Developing railway station choice models to improve rail industry demand models.

Marcus Young
PhD student, Transportation Research Group
5 January 2017
Outline

• Research background
• Data considerations
• Model results
• Model appraisal
• Conclusions & future work
Research background
A rail renaissance

Rail passenger journeys 1950 - 2015/16 (ORR)

57% growth in last 10 years

1982 = 630

2015-16 = 1687
New stations

- Increasing interest in using rail to meet transport needs or drive economic growth
- Need accurate demand forecasts
Demand models – defining catchments

- Trip end and flow models
  - Must define a catchment first:
    - circular (buffer) around station
    - nearest station – zone based
- Choice of station is deterministic
- Catchments are discrete, none overlapping
Catchments in reality

- 2km circular catchments account for 57% of trips, 0-20% for some stations (Blainey & Evens, 2011)

- 53% of trip ends located within nearest station zone-based catchments (Blainey & Preston, 2010)

- 47% of passengers in the Netherlands do not use their nearest station (Debrezion et al., 2007)
Catchments in reality

• Catchments are not discrete. They overlap and stations compete

• Catchments vary by access mode, station type and destination

• Station choice is more complex than definitions allow
Probability-based catchments

• For each zone calculate the probability of each competing station being chosen

• Allocate zonal population to each station based on the probabilities

• Develop transferable station choice models
Data considerations
Data – passenger surveys

• Welsh Government (WG) 2015
  – South East Wales (Cardiff, Newport, Valleys) and Swansea
  – 7,000 observations

• LATIS, Transport Scotland 2014 & 2015
  – All of Scotland, concentrated in Central Belt
  – 50,000 observations
Data – cleaning and validation

• Missing unit-level postcodes
  – address matching and location estimation
  – OS AddressBase (28 million addresses)

• Automated trip validation
  – excessive access/egress legs
  – illogical trips (reversed or back-track)
## Data – explanatory variables

<table>
<thead>
<tr>
<th>Access journey</th>
<th>Origin station facilities/services</th>
<th>Train leg</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Drive distance(^1)</td>
<td>• Staffing level(^2)</td>
<td>• Total duration(^1)</td>
</tr>
<tr>
<td>• Drive time(^1)</td>
<td>• Service frequency(^5)</td>
<td>• On-train time(^1)</td>
</tr>
<tr>
<td>• Walk time(^1)</td>
<td>• CCTV (y/n)(^2)</td>
<td>• Waiting time(^1)</td>
</tr>
<tr>
<td>• Bus time(^1)</td>
<td>• Car parking spaces(^2)</td>
<td>• No. of transfers(^1)</td>
</tr>
<tr>
<td>• Nearest station (y/n)(^4)</td>
<td>• Free car park (y/n)(^2)</td>
<td>• Fare(^3)</td>
</tr>
<tr>
<td>• Difference in bearing(^4)</td>
<td>• Toilets/Ticket machine(^2)</td>
<td></td>
</tr>
</tbody>
</table>

Data sources:
1. OpenTripPlanner (bespoke route planner - OpenStreetMap/Transit schedules)
2. NRE Knowledgebase XML Feed
3. BR Fares website
4. Derived from data
5. GTFS feed for GB rail services
Data - Automatic processing framework

Models
Model details

• Discrete choice models – RUM theory

• Choice set varies by individual, defined for each origin
  – 10 nearest stations by drive distance plus nearest major station (97% of observed choice)

• Multinomial logit (MNL) and random parameter (mixed) logit (RPL)

<table>
<thead>
<tr>
<th></th>
<th>No. of choice situations</th>
<th>No. of cases</th>
<th>Av. choice set size</th>
</tr>
</thead>
<tbody>
<tr>
<td>LATIS</td>
<td>9,367</td>
<td>97,838</td>
<td>10.44</td>
</tr>
<tr>
<td>WG</td>
<td>5,680</td>
<td>59,833</td>
<td>10.53</td>
</tr>
</tbody>
</table>
## Model results – MNL

<table>
<thead>
<tr>
<th>Model</th>
<th>Significant variables</th>
<th>LogLik</th>
<th>Adj R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>LATIS TE17</td>
<td>• Nearest station&lt;br&gt;• Mode-specific access time&lt;br&gt;• Staffing level&lt;br&gt;• Train frequency&lt;br&gt;• CCTV&lt;br&gt;• Parking space (car mode)&lt;br&gt;• Free car park (car mode)&lt;br&gt;• Ticket machine&lt;br&gt;• Toilets</td>
<td>-6764</td>
<td>.69</td>
</tr>
<tr>
<td>WG TE17</td>
<td>As above</td>
<td>-3733</td>
<td>.72</td>
</tr>
<tr>
<td>LATIS FM2</td>
<td>TE17 plus:&lt;br&gt;• On-train time&lt;br&gt;• Waiting time&lt;br&gt;• Bearing difference</td>
<td>-5243</td>
<td>.76</td>
</tr>
<tr>
<td>WG FM2</td>
<td>As above</td>
<td>-3247</td>
<td>.76</td>
</tr>
</tbody>
</table>
Model results – RPL

Initial models determined that access time parameters had significant standard deviation (using a log normal distribution)

<table>
<thead>
<tr>
<th>Model</th>
<th>Specified as random parameters</th>
<th>Significant non-random parameters</th>
<th>LogLik</th>
<th>Adj R²</th>
</tr>
</thead>
</table>
| LATIS RPL2  | • Mode-specific access time                        | • Nearest station  
• Staffing level  
• Train frequency  
• CCTV  
• Parking space (car mode)  
• Ticket machine  
• Toilets | -6553  | .70   |
| WG RPL2     | • Access time (walk, bus and car modes)             | • Nearest station  
• Cycle access time (cycle mode)  
• Staffing level  
• Train frequency  
• CCTV  
• Parking space (car mode)  
• Free car park (car mode)  
• Ticket machine | -3649  | .73   |
Model appraisal
Predictive performance

• For each station:
  – difference between number of times actually chosen and sum of probabilities from the model

• For entire model
  – absolute difference as % of total choice situations (lower is better)

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictive performance (abs. diff. as % of total choice situations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LATIS</td>
</tr>
<tr>
<td>Base model (prob. nearest = 1)</td>
<td>50.91</td>
</tr>
<tr>
<td>TE17</td>
<td>23.53</td>
</tr>
<tr>
<td>FM4</td>
<td>14.61</td>
</tr>
<tr>
<td>RPL2 (TE)</td>
<td>23.58</td>
</tr>
<tr>
<td>RPL4 (FLOW)</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Predictive performance

LATIS - base model (Pr. Nearest = 1)
Predictive performance

LATIS – FM1 model
Predictive performance

LATIS - base model (Pr. nearest = 1)

LATIS – FM1 model
Transferability

Comparing Flow Station Choice Models

Variable

Parameter

transferability

24
Transferability

Latis - Predictive Performance using WG FM2 (excl. dayfreq)

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictive performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LATIS</td>
<td>50.91</td>
</tr>
<tr>
<td>WG</td>
<td>40.99</td>
</tr>
<tr>
<td>TE17</td>
<td>23.53</td>
</tr>
<tr>
<td>FM4</td>
<td>14.61</td>
</tr>
<tr>
<td>FM2 (model calibrated on other dataset)</td>
<td>20.16</td>
</tr>
<tr>
<td></td>
<td>34.80</td>
</tr>
</tbody>
</table>
## Transferability

### WG - Predictive Performance using Latis FM2

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictive performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LATIS</td>
<td>50.91</td>
</tr>
<tr>
<td>Base model (prob. nearest = 1)</td>
<td>40.99</td>
</tr>
<tr>
<td>TE17</td>
<td>23.53</td>
</tr>
<tr>
<td>FM4</td>
<td>22.35</td>
</tr>
<tr>
<td>FM2 (model calibrated on other dataset)</td>
<td>20.16</td>
</tr>
<tr>
<td></td>
<td><strong>34.80</strong></td>
</tr>
</tbody>
</table>
Conclusions & future work
Conclusions

• Station choice models calibrated using two independent datasets

• Models predict station choice substantially better than a base model that assumes nearest station is always chosen

• Reasonable correspondence in parameter estimates across the two datasets - problematic variables (e.g. CCTV)

• RPL model probably not worth the added complexity, but need to address proportional substitution behaviour of the MNL model
Future work & industry implications

- Main focus to integrate station choice models into trip end and flow rail demand models. When completed this should:
  - improve the models used to assess proposals for new railway stations
  - enable better forecasting of the effects of changing service patterns (e.g. open access services); and
  - provide a methodology that can be incorporated into the industry Passenger Demand Forecasting Handbook
The author wishes to thank the Welsh Government and Transport Scotland for providing passenger survey data; Paul Kelly for permission to use the brfares.com API; and Ordnance Survey Ltd for providing AddressBase data. This work forms part of a PhD funded by EPSRC DTG Grant EP/M50662X/1.