Reducing Garbage-In...for Discrete Choice Model Estimation

David Kurth*
Cambridge Systematics, Inc.
999 18th Street, Suite 3000
Denver, CO 80202
P: 303-357-4661 F: 303-446-9111
dkurth@camsys.com

Marty Milkovits
Dan Tempesta
Cambridge Systematics, Inc.
100 CambridgePark Drive, Suite 400
Cambridge, MA 02140
P: 617-354-0167 F: 617-354-1542
mmilkovits@camsys.com
dtempesta@camsys.com

Jason Lemp
Cambridge Systematics, Inc.
9015 Mountain Ridge Drive, Suite 210
Austin, TX 78759
P: 512-691-8501 F: 512-691-3289
jlemp@camsys.com

Pat Coleman
AECOM
2101 Wilson Blvd, Suite 800
Arlington, VA 22201
P: 703-340-3036 F:703-340-3101
patrick.coogan@aecom.com

* Corresponding author

Words (including references): 2,247
Tables: 2 × 250= 500
Figures: 1 × 250= 250
Total Equivalent Words 2,997

Submitted for Presentation at the Innovations in Travel Modeling 2012 Conference
April 30 – May 2, 2012, Tampa, Florida
Introduction

Estimation of discrete choice models requires information on the alternatives available to individual travelers and the choices those travelers make in response to the characteristics of the alternatives available. Estimation datasets for many discrete choice models (e.g. mode choice and destination choice models) used in trip-based models (TBMs) and activity-based models (ABMs) are constructed by posting modeled auto and transit skim information on survey data for individual travelers. Good model development and model validation practice dictate that the roadway and transit networks and network skimming processes be validated prior to the marrying of the skim data with the choice data.

The Travel Model Validation and Reasonableness Checking Manual – Second Edition <1> suggests several aggregate checks for the reasonableness of auto and transit skims. The checks focus on the inspection of frequency distributions of the variables for the mode alternatives used in the modeling process – in-vehicle travel time, distance, cost, implied interchange speed, number of transfers, out-of-vehicle time, etc. – for outliers. While such tests are important, they do not provide information regarding the veracity of modeled interchange travel impedances with the alternatives actually faced by the individual travelers.

The validation manual also suggests a disaggregate check for transit skims based on the comparison of reported boardings by individual transit travelers surveyed in a home-interview or on-board survey to modeled boardings for the same interchange. Such an effort was used by the Denver Regional Council of Governments (DRCOG) to validate the transit networks used for the development of their ABM <2> and by the North Central Texas Council of Governments (NCTCOG) to calibrate the path-building parameters used in their transit skimming process <3>. Such disaggregate checks do, in fact, provide information regarding the veracity of modeled transit interchange travel impedances for the individual travelers included in the estimation dataset.

This paper reports on the use of prediction-success tables of modeled versus reported boardings for individual travelers for an ABM development project for the Houston-Galveston Area Council (H-GAC) and a TBM update project for the Southeast Michigan Council of Governments (SEMCOG). In both efforts, path-building parameters and procedures were, in effect, calibrated to improve the prediction-success results. The paper also describes a disaggregate validation test of the auto skims used for the H-GAC project. While the auto test was not used to adjust path-building parameters, it was used to identify outliers in the estimation dataset so they could be removed from model estimations that could be negatively impacted by their use.

Transit Network Processing

H-GAC network validation efforts

The H-GAC transit network and path-building validation for use in the ABM estimation was initiated by preparing transit trip tables from Houston Metro’s 2007 on-board survey and assigning them to the transit network. Using Cube’s Public Transport (PT) path-building procedures from HGAC’s current TBM, path-building parameters were refined and the process
was repeated until the modeled boardings reasonably replicated the observed transit boardings from the survey. One primary effort in this initial effort was reducing the number of unassigned trips from about ten percent of the total trips in the initial effort to about one percent with the final path-building parameters.

The focus then shifted to improving boardings by submode to match the survey results. Transit paths were built for the following access mode - submode combinations expected to be modeled in the ABM for each time-of-day (peak or off-peak):

- Walk to local bus
- Walk to express/premium bus
- Walk to rail
- Drive to local bus
- Drive to express/premium bus or rail

In order to eliminate the paths without rail in the rail paths, PT’s MUSTUSEMODE parameter was implemented. This parameter forces the path builder to search for paths until a rail mode is part of the resulting path. This parameter was also added to the express/premium paths to avoid paths using local bus only.

The second step in the validation of the transit network and path-building procedures was the preparation of prediction-success tables of modeled boardings by interchange to reported boardings. To perform this analysis, the modeled numbers of boardings for each interchange and the appropriate time-of-day were posted on individual survey trip records and cross-tabulations of modeled-to-observed boardings for the unexpanded survey records were prepared.

Table 1 summarizes the resulting prediction-success tables aggregated for overall walk and drive access paths. Overall, the modeled boardings matched the reported boardings 59 percent of the time for walk access paths and 72 percent of the time for drive access paths, for an overall prediction-success rate of 62 percent. The prediction-success rates for H-GAC were slightly lower than the result reported in the DRCOG effort (67 percent), albeit with a less complicated path-building procedure. The DRCOG results were based on time-of-day transit paths being built for seven combinations of modes for walk access (local bus, premium bus, rail, local/premium bus, local bus/rail, premium bus/rail, local bus/premium bus/rail).

The results of the prediction-success testing provides some insight for areas that may be investigated to improve the path-building process. Overall, modeled 25 percent of the walk to transit paths had fewer modeled boardings than reported boardings while only 15 percent of the walk access paths had more modeled boardings than reported boardings. One would expect some natural random differences due to detailed locational considerations of individual travelers contrasting with modeled zone-to-zone paths. However, in that case, the rates of underpredicting and overpredicting modeled boardings should tend to be more balanced. Thus, these results suggest that transferring might be slightly over-penalized in the walk access transit path-building.
### Table 1. Transit Boarding Prediction-Success Table – H-GAC ABM Model Development

<table>
<thead>
<tr>
<th>Walk Access</th>
<th>Reported Boardings</th>
<th>Modeled Boardings</th>
<th>Summary Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>No Modeled Paths</td>
<td>1.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported &gt; Modeled</td>
<td>24.8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported &lt; Modeled</td>
<td>15.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported = Modeled</td>
<td>58.7%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drive Access</th>
<th>Reported Boardings</th>
<th>Modeled Boardings</th>
<th>Summary Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>No Modeled Paths</td>
<td>2.1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported &gt; Modeled</td>
<td>11.8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported &lt; Modeled</td>
<td>13.8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported = Modeled</td>
<td>72.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>Reported Boardings</th>
<th>Modeled Boardings</th>
<th>Summary Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>No Modeled Paths</td>
<td>1.4%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported &gt; Modeled</td>
<td>21.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported &lt; Modeled</td>
<td>15.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported = Modeled</td>
<td>62.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

Note: Shaded cells indicate modeled boardings equal reported boardings.

**SEMCOG network validation efforts**

The preparation of prediction-success tables for validating the SEMCOG transit network and path-building parameters varied somewhat from those used for H-GAC since SEMCOG uses TransCAD’s Pathfinder multi-path path-building routine. The multi-path routine determines fractional numbers of boardings for specific interchanges based on the proportions of travelers likely to use each identified transit path. Thus, unlike the shortest-path routine used for H-GAC where individual reported trips for an interchange could be compared to a single, integer number of boardings modeled for the interchange, the SEMCOG prediction-success processing had to determine whether the reported number of boardings matched the boardings for one of the feasible paths found for the interchange.

To implement the process for SEMCOG, true/false tables were built by running the path building process multiple times with “Maximum Number of Transfers” set to zero, one, two, or three:
• The initial set of paths was built restricting the paths to those with zero transfers. If paths could be built, the “one-boarding” matrix was marked as true and the average number of boardings saved in a separate matrix; otherwise, the one-boarding matrix was marked as false.

• The paths were rebuilt restricting the number of transfers to one. If paths could be built for an interchange and the average number of boardings for the interchange was greater than that obtained for the path with zero transfers allowed, then the “two-boarding” matrix was marked as true and the average number of boardings saved; otherwise, the two-boarding matrix was marked as false.

• The process was repeated for maximum number of transfers set to two and three and the “three-boarding” and “four or more-boarding” matrices set accordingly.

• If paths could not be built for any of the reasonable numbers of transfers, the no-transit path matrix was marked as true.

The interchange-transfer true/false tables were joined with the on-board survey data by time-of-day and interchange, and prediction-success was set as true if the reported number of transfers matched one of the true/false tables. If the reported number of transfers was not found in the true/false tables for the interchange, the “no-path” case was associated with the survey record. Prediction-success tables were built for walk access and drive access to bus. No rail transit currently exists in the SEMCOG region.

Prediction-success tables were constructed using the transit path-building parameters specified for the existing SEMCOG travel model. The prediction success rate was 67 percent for walk access paths. On average, the modeled boardings per trip\(^1\) (1.62) were quite similar to the average reported boardings per trip (1.65). However, based on the prediction-success tables, four percent of the reported trips did not have an associated modeled path which indicated that the maximum walk access/egress times allowed in the existing model were too restrictive. Doubling the maximum walk times reduced the no-path trips to two percent and increased the prediction-success to 72 percent, but decreased the average number of boardings per trip to 1.60. To address this issue, the transfer penalty parameter was cut in half to three minutes. Based on this change, the average modeled boardings per trip matched the average reported boardings per trip and, as shown in Table 2, the prediction-success improved slightly to 73 percent. The prediction-success analysis was also instrumental in identifying 391 reported trips (4.4 percent of the reported walk-access trips) with walk access distances greater than five miles. Those reported trips were removed from the model estimation dataset.

---

\(^1\) The average modeled number of boardings per linked trip must be estimated from the path-building process, not from the prediction-success tables, since the prediction-success tables are based on unexpanded survey data.
Table 2. Transit Boarding Prediction-Success Table – SEMCOG TBM Model Development

<table>
<thead>
<tr>
<th>Walk Access</th>
<th>Modeled Boardings</th>
<th>Summary Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reported</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Boardings</td>
<td>1</td>
<td>0.8%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.0%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Shaded cells indicate modeled boardings equal reported boardings.

Highway Network Processing

Substantial effort was expended updating roadway network speeds for H-GAC based on observed, link specific speed data. Based on the effort, it was concluded that reasonable link specific congested and uncongested roadway speeds had been developed for the network. Since there were no current speed run data that could be used to validate selected paths (and be used to infer validity for the rest of the paths), a disaggregate validation test of the veracity of the skimmed auto times was designed.

The validation test consisted of posting modeled interchange travel times on individual surveyed trip records for auto drivers (from the H-GAC household travel survey), calculating the difference between the reported travel time and the modeled travel time, and plotting the distribution of travel times differences. The reported travel time data were derived from the ending time of an activity at one location and the starting time of a subsequent activity at a different location. Of course, the reported starting and ending time data have the well documented reporting issues of rounding error (to the nearest 5, 10, 15, or even 30 minutes on a clock face). While this was, of course, a concern in evaluating individual reported travel times, the impacts of “upward” and “downward” rounding were assumed to cancel each other when comparing large numbers of individual trip records.

The H-GAC household travel survey did not include the parking or accessing of an auto as specific activities. Thus, the derived reported travel times implicitly included terminal times and skimmed travel times had to be adjusted to include terminal times².

Figure 1 shows the distribution of the differences between the skimmed and reported travel times (excluding terminal times). The plot shows smoothed lines drawn through histograms of the frequencies of surveyed auto driver trips aggregated by five minute travel time difference bins. Several conclusions can be drawn from the plot:

² Accounting for these differences provided the ancillary benefit of being able to estimate a simple terminal time model based on the origin and destination zone area types.
• The distributions of the peak and off-peak differences in travel times are similar, suggesting that there is not a systematic error in the estimated congested or uncongested speeds.

• The modes for the peak and off-peak distributions are -2 minutes for the peak and -3 minutes for the off-peak. The medians of the distributions are, likewise, negative. These results suggest that reported travel times tend to be slightly less than modeled travel times. This could be due to modeled speeds being slightly lower than they should be.

• The relative symmetry of the distribution of the differences in travel times suggests that there is not a systematic overestimation or underestimation of interchange travel times. Thus, the differences can, in effect, be considered to be random error.

**Figure 1. Distribution of Difference Between Reported and Skimmed Travel Times**

Based on this validation test, approximately 1,000 auto driver records (approximately three percent of the total auto driver records) with differences in skimmed and reported travel times in excess of ±30 minutes were identified. These records were considered to be outliers resulting from either errors in geocoding of destinations, unreported intermediate stops on a trip, or abnormal incidents affecting the reported travel times. The outlier driver records, along with any accompanying passenger records, were removed from estimation data sets for destination choice and mode choice models.
Discussion

This paper has described two relatively simple disaggregate tests of modeled skim data used for the estimation of travel models. The tests have been used to improve the modeled skim data used for model estimation, eliminate reported travel data with unresolved differences between modeled and reported travel impedances, and to quantify error in modeled data used for model estimation. The increased use of discrete choice models to explain traveler behavior rather than aggregate travel patterns makes it imperative that the travel characteristics that affect the traveler behavior be modeled correctly.

References

