Developing disaggregate transport prediction models from aggregate survey data

Stuart Bain and David A. Hensher

Institute of Transport and Logistics Studies
Faculty of Economics and Business
University of Sydney, Sydney, NSW 2006, Australia
{S.Bain, D.Hensher}@itls.usyd.edu.au

Abstract: Different transportation modes (such as road, rail, coach and air) have vastly different economic and environmental impacts. Understanding the interrelationships between transportation networks, the economy and the environment therefore requires an understanding of the modal shares of the transportation network. The acquisition of mode-choice decision data is however, a costly and time-consuming process, often necessitating a reliance on secondary data sources. Such secondary sources may well be at a higher level of aggregation than required, again due to the costs and time involved in fine-grained data collection.

In such cases, one method to satisfy the need for disaggregate data is to predict the lower level mode-choice decisions from the known mode-choice decisions at the more aggregate level. This paper examines such a scenario, using a secondary data set collected at a high level of aggregation (15x15 OD-pairs) to develop a model to predict transport behaviour at a lower level of aggregation (154x154 OD-pairs).

Keywords: regional transport, data disaggregation, prediction, multi-modal passenger travel, secondary data

1 INTRODUCTION

Over the past two decades there has been significant research devoted to the construction of land use and transport models for major metropolitan areas with relatively high population densities (Wegener, 2004). In contrast, the development of analogous models for rural and regional areas has not attracted the same level of interest, with the notable exceptions of a small number of national models developed primarily in Europe (Daly, 2008) These systems evidence the usefulness of models capable of evaluating policy instruments outside of metropolitan areas, but no such capability currently exists in Australia.

The development of such models requires both demand- and supply-side information about transportation movements and networks. Furthermore, this information must be available in the same geographic units as used in the model. Whilst much supply-side information is publicly available (for example, census data on vehicle ownership, public transport timetables, airline route schedules etc.), demand-side data is more difficult to obtain. The reasons for this are many: trips by personal vehicle are not centrally recorded; commercial operators may consider their patronage figures to be commercially sensitive, and governments may consider figures from nationalised operations to be politically sensitive.

In the absence of actual demand-side figures, travel surveys become one of the main methods of obtaining demand-side data. The collection of primary survey data is a costly process however, particularly if the survey is to be conducted at a high level of geospatial disaggregation. It must be remembered that if origin-destination (OD) data is to be collected, the number of OD-pairs increases quadratically with respect to the number of zones.

If sufficient resources are not available to commission a survey especially for the model to be developed, a practitioner may be forced to rely on a survey undertaken for an entirely different purpose. There is no guarantee that the available survey data will employ the desired geospatial partitioning. The survey data may have been collected at a higher level of aggregation or perhaps using an overlapping geography to that which is to be used.

This study considers the former case, examining the disaggregation of mode choice trip data collected for large, highly aggregate geospatial areas to smaller, less aggregate geospatial units. Using a secondary survey data source collected at a high level of geospatial aggregation, trip-prediction models for four different modes of transportation are estimated. These models are then applied at the desired (more disaggregate) level to obtain trip predictions for this level. The predictions can then be used in a subsequent stage to develop an integrated land use and transport model system.

2 METHODOLOGY AND RESULTS

2.1 Data source

The primary data source discussed in this paper is the *National Visitor Survey* (NVS) conducted annually by Tourism Research Australia (2008). Each year, 120,000 Australian residents are surveyed using random digit dialing and a computer aided telephone interview (CATI). This survey is the principal source of information on domestic tourism movements in Australia. Information about the movements of international visitors is collected separately. As "tourism"" in this context is defined to include holidaying, visiting friends and relatives and travel for business, education or employment, the survey covers a broad range of passenger movements.

Data from the survey is divided into two parts: movements involving an overnight stay, and those that do not. Whilst this division is relevant from a tourism perspective (overnight stays lead to economic activity in the accommodation and hospitality sectors), from a transport perspective no such distinction is necessary. For this reason, all trips, leading to an overnight stay or not, were combined into a single data set for the purposes of this study[1]. The survey considers five different modes of transport: car (both private and rental), bus, plane, and other (train). The two car modes were combined into a single mode.

Data from the NVS is reported at the *Tourism Region* (TR) level. Tourism regions are defined by State and Territory tourism authorities and are comprised of sets of *Statistical Local Areas* (SLA), the smallest spatial unit in the Australian Standard Geographical Classification in non-Census years[2] (ABS, 2008). This study focuses on passenger movements within New South Wales (NSW) and the Australian Capital Territory (ACT). In total, there are 15 tourism regions within NSW and the ACT (see Figure 1).

Details of the trip data collected in the 2007 NVS is summarised in Table 1. As TRA does not publish estimates for the total number of trips wholly contained within individual states,

[1] Tourism Australia provided data on trips undertaken in the one day as a separate file on the understanding that sample sizes are relatively small. This data is normally not released.

[2] In census years, a smaller unit denoted a Census Collection District (CD) is defined for ease of collection and dissemination of census data.

Figure 1. NSW and ACT Tourism Regions and LGA boundaries.

the assumption has been made here that survey trips (within NSW/ACT) translate into estimated total trips at the same rate as they do at a national level. These figures would therefore differ from those obtained by TRA, which are calculated with consideration to other demographic variables such as age, sex and household size.

These trips are reported by Tourism Region OD-pairs, of which there are a total of 225 considered in this study. Whilst a substantial number of trips are available for the car mode, other modes, particularly air, provide comparatively few observations. Surveyed trips are not uniformly distributed across OD-pairs, with those involving a state capital (Sydney or Canberra) as an origin or a destination much more often represented in the data. Particularly for the non-car modes, numerous OD-pairs exist where the number of trips observed (<40) is smaller than what TRA considers necessary to draw statistically significant conclusions; however at this stage these observations have not been removed from the data set. There are also OD-pairs where no observed trip data is available and cannot be used in the estimation of models. The number of OD-pairs for which data exists was: car, 218; bus, 113; air, 86; and other, 105.

2.2 Aggregation bias and the ecological fallacy

An ecology fallacy is committed when conclusions about subsets of a population (or an area) are made on the basis of information about the whole population (or an aggregate area). The use of prediction models estimated on aggregate geospatial areas to make predictions about disaggregate areas immediately raises the possibility of having committed such a fallacy. The implication of this is that any such predictions made may contain an amount of aggregation bias due to unobserved heterogeneity at the disaggregate level. It is not possible to test for aggregation bias without accurate data at the disaggregate level for the dependent variable.

Table 1. NVS 2007 statistics - Australia-wide and intra-NSW/ACT

	Overnight	Day Trips
Total survey respondents reporting travel:	**38,633**	**20,264**
Total survey trips (Australia-wide):	**73,750**	**147,737**
Estimated total trips (Australia-wide):	**73,800,000**	**147,700,000**
Intra-NSW/ACT survey respondents (total)[3]:	**9,992**	**5,961**
- by car	8,804	5,537
- by bus	268	186
- by plane	432	28
- by other	517	390
Intra-NSW/ACT survey trips (total)[3]:	**20,098**	**47,718**
- by car	17,816	43,068
- by bus	553	1,409
- by plane	847	239
- by other	952	3,310
Estimated intra-NSW/ACT total trips:	**20,111,625**	**47,706,049**

Paradoxically, if such data were universally available it would negate the need to make disaggregate-level predictions and similarly the need to test for aggregation bias.

It can be readily seen that LGAs in NSW/ACT are not homogeneous insofar as their demography, geography or commercial activities are concerned. To what extent these differences produce heterogeneity in the consumption of long distance transport is difficult to quantify without data from comprehensive travel surveys covering multiple disparate geographic regions. The use of a single model (for each mode) when there are multiple TRs within NSW means that even making predictions about the aggregate regions assumes a degree of homogeneity between these aggregate geographic regions. At the aggregate level it is possible to quantify the amount of variance accounted for by the model as the actual trip values are known. This information may provide some guidance as to the reliability of the model at the disaggregate level.

Whilst it is necessary to consider the theoretical possibility of aggregation bias in the chosen methodology, this must be weighed against more practical considerations, such as the cost of undertaking large-scale travel surveys. This study should therefore be considered a first-pass attempt to produce a highly-disaggregate description of long distance travel within NSW and the ACT[4], with a view to future model refinement and improvement as further information and resources become available.

2.3 Model Estimation

Models for trips between TRs were estimated for all four modes. The available explanatory variables included a range of demographic data: population (individuals and households), household size, average personal income, four household income category variables, and number of vehicles per household categories. These variables were considered for both origin and destination TRs. Inter- and intra-zone distances were also available. Inter-zone distances were calculated from the population weighted centroids of each TR. As this definition would lead to an unrealistic (zero) distance for trips between the same TRs, intra-zonal distance was

[3] Individual values do not sum exactly due to multi-mode respondents.
[4] The most recent National Travel Survey was conducted in 1976, over 32 years ago.

Table 2. Descriptive profile of the explanatory variables

Variable	Mean	Median	Std Dev.	Min.	Max.
Distance (km)	385.41	328.37	234.52	23.16	1032.21
Population	457,487	237,545	919,941	33,791	3,747,179
Num. Households	163,007	87,017	314,533	12,335	1,285,363
Avg. Household Size	2.71	2.70	0.08	2.59	2.92
Median Household Income	$913.72	$849.83	$224.72	$685.75	$1534.26
% HHs Income < $500	26.14%	26.76%	5.74%	11.72%	34.45%
% HHs $500 < Income < $1000	28.06%	29.09%	3.77%	18.48%	33.09%
% HHs $1000 < Income < $2000	31.38%	32.14%	2.48%	26.66%	35.59%
% HHs Income > $2000	14.42%	11.35%	7.49%	7.50%	34.21%
% HHs 0 vehicles	9.69%	9.48%	2.03%	7.06%	14.04%
% HHs 1 vehicles	39.04%	38.78%	2.93%	35.26%	44.35%
% HHs 2 vehicles	36.21%	36.48%	2.15%	32.46%	39.03%
% HHs 3 vehicles	10.52%	10.42%	1.34%	8.82%	12.58%
% HHs > 3 vehicles	4.55%	4.38%	0.91%	3.29%	6.22%

instead defined to be the 'radius' of the zone, assuming that the zone was circular. Finally, dummy variables were introduced to account for trips where the origin and/or destination was Sydney, due to the disparity between this TR, which is metropolitan, and the other TRs, which are all regional/rural. A descriptive profile of the explanatory variables (at the TR level) is shown in Table 2. The variables "*Pop*", "*Dist*", "*HSize*" and "*Syd*" refer to zone populations, distance, household size and the Sydney dummy respectively; the subscripts "*O*" and "*D*" denote origin and destination respectively.

Models were estimated using LIMDEP v9.0 and the final model specifications are shown in Table 3. Origin and destination populations feature in all models with positive coefficients as expected. The use of double log models permits direct interpretation of the elasticities for the explanatory variables. Distance exhibits negative coefficients for both the road-based modes, with car trips having a greater sensitivity to distance than bus trips. Distance occurs as linear and quadratic terms in the Air and Other models, but with opposite signs. The result of this is that air travel is more appealing for medium-range travel (~500km); conversely, train (other) travel is least appealing over these distances. Anecdotal evidence suggests that air travel within NSW over longer distances would involve non-trunk routes and often a transfer to a smaller aircraft, which is likely to be a disincentive. The presence of household size in the car model is explained by the cost effectiveness to larger families in using a private motor vehicle in comparison with other modes. Finally, the reduced travel by car and increased travel by train for travel wholly within Sydney (as evidenced by the presence of the Sydney dummy), can be explained by the presence of a comprehensive urban transport network within this region that is not present in the regional and rural TRs.

Plots of predicted versus actual trips, along with adjusted R-squared values of the estimated models, are shown in Figure 2.

Table 3. Final model specifications by mode

$\ln(Trips_{Car}) = -5.7526+0.68669\ln(Pop_O)+0.53150\ln(Pop_D)-1.5038\ln(Dist_{OD})+3.3058\ln(HSize_O)-1.193Syd_{OD}$

$\ln(Trips_{Bus}) = -0.4232+0.41396\ln(Pop_O)+0.18647\ln(Pop_D)-0.97825\ln(Dist_{OD})$

$\ln(Trips_{Air}) = -15.648+0.60768\ln(Pop_O)+0.64014\ln(Pop_D)+0.00523Dist_{OD}-0.40981E\text{-}05Dist_{OD}^2$

$\ln(Trips_{Other}) = -13.1897+0.63759\ln(Pop_O)+0.69911\ln(Pop_O)-0.01124Dist_{OD}+0.10630E\text{-}04Dist_{OD}^2+0.95811*Syd_{OD}$

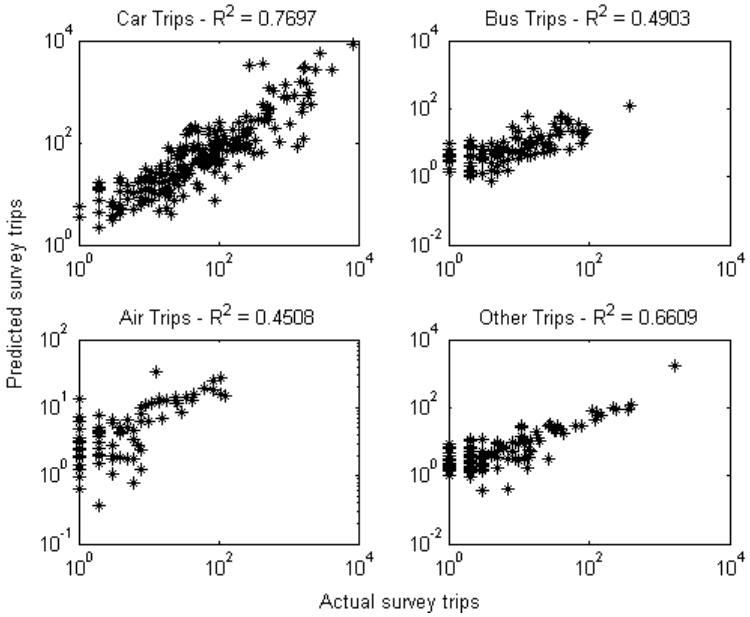

Figure 2. Predicted vs Actual trips (by mode) at TR-level

2.4 Model Application

The models estimated previously were subsequently applied at the LGA-level to obtain LGA-LGA trip estimates. All data items from the TR-level were similarly available at the LGA-level, with inter-/intra-TR distances being replaced by analogous inter-/intra-LGA distances.

A problem was immediately apparent with this approach, owing to the substantial size and proximity differences between TRs and (mainly urban) LGAs. The negative distance coefficient for all land-based modes, coupled with the significantly lower distances caused by the change from TR- to LGA-level distances, led to predictions of an excessive number of trips within and between small and close LGAs. Whilst all models suffer from this problem, it was most apparent for the car model due to the much higher number of car trips than for any other mode. Figure 3 shows the scale differences between this variable at the two levels. Close LGA pairs can be seen to exhibit distances more than two orders of magnitude smaller than their corresponding tourism regions.

Table 4. Model coefficient t-ratios

	$\mathbf{Trips_{Car}}$	$\mathbf{Trips_{Bus}}$	$\mathbf{Trips_{Air}}$	$\mathbf{Trips_{Other}}$
Constant	**-2.396**	-0.236	**-6.676**	**-5.872**
$\ln(Pop_O)$	**11.072**	**4.933**	**6.022**	**6.100**
$\ln(Pop_D)$	**10.893**	**2.317**	**5.951**	**7.336**
$Dist_{OD}$	–	–	**2.906**	**-6.845**
$Dist_{OD}^2$	–	–	**-2.265**	**5.701**
$\ln(Dist_{OD})$	**-14.207**	**-9.119**	–	–
$\ln(HSize_O)$	1.442	–	–	–
Syd_{OD}	**-4.8**	–	–	**2.112**

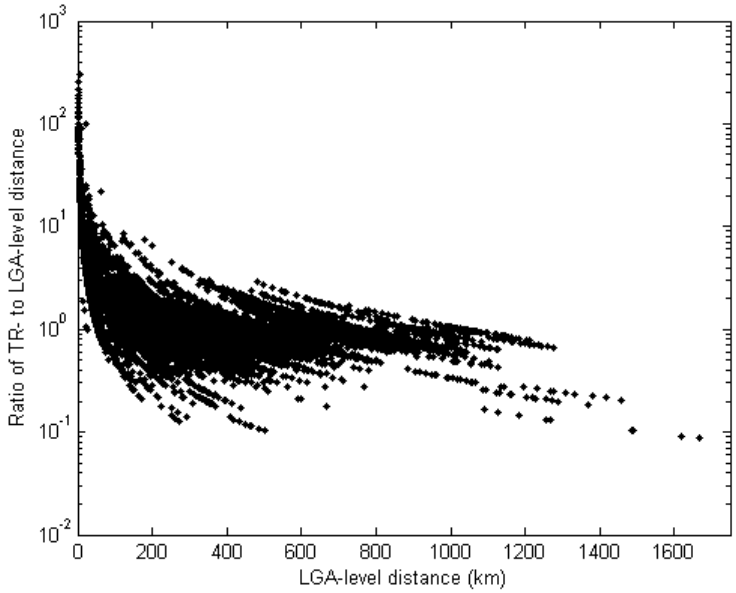

Figure 3. Ratio of applicable TR- to LGA-level distance for all LGA-LGA pairs

In order to resolve this, LGA-level distances were replaced by the applicable TR-TR distances for the TRs containing the origin and destination LGAs. This is of little practical significance for widely separated LGAs, but for small or close LGAs it provides a measure of distance in line with that originally used to estimate the models. Whilst this serves to correct the extreme over-predictions caused by small and/or close LGAs, it must be noted that values below the unity (10^0) line demonstrate that this method will also increase the predicted number of trips between some LGA-LGA pairs. Following this adjustment, the models were re-applied at the LGA level in order to obtain LGA-LGA trip estimates for each mode.

Ideally, predictions made at the LGA-level, when aggregated, should reproduce the estimates observed at the TR-level. As Figure 4 shows, this is not the case, with most TR-TR pairs exhibiting a greater predicted number of trips when the prediction is produced by aggregating up from predictions made at the LGA level. Ortúzar and Willumsen (1994) discusses the process of *naïve aggregation* in order to make predictions at a higher aggregate level from a disaggregate model. Although linear models applied at different levels of aggregation will produce equivalent results, naïve aggregation involving a non-linear model will usually introduce *aggregation error*[5], resulting in different predictions at different levels of aggregation. The method being used herein is analogous to, but the reverse of, such a process and suffers from identical problems.

The principal source of this error is the use of logarithms of the population variables in the model equations. Although distance and household size are also present as logarithmic terms, their values at the aggregate level are not calculated by summation of disaggregate level values (unlike the population measures). The use of TR-level distances at the LGA level and the fact that household size is an average indicates that the values of these variables are similar at both the TR and LGA levels. Unfortunately, linear models were found to give substantially worse model fits than those containing the logarithmic terms.

[5] The term "aggregation error" is used here rather than Ortúzar and Willumsen's "aggregation bias", to distinguish this error from that occurring in disaggregate estimates made from a heterogeneous aggregate population (the so-called ecological fallacy).

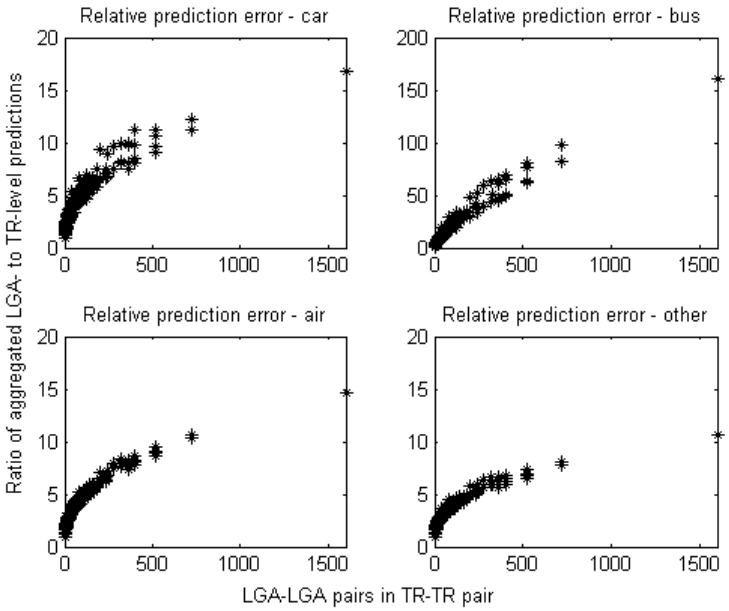

Figure 4. Ratios of aggregated LGA-LGA to TR-TR trip predictions

Consider the differences in the equations for trip predictions at the different levels of aggregation (without loss of generality, a function of only one independent variable is considered, for clarity of notation):

$$y_{agg} = \exp\ \beta \ln \sum x_i$$

$$y_{disagg} = \sum_{i=1}^{n} \exp\ \beta \ln x_i$$

These lead respectively to:

$$y_{agg} = x_1^{\beta} \cdot \prod_{i=2}^{n} \left(1 + \frac{x_i}{\sum_{j=1}^{i-1} x_j} \right)^{\beta}$$

$$y_{disagg} = x_1^{\beta} \cdot \prod_{i=2}^{n} \left(1 + \frac{x_i^{\beta}}{\sum_{j=1}^{i-1} x_j^{\beta}} \right)$$

In order to explain why predictions made at the disaggregate level and summed result in a larger number of predicted trips, it is necessary to show why $y_{agg} \le y_{disagg}$. The full proof has been omitted for brevity, but essentially if $0 < \beta < 1$ (as is the case for all coefficients of the population variables in the estimated models) exponentiating by β causes both a reduction in the value of the base $\left(1 + x_i \Big/ \sum_{j=1}^{i-1} x_j \right)^{\beta} \le 1 + x_i \Big/ \sum_{j=1}^{i-1} x_j$, and larger (aggregate) values to grow at a smaller rate than smaller (disaggregate) values. As the x_i values can be considered to be arranged in descending order (without loss of generality), the following inequality holds:

$x_{i+1}/x_i \le x_{i+1}^{\beta}/x_i^{\beta}$. A proof by induction leads to $1+x_i \Big/ \sum_{j=1}^{i-1} x_j \le 1+x_i^{\beta} \Big/ \sum_{j=1}^{i-1} x_j{}^{\beta}$ and therefore $y_{agg} \le y_{disagg}$.

A prediction made at the disaggregate level and summed will therefore always be larger than the prediction made at the aggregate level, explaining the errors of scale observed in Figure 4. Whilst the exact magnitude of the error depends on the specific model equation, smaller values of β will lead to a larger error. This is caused by the compounding effect of the outer β exponent in the aggregate model. For example, note that the scale of the error for bus trip predictions (Figure 4) is significantly higher than for other three modes as it has smaller coefficients on its population parameters.

In order to address this error, it is necessary to constrain the total trips predicted at the LGA-level to either the actual or predicted trips at the TR-level. Any missing observations in the actual trip data will preclude scaling to actual trip counts. Scaling is a simple process of linear reduction to recreate the TR-level predicted (or actual) values:

$$trips'_{LGA_O LGA_D} = \frac{trips_{LGA_O LGA_D} trips_{TR_{\hat{O}} TR_{\hat{D}}}}{\sum_{i \in \hat{O}} \sum_{j \in \hat{D}} trips_{LGA_i LGA_j}}$$

One way of looking at this process is to consider that trip generation is occurring at the aggregate level but trip distribution is occurring at the disaggregate level. The final LGA by LGA trips data for each model is summarized in Table 5. Although some of the values presented in Table 5 may seem rather low, it must be remembered that the estimated trips are being distributed over the full 154x154 matrix of OD-pairs, many of which would have few, if any, tourism movements.

Table 5. Descriptive statistics of annual LGA-LGA level trip predictions

Mode	Mean	Median	Std Dev.	Min.	Max.
Car	2542.75	335.11	25291.55	5.906	3255531.00
Bus	65.01	23.91	416.07	1.66	52290.65
Air	35.18	18.38	66.71	0.46	1984.95
Other	142.34	19.57	431.83	0.33	32089.28

These annualised trip rates by mode for each LGA-LGA pair provide the best estimates of the dependent variables for a next stage model development at the LGA level. This was the only missing data in our full matrix of LGA by LGA trip, origin, destination, and person characteristics.

3 CONCLUSIONS AND FUTURE WORK

This paper has described a method by which travel behaviour may be predicted at a highly disaggregate level from more aggregate survey data. Trip prediction models estimated at an aggregate geographic level were subsequently applied at a disaggregate level in order to predict trip counts at that level to overcome the need to obtain survey data at that level.

A number of adjustments were shown to be necessary in order to obtain 'realistic' trip estimates at the disaggregate level. Firstly, variables that change significantly in scale between the two levels (in this case, distance) need to be corrected. The solution employed here was to use the aggregate level distances at the disaggregate level. Secondly, the prediction differences due to a non-linear model system were discussed and it was shown that the presence of logarithmic terms in the model will always lead to higher predictions if calculated at a more disaggregate level (subject to the coefficients of the explanatory variables). This difference was eliminated by constraining disaggregate level trip estimates to aggregate level predictions.

One method of validating the performance of this method would be to conduct a disaggregate level travel survey of a number of aggregate zones to permit comparison of actual trip counts with predicted values. This remains an object for future work.

The resulting predictions of trips between each LGA pair for each of car, bus, train and plane, can be used to develop a system of multi-modal models in which the production and attractions characteristics together with accessibility descriptors are explanatory variables. In ongoing research, we have used the prediction LGA to LGA trip matrices to develop empirical models of modal activity and embedded these within a new transport and environmental strategy impact simulator for regional NSW called R-Tresis (Hensher et al. 2008).

REFERENCES

Australian Bureau of Statistics (ABS, 2008), "Australian Standard Geographical Classification", Cat. no. 1216.0, ABS, Canberra.

Daly, A. (2008) *National models*, in Hensher, D.A. and Button, K.J. (eds.) *Handbook of Transport Modelling*, 2nd ed., Elsevier, Oxford, 489-502.

Hensher, D.A., Bain, S. and Li, Z. (2008) R-Tresis: Developing a Demand and Supply Modelling Capability for an Integrated Transport and Land Use Model System for Regional New South Wales, *Institute of Transport and Logistics Study*, University of Sydney, April.

Ortúzar, J. de. and Willumsen, L. (1994) *Modelling Transport*, 2nd ed., John Wiley & Sons, New York.

Tourism Research Australia (2008) *Forecast – 2008*, Issue 2, Tourism Forecasting Committee, Tourism Research Australia, Canberra.

Wegener, M. (2004) *Overview of Land-Use Transport Models*, in David A. Hensher and Kenneth Button (eds.): *Transport Geography and Spatial Systems*, Pergamon/Elsevier Science, Oxford, Ch. 9, 127-146.