THE VEHICLE KILOMETRES TRAVELLED (VKT) BY PRIVATE CAR: A SPATIAL ANALYSIS USING GEOGRAPHICALLY WEIGHTED REGRESSION

CORINNE MULLEY[1] AND MICHAEL TANNER[2]

ABSTRACT

The existing Sydney VKT regression model uses an ordinary least squares (OLS) approach for modelling with the intention of providing the 'back-office' to a future web-based planning tool that would allow Household VKT to be predicted on the basis of to land-use and other factors.

The paper replicates the existing Sydney VKT model to demonstrate that taking account of spatial variability in a local model estimation, using the Geographically Weighted Regression (GWR) methodology, enhances the explanation of household VKT and improves the fit of the data to the model This suggests that GWR would be a superior alternative for the web-based tool. The paper explores the implications of the local estimation thereby showing the depth of information that can be gleaned from local estimation as well as identifying a number of steps to improve the model's theoretical base and performance in the future.

The paper establishes that travel behaviour attributes have a spatial context and this has wide reaching implications for other analysis within the transport sector. Moreover, GWR, as a methodology, has the ability to perform spatial analysis with the added advantage of allowing the visual interpretation of parameter results based on geography

1. INTRODUCTION

The existing Sydney VKT regression model, as described by Corpuz et al (2006) uses an ordinary least squares (OLS) approach for modelling. The purpose of the model was to provide a base to a future web-based planning tool that would allow Household VKT to be predicted on the basis of changes principally to land-use but also other factors, such as levels of car ownership and accessibility to public transport. The difficulties of using household level data as explanatory variables are well described in Corpuz et al (2006).

The motivation for this paper was a belief that to properly understand the relationship between household VKT and potential explanatory variables, it is necessary to deal with spatial data more specifically. In particular, OLS methodology assumes that the observations in the regression are independent of one another and this is unlikely to be the case when using data likely to exhibit spatial autocorrelation (correlation of a variable with itself through space). OLS therefore looks for similarities in different spatial areas are effectively concentrated in an 'average' figure to cover the whole space.

This paper is based on the hypothesis that there may be different causal impacts on household VKT, depending on the locality under consideration and uses the methodology of Geographically Weighted Regression (GWR) to take account of spatial autocorrelation by adding a co-ordinate to each observation in the data set which then allows household VKT, as the dependent variable, to be explained by a number of spatially defined factors. The paper explores whether allowing spatial variability to be included in the model enhances the explanation of the dependent variable and improves the fit of the data to the model. As a basis for a web-based tool, GWR could instead be used as the 'back-office' modelling technique if it were to show superior predictive qualities.

[1] Chair in Public Transport, Institute of Transport and Logistics Studies, University of Sydney, NSW

[2]Manager, Spatial Information, Transport Data Centre, Ministry of Transport, NSW

This paper aims to demonstrate that looking for causal explanations of travel behaviour, in this case household VKT, need to be considered spatially. For transport activities more generally, it lays the foundation for spatial attributes to be considered more specifically in the modelling process.

The paper is organised as follows. The next section describes the methodology of GWR and contrasts this with OLS. Section 3 looks and compares the results for the two different models before identifying the next steps in Section 4. The final section, Section 5 concludes.

2. METHODOLOGY

The existing model, as described by Corpuz et al (2006) was the result of empirical investigation to identify explanatory variables giving the best fit to the data of household VKT in the Sydney area. This paper does not repeat this investigation but takes the existing model as a starting point to demonstrate the opportunities offered by spatial analysis in the transport domain.

2.1 The existing household VKT model

The model used looks at household VKT as a function of broad socio-economic, transport and land-use variables as follows:

$$VKT = f(V, A, L, E) \quad (1)$$

where V relates to vehicle ownership, A to the transport accessibility of the household to high quality public transport, L to the land-use mix and density of building in the area of the household and E to the proximity of employment possibilities for the household. These variables were chosen by reference to the literature as described in Corpuz et al (2006) and by a need to avoid multicollinearity.

The data used to represent the variables in equation (1) are as shown in Table 1 below.

Table 1: Variables: their definition and measurement

Data	Variable Name	Description
Number of vehicles in the household	HH_vehicle	Average number of vehicles in the household for the Collection District (CD) in which the household is located
Accessibility	KmCBDC	The shorter of the two road distances, measured in km, from the centroid of the Collection District (CD) to the nearest centre or CBD
	Access_pt	Walk + wait time in minutes to the nearest high frequency bus, train, light rail or ferry from the centroid of the Collection District (CD). Walking time estimated at 15m/km and wait time as 0.5 of the frequency.
Land use	Landuse_mix	A weighted measure based on the Local environment plan (LEP) that considers the proportion of different land-use types within 1km of the centroid of the Collection District (CD) as a means of standardising for different land uses
	HHDens	Net housing density in terms of dwellings per hectare (residential and business) excluding green space within 2Km of the Collection District (CD) centroid.
Employment opportunities	Employment	Number of jobs (measured in '000) within 5km of the centroid of the Collection District (CD).

The data used in the development of the model was the Transport Data Centre (TDC) Household Travel Survey (HTS). Seven years of survey data were compiled (June 1997 to June 2004): and included over 16,000 household's information from the travel diary undertaken on the

day of survey, together with socio-economic information. Additional information relating to employment and household density was added from the Australian Bureau of Statistics' 2001 Census and the Department of Planning Local Environment Plan (LEP) provided information on land use. Accessibility to public transport was based on the most up to date information at the time of the analysis (2006). The data, whilst collected at the household level with associated statistics identified at the Collection District (CD) level was aggregated to the travel zone (TZ) geography by appropriately averaging the number of household observations in each TZ. There are 872 TZs in the study area. The dependent variable was transformed to its square root to remove the heteroscedastic (non-constant variance) nature of the error term. Thus the model estimated was as follows:

$$\sqrt{\text{VKT}} = \alpha + \beta_1 \text{HH\_vehicle} + \beta_2 \text{KmCBDC} + \beta_3 \text{Access\_pt} + \beta_4 \text{Landuse\_mix} + \beta_5 \text{HHdensity} + \beta_6 \text{Employment} \qquad (2)$$

The results of this model are shown below in the Section 3.1 and are referred to as the Global regression model.

2.2 Geographically weighted regression (GWR)

The traditional multiple regression OLS model assumes that the relationship to be modelled holds everywhere uniformly in the study area under consideration. The type of geographical data used as explanatory variables in this model are likely to give rise to spatial effects. Spatial effects may occur in two different forms: one is concerned with spatial dependency, or its weaker expression, spatial autocorrelation (they are not identical though they are often used interchangeably in the literature) and the other form is spatial heterogeneity, namely spatial non-stationarity (Anselin 1999). Spatial autocorrelation can be seen as spatial interaction whilst spatial heterogeneity (spatial non-stationarity) refers to spatial structure (Anselin 1999). Spatial dependency and spatial non-stationarity have been the major challenges in spatial data analysis (Fotheringham et al. 2002). One of the advantages of GWR is that tackles both spatial non-stationarity by accounting for coordinates in parameter estimates, but also spatial dependency by taking into account of geographical location in the intercepts.

In the context of modelling household VKT, part of the spatial variation will be due to differences in attitudes, or preferences of the different households which are distributed over space. Other spatial variation will arise from the different effects of administrative or political boundaries on, for example, land use mix.

GWR, developed by Fotheringham et al (2002) is a relatively new technique for spatial data analysis and has been applied in the transport sector for the analysis of land value uplift (Du and Mulley (2007) and in trip end analysis of transport demand for rail (Blainey (forthcoming)).

The extension provided by GWR is demonstrated in the context of a traditional cross-sectional regression model, despite the fact that the data used in the analysis covered by this paper is pooled cross-section/time series data. The cross section model, as described by Fotheringham et al (2002) can be written as

$$Y_i = \beta_0 + \sum_k \beta_k \beta_{ik} + \varepsilon_i \qquad (3)$$

to a model in which local variations in the parameter values can be revealed by taking into account of coordinates of the variable. If the dependent variable has coordinates (u_i,v_i), the model expressed in (3) above can be rewritten as:

$$Y_i (u_i v_i) = \beta_0 (u_i v_i) + \sum_k \beta_k (u_i v_i) \beta_{ik} + \varepsilon_i \qquad (4)$$

This can be then fitted using a weighted least squares method to give an estimate of the parameters at the location (u_i, v_i) and a predicted value of y where the weights are chosen so that observations near the point in space where the parameter estimates are chosen, (u_i, v_i), have more influence on the result than observations further away. By this geographically weighted

calibration, estimates of the parameters can be made for each data point with coordinates, which then can be mapped. This GWR process provides the unique advantage over some other spatial methods since each observation is treated as an individual observation as opposed to lying within a particular boundary (for example within a particular political boundary) as required by other methodologies used in spatial analysis (for example, multi-level modelling).

Estimation is a trade-off between efficiency and bias in the estimators. The weighting process is undertaken by the use of spatial kernels which capture the data points to be regressed by moving the regression point across the region. The weight is measured through the bandwidth against the distance from data point j to regression point i. Regression results are very sensitive to the choice of bandwidth and the GWR software allows the choice of adaptive spatial kernels to be used so that the bandwidth is narrow when data are dense but that the bandwidth is allowed to spread where data are sparse. The Akaike Information Criterion (AIC) is a useful measure for evaluating the explanation given by GWR models as, in addition to measuring goodness of fit, it takes the complexity of the model into account. A rule of thumb is if the AIC of two models differ by more than 3 then they are statistically significantly different with the lower AIC suggesting a better fit (Fotheringham et al (2002)). If adaptive kernels are used in the estimation process, the GWR software chooses bandwidths so as to minimise AIC. This is explained in more detail in Fotheringham et al (2002).

The GWR methodology requires a spatial location in terms of a Cartesian or Geographic coordinate. The TDC uses a Cartesian system based on The Geocentric Datum of Australia (GDA) and use the map grid of Australia Zone 56 – a Universal Transverse Mercator projection, using the GRS80 ellipsoid. For this paper, all data was treated as if located at the centroid of the TZ with the easting and northings of the centroid being generated using ArcGIS [3] software using population data from the Australia Bureau of Statistics 2001 CD and land use classified under the LEP.

3. RESULTS

3.1 The global model estimation

The results of the global regression model are shown in Table 2 below. This is exactly the same results as the existing model estimated by Corpuz et al (2006), demonstrating that GWR estimation of the global model is equivalent to OLS method. The adjusted R^2 is 0.728, suggesting that 72.8% of the variance in the household VKT is explained. The AIC for this model is 2999.

Table 2: Estimated parameters for the Global regression model

Parameter	Estimate	t-statistic
Intercept	3.940	15.28
HH_vehicle	2.450	22.56
KmCBDC	0.012	1.64
Access_pt	0.008	7.35
Landuse_mix	-1.804	-5.20
HHDens	-0.01 0	-3.66
Employments	-0.002	-3.55

As the t-statistics in the final column show, all parameters except KmCBDC are statistically different from zero at the 5% level of significance and KmCBDC is relatively highly correlated with Access_pt (at 0.64). The insignificance of the parameter estimate might be due to multicollinearity between KmCBDC and Access pt (a correlation between these two variables of 0.63). It can be seen from these results that the dominant driver of household VKT is the

[3] ArcGIS is a suite consisting of a group of geographic information system (GIS) software products produced by Environmental Systems Research Institute.

average number of cars in a household: an increase in the average number of cars in a household by one leading to just over 6 km[4] per day increase in household VKT. At the mean household VKT for the sample, this would represent an increase of approximately 11%. Increasing walk and wait time to high quality public transport increases household VKT (an increase of one minute leads to a 0.001 increase in household VKT per day whilst increasing density of land use and employment serves to decrease household VKT.

The traditional method of investigating spatial model errors is to map the residuals of the regression. For this model, larger residuals (both positive and negative) are associated with outer areas of Sydney where residential density is low, as compared to more central locations. In the inner rings, closer to the Sydney CBD, there are also large positive and negative residual and these too are associated with TZ areas which have low residential activity (eg hospitals, university areas). The non-random nature of the residuals justifies the use of a spatial modelling technique for use with this data.

3.2 The local model estimation

One of the advantages of the GWR model is the ability to examine the spatial variability of the independent variables. Some independent variables therefore might be non-significant at the 5% level in the global regression model which hides the spatial variability. Independent variables might vary significantly over the geographical area and be revealed as significant local parameters by the GWR modelling.

The GWR software provides diagnostic information to assess whether the local model is an improvement over the global model described above. In this model, the local model benefits from a higher adjusted coefficient of determination (adjusted R^2) at 0.785. In addition, the Akaike Information Criterion (AIC) is also statistically significantly lower at 2821 implying that the GWR local model gives a better explanation, after taking the complexity of the model into account. A further diagnostic is provided by the Monte Carlo simulation, provided within the GWR software, which tests whether the geographical variation in the local parameters is significant: in the case of this local model, the geographical variation of all variables except KmCBD and Access_pt is statistically significant at the 5% level (or better).

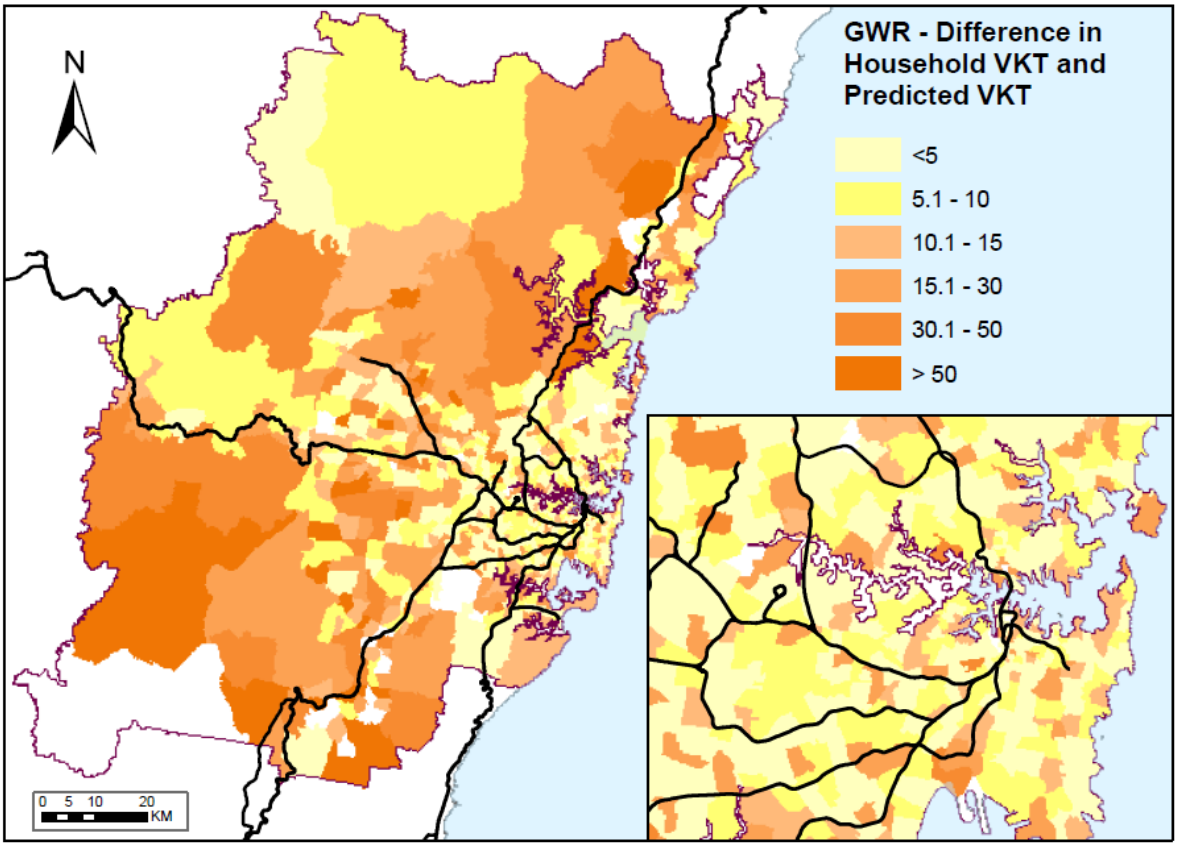
3.2.1 Actual and Predicted Household VKT

The motivation behind the original model was to be able to use it to predict the effect on household VKT following some change in the explanatory variables. It is instructive, therefore, to look at the ability of the local GWR estimation's performance in this context. Figure 1 shows the difference in performance between the global (or OLS) regression model and the local model by looking at the absolute difference between the observed household VKT and the household VKT predicted by the respective models. It can be seen that the absolute error of the model appears to be less using GWR.

The practical difference offered by GWR is a separate equation being estimated for each travel zone giving 872 separate regressions in this case. So for example, taking two travel zones such as Vauclause in the eastern suburbs of Sydney (zone 104) and Macmasters Beach in the Central Coast (zone 1696), Table 3 shows how the estimates differ in each TZ as well as being different from the global (OLS) parameter estimates reported in Table 3.

[4] The dependent variable is the square root of VKT. For the interpretation of the regression, the effect of changes in the explanatory variable are shown on household VKT, rather than its square root. Hence the parameter estimate of 2.46 is squared to become just over 6 in the interpretation of the impact of an increase

Figure 1: A comparison of the error in prediction of the global (OLS) model and the GWR local model for household VKT.



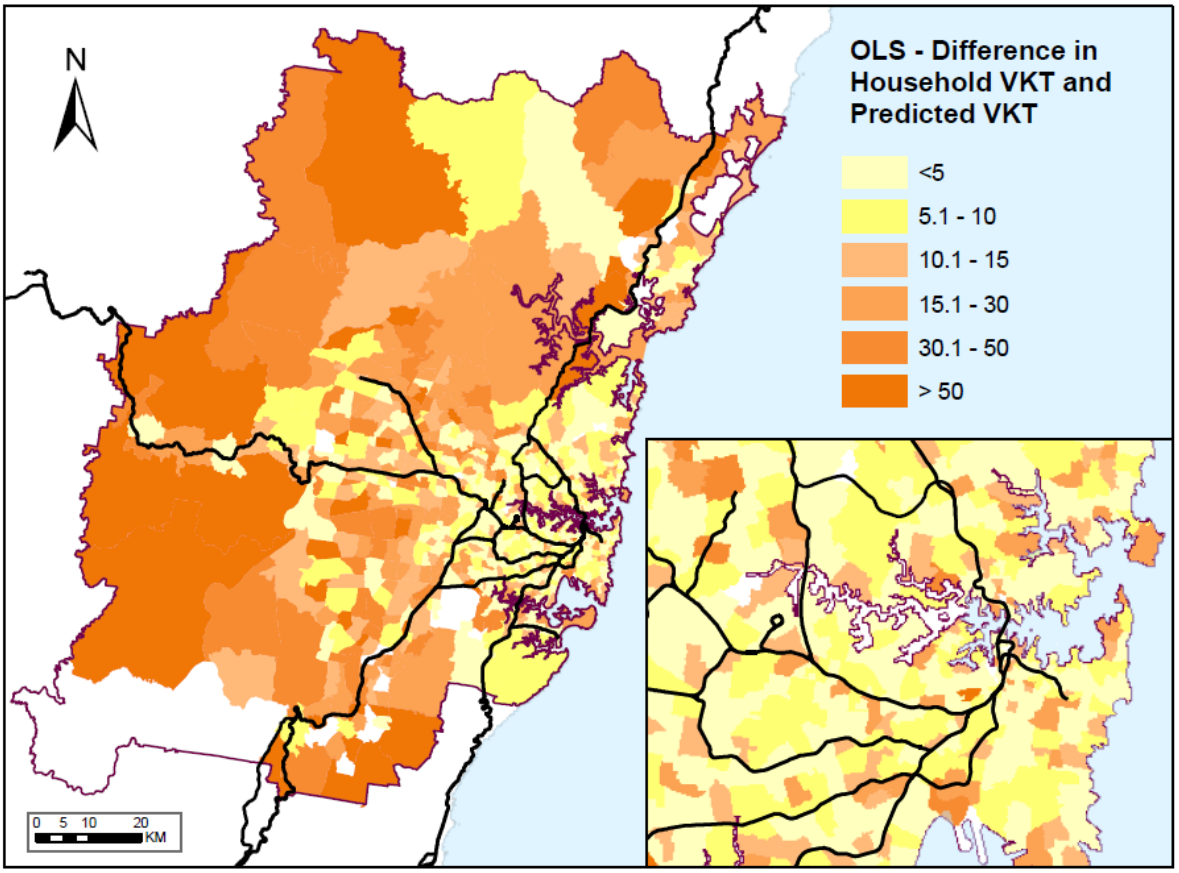


Table 3: Estimated parameters for the Global regression model, and GWR example results for zones 104 and 1696

Parameter	Global (OLS) estimate	Zone 104 estimate	Zone 1696 estimate
Intercept	3.940	6.790	6.790
HH_vehicle	2.450	1.2340	2.804
KmCBDC	0.012	-0.023	0.017
Access_pt	0.008	0.009	0.011
Landuse_mix	-1.804	-2.437	-0.366
HHDens	-0.01 0	-0.022	-0.006
Employments	-0.002	-0.004	-0.001

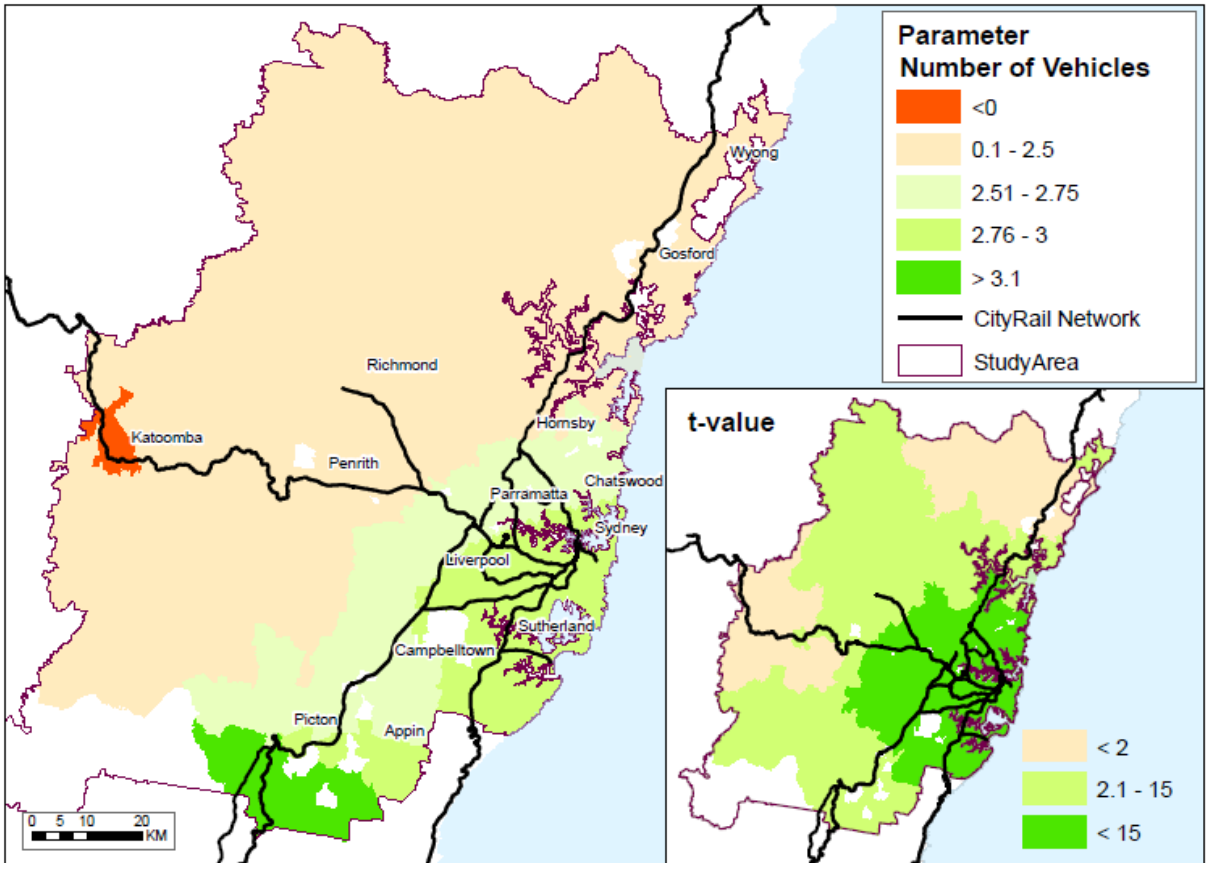
3.2.2 Interpretation of the local GWR model

The interpretation of the local GWR model is conveniently undertaken through mapping the local parameters, in this case at the centroid of the TZ. The value of the parameter, estimated at this centroid, is assumed to hold throughout the zone.

As a result of the GWR methodology, there are different values for the parameter estimates over space. In exactly the same way as it is normal to assess whether a global estimate is statistically different from zero, local estimators like-wise have to be assessed for their statistical significance.

In this section, the interpretation of the local parameter for the average number of vehicles, depicted in is given in detail prior to commenting on the information shown by the other local parameters.

Figure 2: Parameter estimates and t-values for the average household number of vehicles

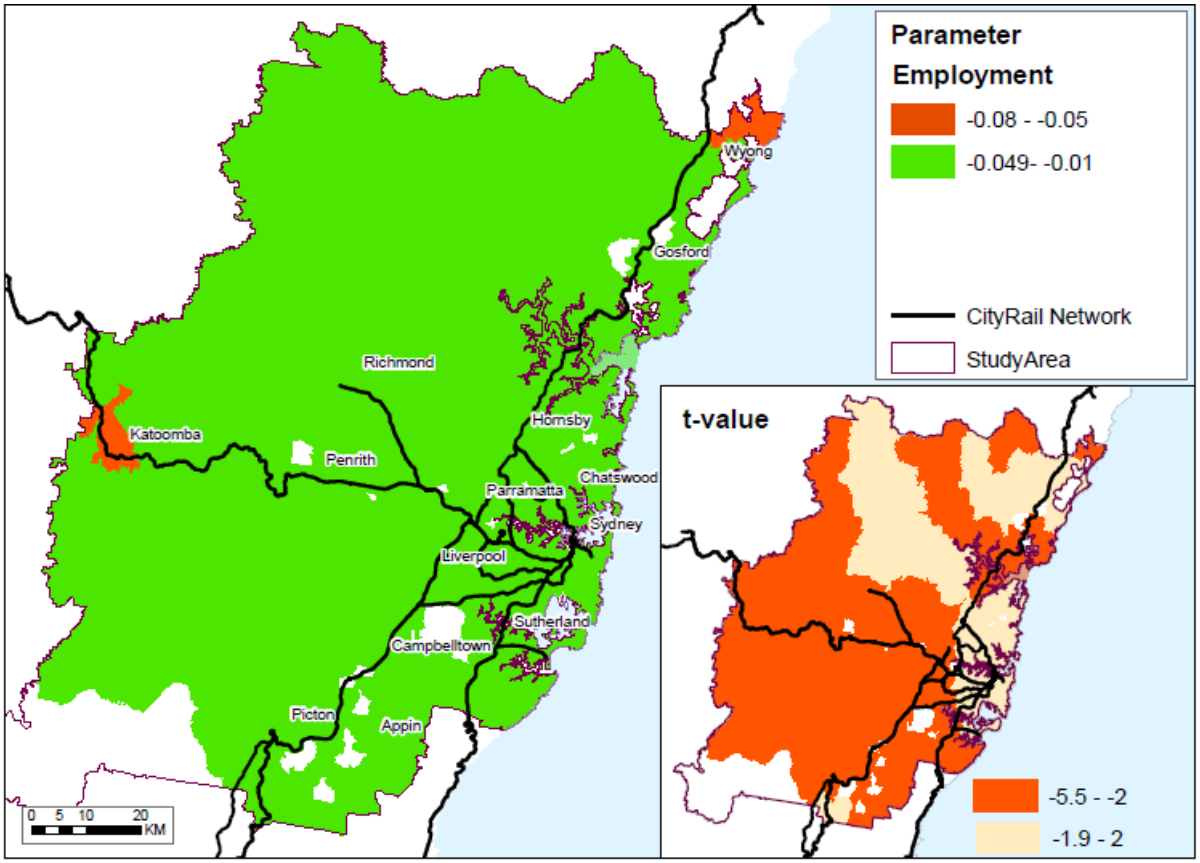


As with global parameter estimates, the local estimates can be considered one by one, assuming that all else is being held constant. Broadly speaking, parameter estimates with an absolute t-value greater than 2 can be considered statistically significantly different from zero at a 5% level of significance. The mapping of the t-values is shown in the insert to Figure 2. In the case of this

parameter, all the t-values are greater than -1.8 and so the negative values lie in the area where a hypothesis test would accept the parameter as not significantly different from zero. But there is a large area where the t-value is greater than 2, indicating that the local parameter estimates for this area are significantly different from zero. The main map depicts the local parameter estimates in terms of their values. It can be seen that these generally become less positive, the further they are from the Sydney CBD. This suggests that the further the household is located from the Sydney CBD, the less of an addition to household VKT will follow from increasing the average number of cars in a household. Indeed this geographical variability shows that the positive effect on household VKT varies between 0% and up to 20% following an increase in one in the average number of vehicles in the household at the mean of household VKT as compared to the 11% suggested by the global parameter (see Section 3.1 above). There are two areas that stand out in the map: in the south, there is a significant area where increasing the average number of cars in the household will have the most positive effect on daily household VKT and an area to the west where an increase in the average number of cars in a household will have a negative effect on daily household VKT. The spatial variation of these estimates allows the next stage of investigation to take place in seeking why such variation might be present and to derive policy that takes these differences into account.

The effect of changes in employment on household VKT is shown in Figure 3 below.

Figure 3: Parameter estimates and t-values for the number of employed (in '000) within 5km of the centroid of the TZ

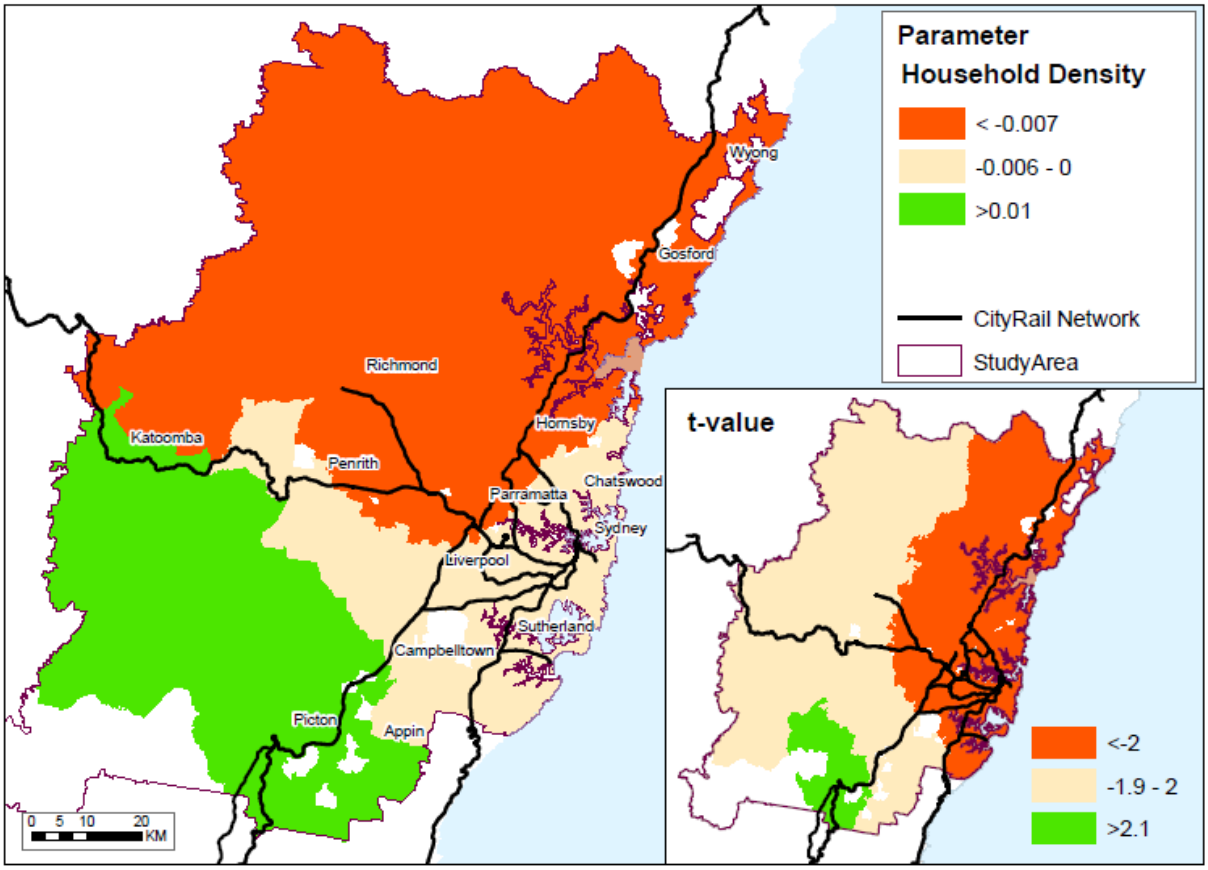


The insert shows that virtually all the local parameters are statistically different from zero with the exception of two large areas which are predominantly national park areas. The main map shows that increases in employment have a negative effect on household VKT suggesting that increasing employment within 5km of the centroid of the TZ will reduce household VKT. The area where this is most pronounced is in the Blue Mountains around Katoomba.

The effect of changes to housing density is shown in Figure 4 below. The insert shows that all there are two areas where the parameter values are statistically significant although there is a large

area where they are not. Turning now to the main map, it can be seen that there is a north-south divide in terms of the impact of a change in household density on household VKT. In the southern part of the map, the area where the parameter values are significant are positive suggesting that an increase in household density would increase household VKT. This area is however very sparsely populated and the greatest part of the geographical areas where the parameter is significant is distant from the rail network. The more interesting geographical area is to the north-west of Sydney where it suggests that the most value would come from increasing housing density (the parameters are both most negative and statistically significant). This figure is also a very good example of the enhancement (and difference) between GWR and the global (OLS) modelling techniques. In GWR, there is the opportunity to interpret over geographical space and this shows areas where there are both positive and negative impacts on household VKT from changes in the single independent variable of household density. The global (or OLS) approach averages these out into a single, and very small, average, negative, effect.

Figure 4: Parameter estimates and t-values for housing density within 2km of the centroid of the TZ



4. NEXT STEPS

The purpose of this investigation was to see if GWR could provide better explanation and fit for household VKT in the Sydney Statistical Division. The results demonstrate that this is the case and supports the criticism of Burke and Brown (2005), noted by Corpuz et al (2006) that regression based approaches 'ignore most of the contributions of location'. By taking account of location, GWR is able to offer improvement.

However, this demonstration has been in the context of data, which is available at the household level, being aggregated to the travel zone (TZ) level to provide an appropriate comparison. The next stage of analysis will be undertaken with more recent data. Census data for 2006 is now available as is data, collected by the TDC, for more recent times. This will allow an updating of the model.

Following this will be an investigation using household VKT and associated variables at the household level, aggregated to the level of the Collection District (CD) level only. It is hoped that this investigation using both the variables identified in this paper and others which are available may provide further improvements.

Further investigation around the nature of the explanatory variables is also pertinent. It is possible that different measurements of accessibility, for example, might be pertinent and available. Accessibility, as an example, has been drawing much attention from both academic and practical viewpoints and whilst it is easy to describe in terms of 'ease of reaching somewhere', there is less consensus on the best way of measuring this. In the model under consideration here, for example, the measurement of access to public transport is in a combined walk and wait time: this is evaluated for every observation even if this results in values of over 300 minutes: this might be better calibrated to a Likert scale which identifies a public transport journey as unlikely if the walk distance (converted to time) is greater than say, 800m, and a wait time in excess of 30 minutes.

In this paper, because of the intention to compare local and global estimation of an existing model, the authors have not been subject to one of the difficulties of using GWR estimation. This difficulty relates to the way in which the best global (OLS) model (defined by the adjusted R^2, the AIC and the inclusion of appropriate and meaningful explanatory variables) may not give rise to the best local model. Thus the further steps will include significant empirical iterations examining global and local models to seek the best compromise between local and global estimations.

5. CONCLUSIONS

Taking account of the spatial variation improves the fit of the model, according to the adjusted R^2 and the AIC statistics. For performance, the GWR local model is better than the global model, as measured by the absolute of difference between the observed and the predicted household VKT.

The superiority of the local model over the global model is demonstrated particularly in the maps examining the variability of the parameter estimates over space. The maps of Section 3.2 also show the depth of information that can be gleaned from local estimation. Of course, this highlights the other important feature of GWR: the output allows the visual interpretation of parameter results based on geography.

The point established by this paper – that travel behaviour attributes have a spatial context – has wide reaching implications for other analysis within the transport sector. Most practitioners would readily accept that transport and spatial characteristics are strongly connected yet most analysis ignores this connection. The availability of software that calibrates local models is a great opportunity in transport analysis to ensure that the spatial elements are properly accounted for.

REFERENCES

Anselin, L (1999), The future of spatial analysis in teh social sciences, *Geographic Informatioin Sciences, vol 5 pp. 67-76*

Blainey, S (forthcoming), Trip end models of local rail demand in England and Wales, *Journal of Transport Geography,* DOI: 10.1016/j.jtrangeo.2008.11.002

Burke, M and Brown, L (2005), Rating the Transport Sustainability of New Urban Developments: a starting point and ways forward, *Papers from the 28th Australasian Transprot Research Forum*

Corpuz, G, McCabe, M and Ryszawa, K.(2006), The Development of a Sydney VKT Regression Model, *32nd Australasian Transport Research Forum.*

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

Fotheringham, A. S. et al. (2002), Geographically Weighted Regression, John Wiley & Sons Ltd.