MODELLING THE DESTINATION CHOICE OF NEWCASTLE COMMUTERS USING LOCAL REGRESSION TECHNIQUES

Bernard Trendle
Economics and Law Research Unit, Faculty of Business, Economics and Law, University of Queensland, Australia. Email: batrendle@hotmail.com

Abstract: Frequently, studies exploring the determinants of commuting flows have adopted global modelling techniques. These techniques estimate a single set of coefficients, implicitly assuming that the same relationship applies across the entire study area. This paper tests that notion, estimating a global spatial interaction model to explain commuting outflows from Newcastle, using census 2016 journey to work data. Results from this model are compared to the information generated using a count data version of a local regression model. Finally, a spatial clustering technique is applied to the estimated coefficients of the local models to identify spatial regimes in the relationship between commuting outflows from Newcastle and the model's estimated parameters.

KEYWORDS: ABS census, commuting, destination choice, regional labour markets.

1. INTRODUCTION

Commuting is an important process with many implications for local communities. It is also an important equilibrating mechanism in local labour markets, and is likely to have a large impact on local economies. Although commuting is a well-researched topic, many studies have had methodological limitations (Kalogirou, 2003). For example, the vast majority of empirical studies on commuting assume that the role of explanatory variables on commuting flows is spatially invariant. This assumption is implicit in the statistical techniques adopted, which provide estimates of a single set of parameters. Implicitly, these models assume that the included variables impact on commuting flows equally across the study area. Although these approaches have proved useful, they may mask geographical variation in the relationship between the explanatory variables and commuting flows.

The development of statistical techniques for spatially disaggregated modelling, such as Geographically Weighted Regression (GWR), allows for investigation of the existence of spatial non-stationarity in the commuting process. These approaches permit the exploration of spatially varying relationships in datasets (Brunsdon *et al.*, 1996). In the current study, a GWR version of the Poisson spatial interaction model (GWRP) is compared to a more conventional Poisson model of spatial interaction.

Poisson and other count data models are frequently applied to commuting flow data (Persyn and Torfs, 2016). However, while these models are an improvement on the application of OLS to flow data, they are limited by the implicit assumption that the process being modelled is invariant across geographic space. This limitation is addressed by the GWR version of the Poisson model applied here (Kalogirou, 2003).

The following section provides a brief literature review, while section 3 provides an outline of the data used in this analysis. Section 4 outlines the statistical techniques used in this study, and section 5 presents the global and local modelling results. A brief conclusion is presented in section 6.

2. LITERATURE REVIEW

Journey to work patterns have been studied widely within the field of regional and urban economics. While some employees may choose to work close to home, long commutes are a common phenomenon (Giuliano and Small, 1993). In order to understand commuting behaviours, these studies have examined the role of several factors, including distance, measured in either time spent travelling or the mileage covered, education level, age, home ownership, income, work status and gender (Giuliano and Small, 1993). Many of these studies have been undertaken using either discrete choice or spatial interaction modelling approaches. Typically, these models are estimated for the entirety of the study area. For the most part, the techniques have yielded global estimates, which implicitly assume that the variables have the same impact on commuter’s decisions across this area.

Studies accessing unit record data and using discrete choice models also typically ignore the role of space or include it via the inclusion of spatial dummy variables, for example, Dijst and Vidakovic (1997) focus on commuter's destination choice decisions using the distance between locations of activity bases as their only spatial variable. More recently, Handy *et al.* (2005) and Hammadou *et al.* (2008) specify their model so that the choice of destination depends upon the characteristics of the zones (i.e. the built environment, land use, neighbourhood characteristics) as well as of the travellers socio-economic characteristics and attitudes. More recent studies using discrete choice approaches have extended the analysis via the incorporation of new data sources, especially social media data (Molloy and Moeckel, 2017; Hasnat *et al.,* 2019), however, the modelling approach has remained unchanged. While this approach means unit record data can be used and a wider range of variables included in the analysis, they frequently fail to take account of the spatial structure of the data while also assuming that the included characteristics have the same impact at all locations.

Other studies of commuting patterns have used gravity or spatial interaction modelling approaches, which are well suited to studying aggregate zonal data showing point to point journeys, such as the census data accessed for this study. For example, Persyn and Torffs (2016) study the impact of regional borders within Belgium on commuting flows, uncovering evidence that language barriers impede such flows. Their work estimates the gravity equation using a negative binomial model to take into account the over-dispersion of residuals from standard estimation approaches likely in cases when many pairs of regions have no commutes recorded between them.

Another problem confronting the analyst using spatial data is the possible presence of spatial heterogeneity. Spatial heterogeneity may arise when there is variation in the relationship being studied over space. In the context of a study like the one conducted here, the factors impacting on destination choice may have different impacts at differing locations.

The application of Geographic Weighted Regression (GWR or GWRP in this case) modelling provides one way to address this problem. Harris *et al.* (2010) note that GWR is a method of exploratory data analysis that reveals geographical variations in the estimated relationship. While traditional methods of regression analysis are essentially non-geographic, assuming that the modelled relationships are stationary across geographical space, GWR begins with the opposite view, anticipating spatial dependence and spatial variation in the relationship. It is a model of spatial heterogeneity.

Initial implementation of GWR involved local versions of OLS models (Brunsdon *et al.,* 1996); however, this has since been extended to Logistic, Gaussian and Poisson models (Kalogirou, 2003). GWR has been applied to study many topics, for example, Cahill and Mulligan (2007) study urban variation in the rates of serious crime in Portland, Oregon, Koutsias *et al.* (2010) study the incidence of wildfires in Southern Europe, while Harris *et al.* (2010) study spatial variation in participation in higher education in England. Transport studies incorporating this approach include Lloyd and Shuttleworth (2005), Paez and Currie (2010) and Blainey and Preston (2014). Lloyd and Shuttleworth (2005) study the relationship between out-commuting distance and origin locations socio-economic variables using data from Northern Ireland, while Paez and Currie (2010) study factors affecting private car use for work commutes in Melbourne. Blainey and Preston (2014) also focus on mode choice, using GWR and other techniques to study the propensity to travel to work by rail in the area around Cardiff in Wales.

3. THE DATA

There are numerous explanatory variables available for modelling the commuting patterns of Newcastle residents, but in this study only four were selected. The choice of variables to include is based on previous work and preliminary analysis where non-significant and highly correlated variables were dropped from the analysis. The final model includes the natural log of distance (*Ln_Dist*), as in many studies of spatial interaction, the natural log of the employment size (*Ln_Size*) of potential destination regions is also commonly include in studies of this kind and is included here. Many studies also include origin size, but this made no sense in this application, where the focus is on commutes from the Newcastle region. In this situation, origin size is the same for every origin-destination pair. The two other variables included are the difference in income between the origin and destination region (*Inc_Diff*), as well as the difference in occupational structure between the origin and destination region (*Occ_Diff*).

The data used in this work are taken from the 2016 Australian Bureau of Statistics (ABS) census. The Newcastle region, defined here (The Newcastle SA3)[1], has, according to data from the 2016 census, 163,881 residents with 123,934 of them (or 75.6%) being of working age. Of these working age residents, 72,075 were employed while another 6,133 are actively seeking work, giving an unemployment rate for this region of 7.8%. Of the 72,075 employed Novocastrians at census time, 49,844 worked in the Newcastle SA3 region, while 22,231 commuted outside this region for work. Thus over 30% of employed persons in the

[1] In this study, the Newcastle area is defined as the Newcastle SA3, comprising the SA2s of Adamstown - Kotara, Beresfield - Hexham, Hamilton - Broadmeadow, Lambton - New Lambton, Maryland - Fletcher - Minmi, Mayfield - Warabrook, Merewether - The Junction, Newcastle - Cooks Hill, Newcastle Port - Kooragang, Shortland - Jesmond, Stockton - Fullerton Cove, Wallsend - Elermore Vale, Waratah - North Lambton and Wickham - Carrington - Tighes Hill.

Newcastle SA3 leave the region for work, and the destination choice of this latter group is the focus of the current study.

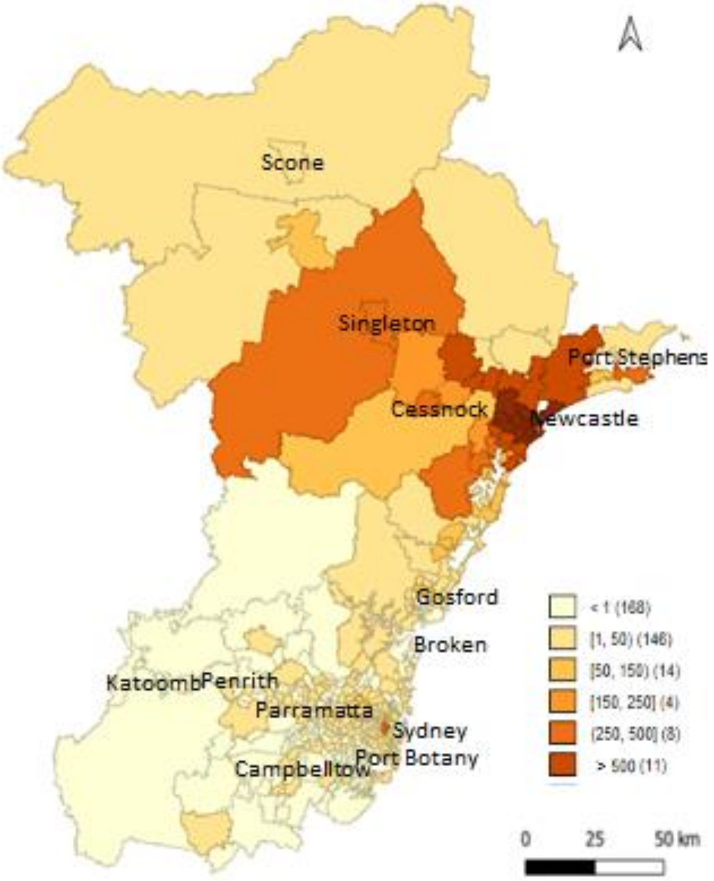

Figure 1. Destinations of Commuters from Newcastle. Source: Australian Bureau of Statistics (2016), Dataset: 2016 Census - Counting Employed Persons, Place of Work (POW).

Figure 1 provides details of the destination of commuters from the Newcastle region. The map indicates that relatively large flows from Newcastle travel to neighbouring areas, especially to the north, in a band from the Williamtown-Medowie-Karuah SA2, to Maitland West SA2 and to the south to the band defined by the Belmont-Bennets Green and Warners Bay-Boolaroo SA2s. Beyond that, an isolated outlier exists, being the SA2 of Sydney-Haymarket-The Rocks, which Journey to Work

(JTW) data indicates is the destination of just over 290 commuters from the SA2s of the Newcastle SA3.

By limiting analysis only to commuters who leave the Newcastle SA3, the analysis undertaken here can focus on the importance of regions external to the Newcastle SA3 to commuters from this area. However, this modelling strategy is a two-edged sword, and because commutes within the Newcastle SA3 area are omitted, this limits the generalisation of the results derived from the analysis, implying that they cannot be applied to people within Newcastle.

Another limitation of the local modelling techniques used in this analysis is the reduced sample size. Only subsets of the data are used to estimate each local model. This necessitates the use of relatively simple regression models with few explanatory variables to ensure adequate degrees of freedom, while the reduced sample size also increases the risk of multicollinearity in local coefficients. Chen *et al.* (2012) also note other limitations of the technique, including the incapability of decomposing the global estimates into local estimates. It is also important to remember that GWR and GWRP are considered data exploration techniques (Fotheringham *et al.,* 1996), with their strength being the ability to highlight variability in the modelled relationship. However, the modelling strategy has little predictive power, with Zhang and Shi (2004) noting that the results should not be extrapolated beyond the region for which the models have been developed.

Figure 2 provides details of the spatial distribution of the explanatory variables chosen for this analysis, with the maps presenting quintiles (divided into 5 equally sized categories). In the top right-hand panel, (2a) the numbers employed (*Ln_Size*) by SA2 are mapped. The chart indicates a small number of SA2s in the highest quintile close to Newcastle, while SA2s in this quintile are more numerous in and around Sydney.

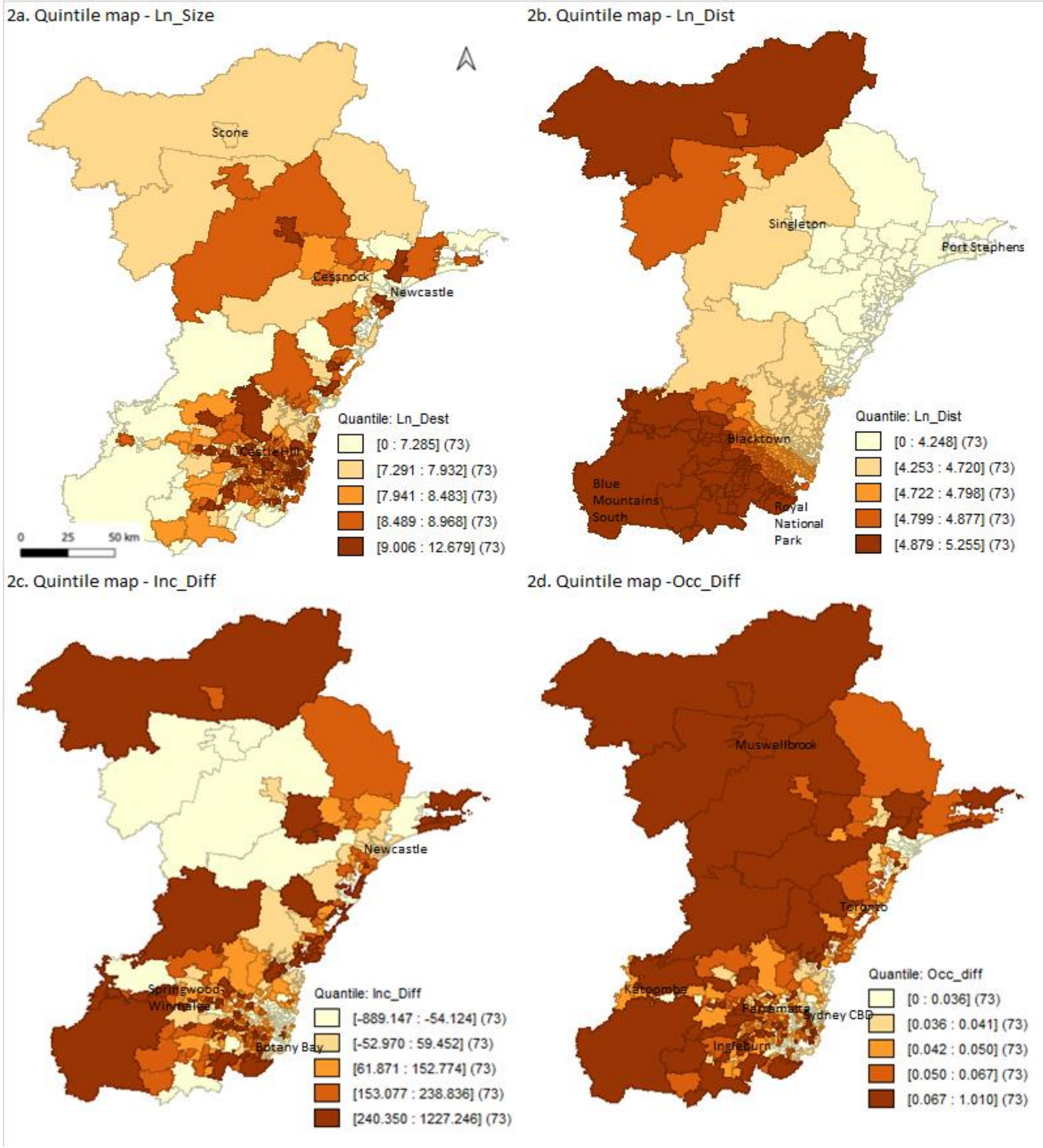

Figure 2. Spatial Distribution of Explanatory Variables. Source: Australian Bureau of Statistics (2016) Dataset: 2016 Census - Counting Employed Persons, Place of Work (POW).

Figure 2b maps the distance from the Newcastle SA3. In the current work, distance is measured as the Euclidean, or the straight line distance. This distance is measured from the centroids of each SA2. The results presented in this figure indicate that shorter distances are clustered around the Lower Hunter Valley, Lake Macquarie, Central Coast and Port Stephens regions. In contrast, the upper Hunter, and SA2s in the south-west of Sydney and in the Blue Mountains are in the category that is the greatest distance from Newcastle.

Figure 2c maps income differences (*Inc_Diff*), measured as the income of workers employed in Newcastle, less the income of workers employed in potential destination SA2s. The average income of workers is derived using census data showing employment by income bands, specifically using the midpoint of each band as suggested by Needleman (1978). In this approach, a special treatment must be adopted for the upper band, which is open ended. In his study of UK income inequality, Needleman (1978) multiplied the starting point of the upper band by 2 to create a mid-point. However, Maxell and Peter (1988), studying income inequality in Australia, have argued that income is more equally distributed here and propose multiplying the starting point of the upper income band by 1.5. While this approach is somewhat arbitrary, lack of better information justifies its use, and it is adopted here. Using this approach, it can be seen that SA2s falling in the lowest quintile tend to have incomes higher than the average income of residents of Newcastle. Further, SA2s falling in this quintile tend to be clustered in the Lower Hunter region, a region dominated by the mining industry. This category also includes SA2s around the Sydney CBD, an area dominated by the Public Administration and Finance industry and having many high level managerial and executive positions.

Finally, figure 2d maps the distribution of the quintiles of Occupational differences (*Occ_Diff*). This variable runs from 0 up to 1 and is measured as the Mean Absolute Percentage Deviation between the occupational structure of persons living in Newcastle and the occupational structure of jobs in potential destination SA2s. After calculation, this variable is normalised by dividing the derived values by the maximum possible value for this variable when there are only 8 (the number of 1st division ANZSCO occupations) categories included in the calculation.

SA2s with an occupational structure close to that of Newcastle have low values, while occupational structures increasingly different to Newcastle have higher values. Apparent from this figure is a large cluster of regions with a different occupational structure to Newcastle lying relatively close to Newcastle in the Hunter Valley, and also west of the Central Coast.

4. GLOBAL VERSUS LOCAL MODELS OF COMMUTING

The current study is concerned with the effect of destination characteristics on attracting Newcastle commuters. In terms of the gravity modelling theory, the destination choice model adopted here is an origin-specific or production constrained model (Nakaya, 2001; Kalogirou, 2003). The literature indicates that the Poisson regression is one of the most appropriate methods when analysing commuting or migration data (see, for example, Flowerdew and Aitkin 1982). In the case of commutes, the dependent variable is a count of individuals who make their decision about the location of work independently. If commutes are relatively small compared with the size of the total workforce, these flows can be considered rare events, thus, the choice of the Poisson distribution.

The simple Poisson model can be written as:

$$C_i = \text{Poisson}[E_i \exp(\square_0 + \square_k x_k)] \tag{1}$$

Where, C_i is the flow of commutes, E_i is the expected number of commutes based on the regional structure, $\square_0$ is the intercept and x_k a vector of k explanatory variables and $\square_k$ their coefficients.

One issue with the use of spatial interaction models is that for the most part, the estimation procedures have tended to ignore the role of geography. While there have been many studies over the last two decades using spatial econometric techniques for simple linear models, fewer applications using count data have adopted appropriate techniques. In this study, an extension of Geographically Weighted Regression for count data, specifically GWRP (Poisson) is applied, with the aim of addressing the problem of spatial autocorrelation while also exploring the variability of the estimated relationship.

GWRP is estimated non-parametrically, giving:

$$C_i = \text{Poisson}[E_i \exp(\square_0 (u_i) + \square_{1(u_i)} x_k)] \tag{2}$$

Where, $\square_0$ and the $\square_k$ are unspecified bivariate functions of geographic location, u_i (= (u_{xi}, u_{yi})) with the u_i being a vector of co-ordinates describing the location of each observation (i) (Nakaya *et al.*, 2004). Although this shifts the model back to linear predictors, it provides interaction between geographical location and functional relationships in the linear predictor. This means that potentially, the model has different coefficients for each location.

It is necessary to note that for the calibration of the data, the weighting scheme adopted is the adaptive kernel using a bi-square function to compute the weights w_{ij} using the function below:

$$w_{ij}=\begin{cases}[1-(d_{ij}/h_i)^2]^2 & \text{if } d_{ij} \leq h_i \\ 0 \text{ otherwise}\end{cases} \tag{3}$$

Where, d_{ij} is the distance between the regression point *i* and each location *j* around *i*; and h_i is the *n*th nearest neighbour distance from *i*. The adaptive bi-square kernel is used in this study because the adaptive kernel choses a bandwidth based on a fixed number of nearest neighbours, rather than a fixed distance. Given the unequal sizes of the SA2s in the area being studied, this is considered the appropriate approach.

It is worth noting that previous work with GWR indicates that the use of different continuous weighting functions does not have much influence on model results (Fotheringham, Charlton and Brunson 2001 and Fotheringham, Brunson and Charlton 2002). However, the selection of the bandwidth used in the analysis can significantly affect model calibration. In application, an optimum kernel bandwidth for the GWR is frequently found by minimising some model fit diagnostic. In this study, the corrected version of the Akaike Information Criterion (i.e. AICc) is adopted. For GWR and GWRP, this approach ensures that models using smaller bandwidths receive a higher penalty than those using large bandwidths. Thus for a GWR with a bandwidth *b*, its AICc can be found from:

$$AIC_c = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left\{\frac{n+tr(S)}{n-2-tr(S)}\right\} \tag{4}$$

Where, *n* is the (local) sample size (according to *b*); $\hat{\sigma}$ is the estimated standard deviation of the error term; and tr(*S*) denotes the trace of the hat matrix *S*. The hat matrix is the projection matrix from the observed *y* to the fitted values, $\hat{y}$. The optimal bandwidth is calculated following an iterative process that tries different bandwidths and, in this case, minimises the AICc.

5. GLOBAL AND LOCAL MODELS OF COMMUTING FROM NEWCASTLE

Table 1. Global and Local model results. Source: ABS 2016 Census of Population and Housing and Author's calculations.

Global model results					GWRP (Poisson) model results					
Number of parameters:	5				Effective no. of parameters:	370.490				
Degrees of freedom	346				Degrees of freedom	284.843				
Deviance:	12115.490				Deviance:	1561.711				
					Optimal bandwidth	51				
AICc:	12125.664				AICc:	1725.328				
% deviance explained	0.891				% deviance explained	0.985				
Variable	Estimate	s.error	z-statistic	p-value	Variable	Min	Lower Quartile	Median	Upper Quartile	Max
Intercept	6.902	0.096	72.230	0.000	*Intercept*	-155.105	-5.692	4.249	28.062	156.235
Ln_Size	0.603	0.009	70.015	0.000	*Ln_Size*	0.575	0.958	1.077	1.364	2.781
Ln_Dist	-2.076	0.009	-242.470	0.000	*Ln_Dist*	-35.550	-8.253	-2.920	-0.683	29.474
Inc_Diff	-0.001	0.000	-27.727	0.000	*Inc_Diff*	-0.007	-0.002	-0.001	0.000	0.007
Occ_diff	3.476	0.304	11.423	0.000	*Occ_diff*	-53.947	-6.684	-2.879	7.087	24.773
					Geographical variability tests of local coefficients					
					Variable	Diff of deviance	Diff of DOF	DIFF of Criterion		
					Intercept	n.a.	n.a.	n.a.		
					Ln_Size	860.609	11.344	-829.889		
					Ln_Dist	277.219	5.341	-261.232		
					Inc_Diff	187.972	11.716	-153.659		
					Occ_diff	186.875	12.490	-150.390		

Table 1 presents the results of the global and local models of commuting estimated in the current study. The left hand panel provides details of the estimated coefficients and diagnostics for the global version of the Poisson model. This model is global in the sense that the estimated relationship is assumed to hold across the entire study area. The model has an intercept and 4 parameters, all of which are statistically significant, while also having the expected sign. Additionally, this model explains a high (0.84) percent of the observed variation in the dependent variable (the number of out-commutes from Newcastle).

The positive coefficient of *Ln_Size*, the natural log of the number employed in destination SA2s is expected and is frequently uncovered using spatial interaction models. The larger the employment size destination region, the more likely commuters will travel there. The negative coefficient of *Ln_Dist*, the natural log of distance, is also commonly found in these sorts of studies. Distance acts as a deterrent, and, all else being equal, the number of commutes is expected to decline at greater distances.

Remembering that *Inc_Diff*, the difference between the average weekly incomes of workers working in the origin and destination regions is derived as $Income_o – Income_d$, then the negative coefficient indicates that flows are expected to be lower when the destination regions income is lower than the origin regions income. This is another common finding and is consistent with a Human Capital explanation of commuting. Incomes greater than the average of the origin region act to increase commutes, while incomes below the average of the origin region act as a deterrent to commuters. For this reason, commuters are likely to earn more than residents who choose to work in the home region, with the higher incomes providing an incentive to undertake the relatively costly commute.

Finally, *Occ_Diff*, the index of differences between occupational structure of workers residing in the origin region and the occupational structure of workers employed in destination regions, has a positive coefficient in the global model. This result might be considered surprising, as one might expect that the likelihood of workers from Newcastle finding employment increases when the industrial base of a potential destination region has a similar occupational structure to the occupational structure of the workforce residing in Newcastle. However, it is also likely that workers who chose to commute, and travel longer distances for work, expect to earn higher incomes than if they accepted a job close to their place of residence, so this may not be such an unexpected result.

For the local modelling, the optimal bandwidth for the GWRP model was calculated at 51, meaning that the 51 nearest SA2s were used in estimating each local parameter. With a bandwidth of 51, an additional 370 parameters were included in the model. The proportion of the deviance explained has increased from 0.891 in the global model, to 0.986 in the local models, while the AICc has fallen from 12,125.664 to 1,725.328.

The final stage of testing comprised determining whether the GWR version of the Poisson model is preferred to the global or a mixed model, incorporating both spatially varying and global parameter estimates. The geographical variability tests (Nakaya, 2016) are presented in the lower right-hand panel of table 1. Generally speaking, Difference criteria (final column of test results) less than 0 are interpreted as providing support for the idea that there is spatial variation in the coefficient estimates. This idea is supported by the results presented here, thus the test supports the idea of the GWRP being superior to the global model.

For the local variable estimates of the GWRP model, Table 1 also provides information on the range, mean, lower and upper quartiles. More information is provided in the detailed output, which includes parameter estimates at each location. The problem with GWR and GWRP is the volume of results available and frequently data reduction techniques are adopted to make sense of the results and this approach is adopted here, with clustering techniques used to categorise or coefficient estimates into regimes. Because this study deals with regional data, the Skater algorithm[2] is used to group the data (Assuncao *et al.,* 2006). Unlike conventional forms of cluster analysis, such as PCA and Hierarchical clustering, the Skater algorithm, explicitly takes into account the contiguity constraints in the clustering process.

A number of clusters/regimes were defined and it was found that five clusters, with a minimum number of 30 SA2s to each cluster, provided the most reasonable result. To compare cluster definitions, a measure

[2] The algorithm carries out a pruning of the minimum spanning tree created from the spatial weights matrix for the observations. The weights of the spatial weights matrix correspond to the pair-wise dissimilarity between observations. Starting from this weights matrix, a minimum spanning tree is obtained. This is a path that connects all observations. The *n* observations are connected by n-1 edges, such that the between observation dissimilarity is minimised. This tree is then pruned by selecting the edge whose removal increases the between group dissimilarity the most.

defined by the ratio of the sum of squares to total sum of squares is frequently used (see Anselin, 2018), and it was found that this ratio was higher for five clusters than for four. In addition, it was also found that the specification of a minimum number of 30 SA2s in each cluster had no impact, as this constraint was not met in this application, with the optimum value of the ratio of the sum of squares to total sum of squares being found where n = 37 for regime 5.

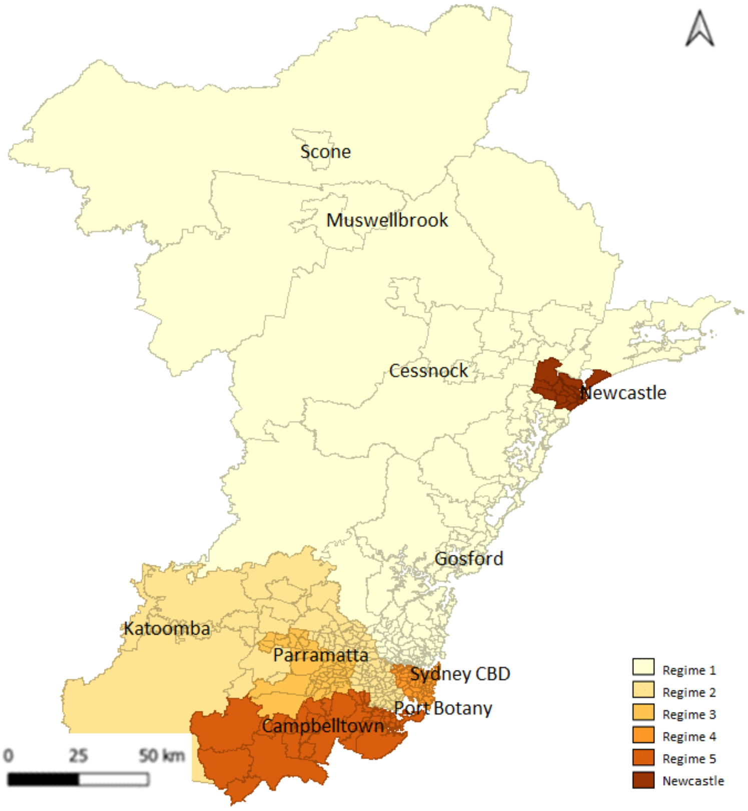

Figure 3. Regimes Identified Using Cluster Analysis. Source: ABS 2016 Census of Population and Housing and Author's calculations.

The five regimes identified by the cluster analysis are shown in figure 3. They are contiguous, i.e., there are no islands, which is an attractive feature of location-based clustering methods such as the Skater algorithm. The largest cluster (Regime 1), with 120 SA2s, comprises and band running from Upper Hunter, east to Port Stephens and south to Central

Coast. This cluster has SA2s with relatively small numbers employed within it, yet is relatively accessible for workers from Newcastle. In contrast, the second largest cluster, with 103 SA2s, is made up of SA2s in the Blue Mountains, north-west Sydney and the western fringe of Botany Bay. Regime 3 comprises 47 SA2s around Parramatta. This is an area which is relatively inaccessible to commuters from Newcastle, with an average distance of 133.7km. Regime 4 centres on the Sydney CBD and Port Botany. This area not only has a large number of jobs (i.e. census data indicates there were 308,508 jobs in this regime), it has a high average income (with average weekly income $117.70 greater than Newcastle average incomes) and, while a great distance from Newcastle CBD (2.5 hours by train), the rail connection is direct and relatively frequent during peak hours. In contrast, regime 5 lies south and south-west of the Sydney CBD and is centred around Campbelltown. This is an area with a relatively low number of jobs (i.e., just over 83,036 jobs or an average of 2.244 per SA2), while also being relatively inaccessible to Newcastle commuters, with an average straight line distance of 142.7 km.

Table 2 provides details of the differences in parameter estimates across the 5 regimes identified in the cluster analysis. To characterize each of the identified regimes, descriptive statistics for actual variable values to describe each area’s structural characteristics, along with parameter estimates to characterize the model of out-commutes from Newcastle, are included. Cahill and Mulligan (2007) note that while significance tests are not appropriate for parameter estimates reported this way, the average values are useful for describing models in different parts of the study area and can be compared to the global model. For the sake of brevity, the discussion here is limited to regimes 1 and 4. Regime 1 lies closest to the origin region of Newcastle, and accounts for the largest number of out-commutes, while regime 4 is the Sydney CBD, the centre of commerce in the state of New South Wales. The details provided in table 2 allow interested readers to interpret the results for the remaining regimes.

Regime 1, the largest cluster, comprises 120 SA2s. It is a large regime, with 419,709 persons working in it and flows from Newcastle to this regime are high, with an average of 176.7 Newcastle workers per SA2, or 21,205 persons in total commuting to the SA2s of this regime. The average straight line distance is 57.7km, making it the closest regime to Newcastle by a considerable margin. The average weekly income of workers employed in this regime is close to Newcastle (about $73.26 per week less). This lower income suggests that the incentive income provides to commute to this regime is low, but that this is offset by the size of the region and the fact it is adjacent to Newcastle.

Table 2. Descriptive Statistics and Structural Measures and Parameter Estimates by Cluster Group. Source: ABS 2016 Census of Population and Housing and Author's calculations.

	Regime 1		Regime 2		Regime 3		Regime 4		Regime 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Structural measures										
Count = 187										
Flow	176.708	487.929	2.990	6.002	1.936	3.448	13.500	43.436	0.892	1.914
Ln_Size	3497.575	2.621	3212.527	3.717	3450.842	2.438	6564.016	3.226	2244.204	5.507
Ln_Dist	57.696	2.071	125.268	1.105	133.689	1.043	113.302	1.027	142.743	1.107
Inc_diff	73.256	241.719	113.010	181.444	168.958	128.581	-117.493	213.985	150.860	229.066
Occ_diff	176.708	487.929	2.990	6.002	1.936	3.448	13.500	43.436	0.892	1.914
Estimated parameters										
Ln_Size	1.009	0.138	1.258	0.280	1.790	0.361	0.90	0.09	1.39	0.35
Ln_Dist	-1.611	1.769	-10.446	8.630	-6.977	4.837	2.89	4.23	1.66	13.72
Inc_diff	-6.605	0.001	0.592	0.001	12.579	0.001	-4.06	0.00	-12.28	0.00
Occ_diff	1.000	11.483	0.000	9.201	0.000	4.052	0.00	1.51	0.00	19.18

The lower part of Table 2 provides details for the coefficient estimates. For this regime, the results indicate that commutes are more sensitive to the size of the destination region (*Ln_Size*) with the average value of the coefficient of this variable at 1.009 in this regime, compared to only 0.603 in the global model. Commuters in this regime are also more sensitive to *Inc_diff*, the difference in incomes between the origin and destination regions, with a coefficient of -0.002, compared to -0.001 in the global model. In contrast, commuters from Newcastle to this regime are less sensitive to *Ln_Dist*, the natural log of distance, with the average parameter estimate at -1.611 in this regime, compared to -2.076 in the global model, while the average estimated coefficient for *Occ_Diff*, the difference in occupational structured between Newcastle and potential destination regions is 6.605, about twice the absolute size and the opposite sign of the value in the global model (3.476). It seems that for destinations that are close to Newcastle, flows are larger where the occupational structure is more similar.

Regime 4 consists of 44 SA2s around the Sydney CBD to the northern side of Botany Bay. Census data indicates that there were 594 commuters from Newcastle to these SA2s (with an average of 13.5 per SA2) and that the average number of persons working in each SA2 was 6,564, or that there were some 288,816 persons working in these 44 SA2s. This regime, with its focus on the Sydney CBD, is the centre of commerce in NSW. Average incomes are higher, with census data indicating that they are $117.49 per week higher than in Newcastle. Despite its distance (with the SA2s being an average 113.3km from Newcastle), these high incomes and a direct rail connection may make this regime relatively attractive for well-qualified Newcastle residents. The coefficient estimates in the lower panel of table 2 indicate that commutes to the Sydney CBD cluster are only slightly more sensitive to employment size than the global model results. In this regime, the average values of the parameter estimates of *Ln_Size* at 0.899, compared to 0.603 in the global model, implying that commuters from Newcastle are more likely to commute to large employing SA2s in this regime. Additionally, the average parameter estimates of *Inc_diff* (-0.001) are identical to the global model. In contrast, distance and occupational structure pull in the opposite manner to the global estimates. In this regime, distance (*Ln_Dist*) does not seem to acts as a deterrent as in the global model, with the average estimated coefficient at 2.890 indicating that the flows are above what would be expected, given this regimes distance from Newcastle. For occupational structure (*Occ_Diff*), it seems that Newcastle commuters are more likely to commute to destinations in this regime that are more similar in occupational structure to the SA2s of Newcastle.

6. CONCLUSIONS

In this paper, spatial interaction modelling has been applied to Newcastle out-commutes, with the aim of investigating the role of the factors influencing destination choice. While spatial interaction modelling is not new, many applications have used less than ideal statistical techniques to conduct analysis. For example, Aroca and Hewings (2002) note that many early studies have linearized the model and applied OLS regression, while Johansson *et al.*, (2003) note that more recent work has often used Poisson or Negative binomial models which provide global estimates, assuming that the relationship is stable across the study area.

Table 1 provides results from a Poisson regression, addressing some of the limitations of traditional applications of spatial interaction models. The estimated coefficients provide few surprises. The coefficient for the natural log of the number of persons working in the SA2 (*Ln_Size*) is positive, in line with results commonly derived in these sorts of studies. Commuters are more likely to commute to regions where there are many jobs. The same conclusion holds for the negative coefficient of the natural log of distance (*Ln_Dist*), there are no surprises there also. The negative coefficient implies that distance acts as a deterrent to commuting. Commuting is costly in time and money, so all else being equal, people will prefer to work closer to their residential location. The coefficient for *Inc_Diff* is also negative. Income difference is measured as the difference between the average weekly incomes of workers in the origin and destination regions and is derived as ($Income_o$ – $Income_d$). The negative coefficient implies that incomes greater than the average of the origin region provide an incentive to commute. Again, this is an expected result. *Occ_Diff*, the index of differences between occupational structure of workers residing in the origin region and the occupational structure of workers employed in destination regions has a positive coefficient. Because workers who leave their local area for work are likely to be high-income individuals, it could be that SA2s with an occupational structure skewed towards high-paying jobs are more attractive destinations.

While Poisson regression is an improvement over traditional OLS techniques for the derivation of parameter estimates for models based on count data, such as commuting flows, a limitation is that the spatial nature of the data is overlooked. In an attempt to address this shortcoming, this paper has implemented an extension of Geographically Weighted Regression version of the Poisson model (i.e., GWRP), which allows the modelling of count data (Nakaya, 2001).

The estimation of parameters that vary at each point across the study area limits the predictive power of these models. For this reason, GWR and GWRP should be seen as data exploration tools. With that in mind, the conclusions highlight the variability in the relationship being modelled. Results from the local modelling and geographical variability test seem to indicate that the assumption of parameter stability, an underlying assumption of global spatial models, is not supported by the data used in this analysis. This conclusion will have a number of implications.

Firstly, while accessibility, included as the natural log of distance (*Ln_Dist*) is important, the results uncovered in the GWRP indicate that the effect of distance on commuting flows is not as simple as might be suggested by the results obtained from the global model. The application of the Skater techniques led to the identification of five regimes over which the average coefficient for distance varies considerably. Commuting flows respond differently to distance at different locations. In regimes 1 (-1.6119), 2 (-10.446) and 3 (-6.977) commuting flows respond as expected, with distance associated with a reduction in commuting flows, while in the remaining regimes, distance is associated with increased flows. In these regimes, it seems that distance does not act as a deterrent as is generally expected.

The effect of the other variables shows similar variation, although the results do not follow the same pattern across the regimes identified in the analysis for any of the variables. For differences between estimates of *Inc_diff* (the incomes of workers in Newcastle, less the income of workers in potential destination SA2s), there was less variation in the coefficient estimates, though regime 3 (0.002) is twice the absolute size and the opposite sign of that found in the global model, while the coefficients for Occupational difference (*Occ_Diff*) also shows considerable variation ranging from 12.579 in regime 3, to -12.280 in regime 5. It seems that only the employment size of the regime (*Ln_Size*) does not vary in sign across the five regimes, though it varies from an average of 1.790 in regime 3, to 0.899 in regime 4.

These findings will have a number of implications, particularly for the study of regional labour markets, regional policy and also the evaluation of various transport infrastructure projects. For example, Persyn and Torfs (2015) note that commuting is an important spatial equilibrating mechanism in the labour market. In standard closed-economy labour market models, commuting reduces disparities in regional labour market outcomes such as unemployment rates and wages, and brings aggregate welfare gains (Borjas, 2001). However, commuting is costly, with these costs including those related to commuting distance. This study has shown that these costs are not as directly related to distance, as concluded by analysis based on global models (Persyn and Torfs, 2015). This has implications for the role of commuting in equilibrating disparities in local labour market outcomes, with the results uncovered in this analysis indicating that distance has varying impacts at different locations and that locations relatively far away, or inaccessible, may play a stronger role than expected in equilibrating labour market disparities.

The results may also challenge the concept of local labour markets, frequently defined as a geographical area within which a high percentage of commuting by residents occurs (Coombes, 2000). Watts (2004) notes that a number of approaches have been adopted to define local labour markets, with all of these approaches attempting to define exhaustive and mutually exclusive local labour markets. This way of conceptualising local labour markets as contiguous self-contained regions is challenged by the variability of the coefficient estimates uncovered by the local models presented here. This work suggests that areas that are relatively close, may be less attractive destinations than relatively distant locations and that the strength of the variables determining the attractiveness of various work locations are not even across geographic space. For any given residential location, the attractiveness of potential destination areas may vary widely, independent of the attractor variable values. This may result in non-continuous local labour markets and local labour markets with overlapping boundaries. Such a finding will make the analysis of local labour markets and regional policy designed to improve regional labour market outcomes more challenging.

Finally, the analysis of commuting data is an important aspect of transport studies and frequently the evaluation of costs and benefits of transport infrastructure involve the use of large-scale, city-wide models. One component of these models are coefficients derived by applying statistical modelling techniques to census or survey data. These models may be derived by treating the whole of the study area as homogeneous, so that a single set of parameter estimates are used to explain the destination choice of residents. The results presented in this study challenge the notion that the factors driving the destination choice of Newcastle workers are invariant across the study area considered. This finding argues for an additional step in the evaluation of transport infrastructure projects. According to the results here, consideration of potential local variation from the impacts derived from the global model should be the focus at this step. Clearly, all potential job destinations are not equal.

REFERENCES

Anselin, L. (2018). Geoda workbook. Online version accessed December 2020, https://geodacenter.github.io/documentation.html.

Aroca, P. and Hewings, J. D. (2002). Migration and regional labour market adjustment: Chile 1977-1982 and 1987-1992. *The Annals of Regional Science*, 36, pp. 197-218.

Assuncao R. M, Neves, M., Camara, G. and Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Journal*

International Journal of Geographical Information Science, 20, pp. 797-811.

Blainey, S. and Preston, J. (2010). A geographically weighted regression based analysis of rail commuting around Cardiff, South Wales. Proceedings of the 12th World Conference on Transportation Research, Lisbon, July 10 to 14.

Brunsdon C., Fotheringham, A. S. and Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28, pp. 281-298.

Borjas, G. J. (2001) Does immigration grease the wheels of the labor market? *Brookings Papers on Economic Activity*, *63*, pp. 69–133.

Cahill, M and Mulligan, G. (2007). Using Geographically Weighted Regression to Explore Local Crime Patterns. *Social Science Computer Review*, 25, pp. 174-193.

Chen, V., Deng, W., Yang, T. and Matthews, S. (2012). Geographically Weighted Quantile Regression (GWQR): An Application to U.S. Mortality Data. *Geographic Analysis*, 44, pp. 134-150.

Coombes, M. (2000). Defining Locality Boundaries with Synthetic Data. *Environment and Planning A*, 32, pp. 1499-1518.

Dijst, M. and Vidakovic, V. (1997). Individual action-space in the city. In Ettema, D, and Timmermans, H. (Eds) *Activity-based Approaches to Travel Analysis*, Pergamon Press, Oxford, UK.

Flowerdew, R. and Aitkin M. (1982). A method of fitting the gravity model based on Poisson distribution. *Journal of Regional Science*, 22, pp. 191-202.

Fotheringham, A., Charlton, M. and Brunsdon, C. (1996). The Geography of Parameter Space: An Investigation of Spatial Nonstationarity. *International Journal of Geographical Information Systems*, 10, pp. 605–627.

Fotheringham, A., Charlton, M. and Brundson, C. (2001). Scale issues and geographically weighted regression. *Geographical Analysis*, 35, pp. 272-275.

Fotheringham, A., Brunson, C. and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley, ISBN: 978-0-471-49616-8.

Giuliano, G. and Small, K. (1993). Is the Journey to Work Explained by Urban Structure? *UrbanStudies* 30, pp. 1485-1500.

Hammadou, H., Thomas, I., Verhetsel, A. and Witlox, F. (2008). How to Incorporate the Spatial Dimension in Destination Choice Models: The Case of Antwerp. *Transportation Planning and Technology*, 31. pp. 153-181.

Handy, S., Cao, X. and Mokhtarian, P. (2005). Correlation or causality between the built environment and travel behavior? Evidence from Northern California. *Transportation Research* D, 10, pp. 427-444.

Harris, R., Singleton, A., Grose, D., Brunsdon, C. and Longley, P. (2010). Grid-enabling Geographically Weighted Regression: A Case Study of Participation in Higher Education in England. *Transactions in GIS,* 14, pp. 43–61.

Hasnat, M., Faghih-Imani, A., Eluru, N. and Hasan S. (2019). Destination Choice Modeling using Location-based Social Media Data. *Journal of Choice Modelling*. 31, pp. 22-34.

Johansson, B., Klaesson, J. and Olsson, M. (2003). Commuters' non-linear response to time distances. *Journal of Geographical Systems*, (5), pp. 315-329.

Kalogirou, S. (2003). Destination Choice of Athenians: An Application of Geographically Weighted Versions of Standard and Zero Inflated Poisson Spatial Interaction Models. *Geographical Analysis*, 48, pp. 191-230.

Koutsias, N., Martinez-Fernandez, J. and Allgower, B. (2010). Do Factors Causing Wildfires Vary in Space? Evidence from Geographically Weighted Regression. *GIScience & Remote Sensing*, 47, pp. 221–240.

Lloyd, C. and Shuttleworth, I. (2005). Analysing commuting using local regression techniques: scale, sensitivity, and geographical patterning. *Environment and Planning A*, 37, pp. 81-103.

Maxwell, P. and Peter, M. (1988). Income inequality in small regions: A study of Australian Statistical Divisions. *The Review of Regional Studies*, 18, pp. 19-27.

Molloy, J. and Moeckel, R. (2017). Improving destination choice modeling using location-based big data, *International Journal of Geo-Information*, 6, pp. 291-306.

Nakaya, T. (2001). Local spatial interaction modelling based on the geographically weighted regression approach. *GeoJournal*, 53, pp. 347-358.

Nakaya, T. (2016). GWR4.09 User Manual, https://sgsup.asu.edu/sites/default/files/SparcFiles/gwr4manual_409.pdf

Needleman, L. (1978). On the approximation of the Gini coefficient of concentration. *The Manchester School*, 46, pp. 105-122.

Paez, D. and Currie, G (2010). Key Factors Affecting Journey to Work in Melbourne using Geographically Weighted Regression, ATRF

2010: 33rd Australasian Transport Research Forum, 29 September to 1 October 2010, National Convention Centre, Canberra, https://www.australasiantransportresearchforum.org.au/sites/default/files/2010_Paez_Currie.pdf.

Persyn, D. and Torfs, W. (2016). A gravity equation for commuting with application to estimating regional border effects in Belgium. *Journal of Economic Geography*, 16, pp. 155-175.

Watts, M. (2004). Local Labour Markets in New South Wales: Fact or Fiction? Working Paper No. 04-12 Centre for Full Employment and Equity, The University of Newcastle, Online version accessed 2 June 2021, https://www.semanticscholar.org/paper/Local-labour-markets-in-New-South-Wales%3A-fact-or-Watts/d180e0c3442f0100ad5f27e191d3bd2c44ac afed

Zhang, L. and Shi, H. (2004). Local Modeling of Tree Growth by Geographically Weighted Regression. *Forest Science*, 50, pp. 225-244.