$ the $x$-coordinate of data point $i$

$v_i$ the $y$-coordinate of data point $i$

$x_i$ the observed value of the variable for observation $i$

$x_{ik}$ the value of the $k$th explanatory variable at data point $i$

$y_i$ the value of the model dependent variable at data point $i$

$y_y$ a dummy variable which takes the value $e^1$ if time period $t$ is year $y$ and 0 otherwise

$z_i$ the normalised value of the variable for observation $i$

$\alpha$ a parameter determined during model calibration

$\beta$ a parameter determined during model calibration

$\varepsilon_i$ the model error term for data point $i$

$\mu$ the mean of the observed values for the variable

$\sigma$ the standard deviation of the observed values for the variable