Developing disaggregate transport prediction models from aggregate survey data

Stuart Bain and David A. Hensher

Institute of Transport and Logistics Studies
University of Sydney
Australia

Australasian Transport Research Forum
Auckland, NZ, September 2009
Motivation

• Need for demand-side information on passenger movements for construction of transport models
• Information often not centrally-recorded or is not publicly available
• Travel surveys become the main method to obtain such information
• A need for OD-data means data requirements increase quadratically with the number of zones
Challenges of secondary data

• Potential solution is to rely on secondary data source
• Such data may employ a different geospatial partitioning to what is required

• This study considers the former case, where the best available data is a secondary source collected at a more aggregate level than what is needed.
The National Visitor Survey

• Secondary source is the National Visitor Survey
• Conducted annually by Tourism Research Australia
• 120,000 Australians surveyed via random digit dialing and CATI
• “Tourism movements” are defined to include holidaying, visiting friends/family, and travel for business or education
• Collected in two parts: trips involving an overnight stay and day trips
The National Visitor Survey (cont)

- 5 different modes of transport are considered: private car, rental car, bus, plane and other (train)
- Data from the survey is collected at the Tourism Region level
- There are a total of 15 Tourism Regions within NSW and the ACT
Map of NSW/ACT Tourism Regions
The National Visitor Survey (cont)

• Total of 225 OD-pairs
• OD-pair data availability for each mode was: Car, 218; Bus, 113; Plane, 86; Train, 105
• Estimated intra NSW/ACT trips: 67.8M annually
• Car, 89.8%; Bus, 2.9%; Plane, 1.6%; Train, 6.3%

• Using this data source we proceeded to estimate regression models for the 4 separate modes
Explanatory variables

• Inter/intra zone distances
• Demographics
  – Total population, Total households, Avg. household size
  – Median household income, percentage of households in 4 income brackets: <$500, $500-$1000, $1000-$2000, >$2000
  – Percentage households in 5 vehicle ownership brackets: 0, 1, 2, 3 and >3 vehicles
• All demographics obtained from 2006 ABS Census
Model estimation

- Started with double-log models, estimated using LIMDEP 9.0

<table>
<thead>
<tr>
<th>ln(Trips)</th>
<th>Constant</th>
<th>ln(PopO)</th>
<th>ln(PopD)</th>
<th>ln(DistOD)</th>
<th>DistOD</th>
<th>DistOD²</th>
<th>ln(HHSizeO)</th>
<th>SydOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>-5.7526</td>
<td>0.68669</td>
<td>0.53150</td>
<td>-1.50380</td>
<td>-</td>
<td>-</td>
<td>3.3058</td>
<td>-1.1930</td>
</tr>
<tr>
<td>Bus</td>
<td>-0.4232</td>
<td>0.41396</td>
<td>0.18647</td>
<td>-0.97825</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Air</td>
<td>-15.6480</td>
<td>0.60768</td>
<td>0.64014</td>
<td>-</td>
<td>0.00523</td>
<td>-0.04098E-04</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other</td>
<td>-13.1897</td>
<td>0.63759</td>
<td>0.69911</td>
<td>-0.01124</td>
<td>0.10630E-04</td>
<td>-</td>
<td>0.95811</td>
<td></td>
</tr>
</tbody>
</table>

1. Population parameters positive
2. Distance parameters negative road modes
3. Interpretation of distance coefficients air/other is that air is preferred over medium distances; other is more preferred over short/long distances
4. Household size present in car due to increased cost effectiveness
5. Reduced travel by car and increased travel by train in Sydney potentially explained by the presence of comprehensive urban transportation
Model validation

- Car Trips - $R^2 = 0.7697$
- Bus Trips - $R^2 = 0.4903$
- Air Trips - $R^2 = 0.4508$
- Other Trips - $R^2 = 0.6609$
Model application

• All explanatory variables were available at LGA level
• Inter/intra-TR distances were replaced with inter/intra-LGA distances

• Problem with over prediction of trips between small and close LGAs
Model application
Model application

• All explanatory variables were available at LGA level
• Inter/intra-TR distances were replaced with inter/intra-LGA distances

• Problem with over prediction of trips between small and close LGAs
• Resolved by replacing LGA level distances by the appropriate TR level distances
Aggregation error

- Ideally, predictions made at the LGA level would aggregate up to match predictions made at the TR level.
- This would be the case in a fully linear model.
- Not referring to “aggregation bias” caused by population heterogeneity.
Aggregation error
Aggregation error

• Ideally, predictions made at the LGA level would aggregate up to predictions made at the TR level. All explanatory variables were available at LGA level.

• This would be the case in a fully linear model.

• Not referring to “aggregation bias” caused by population heterogeneity.

• Main cause of this problem in this instance is the non-linearity in the population variables.
Aggregation error (proof 1)

• Consider the functional form of the model:

\[ y_{agg} = \exp(\beta \ln \sum x_i) \quad y_{disagg} = \sum_{i=1}^{n} \exp(\beta \ln x_i) \]

• These lead respectively to:

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

• Possible to show:

\[ \left( 1 + x_i / \sum_{j=1}^{i-1} x_j \right)^\beta \leq 1 + x_i / \sum_{j=1}^{i-1} x_j \quad \text{and} \quad 1 + x_i / \sum_{j=1}^{i-1} x_j \leq 1 + x_i^\beta / \sum_{j=1}^{i-1} x_j^\beta \]

• Meaning that:

\[ y_{agg} \leq y_{disagg} \]
Addressing aggregation error

• To address this total trips at the disaggregate level must be constrained to those at the aggregate level
• Simple matter of scaling the number of trips at the disaggregate level so that if they are aggregated they reproduce the aggregate level predictions

• One way of looking at this is to view trip generation as occurring at the aggregate level, but trip distribution at the disaggregate level
### Results

<table>
<thead>
<tr>
<th>Mode</th>
<th>Mean</th>
<th>Median</th>
<th>Std Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>2542.75</td>
<td>335.11</td>
<td>25291.55</td>
<td>5.90</td>
<td>3255531.00</td>
</tr>
<tr>
<td>Bus</td>
<td>65.01</td>
<td>23.91</td>
<td>416.07</td>
<td>1.66</td>
<td>52290.65</td>
</tr>
<tr>
<td>Plane</td>
<td>35.18</td>
<td>18.38</td>
<td>66.71</td>
<td>0.46</td>
<td>1984.95</td>
</tr>
<tr>
<td>Other</td>
<td>142.34</td>
<td>19.57</td>
<td>431.83</td>
<td>0.33</td>
<td>32089.28</td>
</tr>
</tbody>
</table>

- The full matrix (154x154) of these predicted OD trip values can now be used as the dependent variables for the next stage of model development.
- Validation is difficult however.
Ecological fallacy

- Potential problem: are the disaggregate zones sufficiently homogenous w.r.t. each other and the aggregate zones?
- Paradox: impossible to test for aggregation bias without accurate disaggregate level data
- LGAs are clearly not homogenous – to what extent does this matter?
- Introduction of aggregation bias must be weighed against the costs of large scale travel surveys
Future work

• Validation of this methodology in a different context
• Further validation of the model as more and better data becomes available

• Use of the model predictions as the dependent variables in a Regional Transport/Environment Strategy Impact Simulator: R-TRESIS
Conclusions

- Possible to predict disaggregate level transport movements from aggregate, secondary survey data
- ... but, more validation is needed
- Have identified that such an approach is susceptible to large changes in the scale of variables between levels
- Have theoretically shown that disaggregate level predictions always meet or exceed aggregate level predictions
Any questions?