Preferred citation style

Data problems, modelling challenges

KW Axhausen

IVT
ETH
Zürich

June 2015
Data challenges
Do we know the numbers? e.g. daily activities in Switzerland

Tokyo Tech 2015
Do we know the numbers? e.g. drivers licence ownership

![Graph showing the share of car driving licence holders by year of birth and gender.](image-url)
Protocols and response
Surveys, observations are „talk“

Two speakers

- managing their „image“
- staying within the rules of talking
- staying within their socially allocated/identified role
- fulfilling social expectations

Talk and report with/to each other

=>

„Maintaining the willingness of the respondent to report“
Response as a function of response burden @IVT, 2015

- Prior recruitment and incentive
- Prior recruitment, no incentive
- No prior recruitment, no incentive

Response Rate vs. Response Burden Score

Tokyo Tech 2015
Response is a non-random process
Known „error“ generating processes
Activities, movement and traces: A full example record

Home
Out of home
Movement
Phone/SMS
Email
Smart card
GPS
Bluetooth
WiFi
Active/passive tracing: Many owners, locations, quality levels

Phone/SMS
Email
Smart card
GPS
Bluetooth
WiFi
CCTV

Tokyo Tech 2015
Filters imposed/suggested by the study: „Trips“

After “trip” filter:

Tokyo Tech 2015
Filters due to the respondent: Forgetting

Home
Out of home
Movement
After forgetting

Tokyo Tech 2015
Filters imposed by the respondent: Soft non-response

6

12

18

Home

Out of home

Movement

After soft non-response

Home

Out of home

Movement

Tokyo Tech 2015
What is left?

True
5 at home
9 Out of home
26 Stages,
11 trips,
1 subtour,
2 tours

After all processes
3 at home
2 Out of home
4 trips,
2 tours

Tokyo Tech 2015
What happens next?
## Geocoding addresses

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal</td>
<td>Street addresses identifying the entry to the network</td>
</tr>
<tr>
<td>Best-case</td>
<td>Unambiguous street addresses</td>
</tr>
<tr>
<td>State of the art</td>
<td>Street address</td>
</tr>
<tr>
<td>State of practice</td>
<td>Street address/mid-street block/street corners; missing conversion of facility names</td>
</tr>
<tr>
<td>Still seen in practice</td>
<td>Arbitrary zonal centroid, e.g. post offices</td>
</tr>
</tbody>
</table>
### Calculating distances & travel time

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal</td>
<td>Complete GPS track for distance and times with pedestrian-networks added</td>
</tr>
<tr>
<td>Best-case</td>
<td>Minimal gaps, and state-of-the-art imputation of GPS tracks and modes</td>
</tr>
<tr>
<td>State of the art</td>
<td>SUE derived travel times and distances (navigation network)</td>
</tr>
<tr>
<td>State of practice</td>
<td>DUE derived travel times and distances (planning networks)</td>
</tr>
<tr>
<td>Still seen in practice</td>
<td>Shortest path on empty planning networks</td>
</tr>
</tbody>
</table>

Tokyo Tech 2015
What should we do?
Next steps

- Query what we really need for
  - Cost-benefit analysis
  - Planning of prices and services
  - Planning for the slow modes
  - Social accounting

- High-quality multi-modal surveys to establish the measurement errors (add bluetooth and wifi senders, noise profile)
  - Error correction models
  - Cross check against third party sources

- Treat survey data as indicators in a measurement model
  - Treat traces as indicators in a measurement model
, but especially

- Treat respondents as partners in a talk, discussion:
  - Frame your request in a way which addresses them in a clearly defined social role (citizen, driver, customer, etc.)
  - Account for their constraints (readability of text, full guidance through the forms, require no calculations – unless necessary, speak their ‘language’)
  - Be as complex, as the topic warrants, requires, but not more so
  - Don’t surprise them with unannounced requests
  - Don’t ask them to do work you can do

- If appropriate, provide an incentive, acknowledgement
Modelling challenges
<table>
<thead>
<tr>
<th>Modelling challenges: The usual worries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Error heterogeneity</strong></td>
</tr>
<tr>
<td><strong>Spatial correlations</strong></td>
</tr>
<tr>
<td><strong>Temporal correlations</strong></td>
</tr>
<tr>
<td><strong>Independence</strong></td>
</tr>
<tr>
<td><strong>Endogeneity</strong></td>
</tr>
<tr>
<td><strong>Error of the second kind</strong></td>
</tr>
<tr>
<td><strong>Validation</strong></td>
</tr>
<tr>
<td><strong>Substance</strong></td>
</tr>
</tbody>
</table>

Tokyo Tech 2015
Modelling challenges: Substance or t-tests?

Adapted from Zilliak and McCloskey (2008)

Tokyo Tech 2015
Modelling challenges: Substance or t-tests?

Yes, but anything new?

Adapted from Ziliak and McCloskey (2008)
Modelling challenges: Substance or t-tests?

Effect size

large

small

Yes, but anything new?

Fair enough, but do we know why?

Adapted from Zilliak and McCloskey (2008)
Modelling challenges: Substance or t-tests?

- Effect size: small, large
- t-stats: 1.96

OK, but are there surprises?

Yes, but anything new?

Fair enough, but do we know why?

Adapted from Zilliak and McCloskey (2008)

Tokyo Tech 2015
Modelling challenges: Substance or t-tests?

- Effect size
  - large
  - small

1.96

t-stats

- Now, this is interesting
- Yes, but anything new?
- OK, but are there surprises?
- Fair enough, but do we know why?

Adapted from Ziliak and McCloskey (2008)

Tokyo Tech 2015
Modelling challenges: Substance or t-tests?

Effect size

large

small

1.96

t-stats

Now, this is interesting

Yes, but anything new?

OK, but are there surprises?

Fair enough, but do we know why?

Adapted from Ziliak and McCloskey (2008)

Tokyo Tech 2015
Choice modelling challenges
### Choice modelling challenges: The usual worries

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error heterogeneity</td>
<td>Is it always checked?</td>
</tr>
<tr>
<td>Spatial correlations</td>
<td>Are they always checked?</td>
</tr>
<tr>
<td>Independence</td>
<td>Do we check the correlations of the independent variables (sample) thoroughly enough?</td>
</tr>
<tr>
<td>Endogeneity</td>
<td>Do we fully account for it? (sample selection)</td>
</tr>
<tr>
<td>Error of the second kind</td>
<td>Do you calculate it?</td>
</tr>
<tr>
<td>Validation</td>
<td>How often do we ask for out-of-sample tests?</td>
</tr>
<tr>
<td>Substance</td>
<td>or do we talk about t-tests?</td>
</tr>
</tbody>
</table>
### Choice modelling challenges: less usual concerns

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error heterogeneity</td>
<td>Why don’t we check them?</td>
</tr>
<tr>
<td>Number of non-chosen alternatives</td>
<td>How much leverage do they have for your problem?</td>
</tr>
<tr>
<td>Number of choice sets</td>
<td>How stable are our estimates?</td>
</tr>
<tr>
<td>Capacity constraints</td>
<td>Do we check for their impact on the parameters? (attribute values of the known (non)chosen alternatives)</td>
</tr>
<tr>
<td>Unit of analysis</td>
<td>Do we have a MAUP problem?</td>
</tr>
</tbody>
</table>
Residuals: False positives of a membership model

Source: Kopp (Forthcoming)
Residuals: MCDEV model of fleet choice

Source: Jäggi (Forthcoming)

Tokyo Tech 2015
Number of non-chosen alternatives: routes

Source: Schüssler (2010)
Number of choice sets: residential choice

<table>
<thead>
<tr>
<th>MEASUREMENTS</th>
<th>DAT1</th>
<th>ESTIMATES</th>
<th>DAT2</th>
<th>DAT3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household</td>
<td>DAT1</td>
<td>ESTIMATES</td>
<td>DAT2</td>
<td>DAT3</td>
</tr>
<tr>
<td>DIST_PREVLOC</td>
<td>-5.440 **</td>
<td>-7.070 **</td>
<td>-8.740 **</td>
<td></td>
</tr>
<tr>
<td>DIST_WORK</td>
<td>-2.460 *</td>
<td>-3.220 *</td>
<td>-3.880 *</td>
<td></td>
</tr>
<tr>
<td>ETA_PREVLOC</td>
<td>0.192 **</td>
<td>0.163 **</td>
<td>0.135 **</td>
<td></td>
</tr>
<tr>
<td>ETA_WORK</td>
<td>0.218 **</td>
<td>0.203 **</td>
<td>0.166 **</td>
<td></td>
</tr>
<tr>
<td>Accessibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIVACC_CAR</td>
<td>-0.233</td>
<td>-0.302 **</td>
<td>-0.187</td>
<td></td>
</tr>
<tr>
<td>PTACC_NOCAR</td>
<td>0.555 **</td>
<td>0.541 **</td>
<td>0.547 **</td>
<td></td>
</tr>
<tr>
<td>Socioeconomic Environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAME_HH_AGE_SHARE</td>
<td>0.782 **</td>
<td>0.684 **</td>
<td>0.634 *</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.508</td>
<td>0.529</td>
<td>0.524</td>
<td></td>
</tr>
<tr>
<td>adj R²</td>
<td>0.500</td>
<td>0.522</td>
<td>0.517</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from: Schirmer (Forthcoming)

Tokyo Tech 2015
Accounting for consistency
Learning approach of the generic one-day transport model

Competition for slots on networks and in facilities

Activity scheduling

Mental map

\[ k(t,r,j)_{i,n} \]

\[ q_i \equiv (t,r,j)_{i,n} \]
Model estimation: $\beta_{i,o} = \beta_{i,n}$? $\beta_{i,n-1} = \beta_{i,n}$?

- Competition for slots on networks and in facilities
- Activity scheduling
- Mental map
- Parameter estimation

$k(t,r,j)_{i,n}$
$q_i \equiv (t,r,j)_{i,n}$
$\beta_{i,o}$
$\beta_{i,n}$
Model estimation: $\beta_{i,o} = \beta_{i,n}$? Route and mode

Source: Vfrtic (2003)
Do we have a MAUP problem?
Do we have a MAUP-like problem for DCM?

- Location choice, obviously
- Route choice, obviously
- Time-of-day choice, obviously

- But also, mode choice
  - Stage
  - Trip
  - Sub-tour
  - Tour
  - Daily schedule
Swiss national travel diary 2010: Main mode by aggregation
Do we have a MAUP-like problem for DCM?

<table>
<thead>
<tr>
<th></th>
<th>Stage</th>
<th>Trip</th>
<th>Subtour</th>
<th>Tour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value of Time Walking CHF/h</td>
<td>152</td>
<td>28</td>
<td>26</td>
<td>24</td>
</tr>
<tr>
<td>Value of Time Bike CHF/h</td>
<td>194</td>
<td>39</td>
<td>43</td>
<td>40</td>
</tr>
<tr>
<td>Value of Time Car CHF/h</td>
<td>135</td>
<td>25</td>
<td>30</td>
<td>27</td>
</tr>
<tr>
<td>Value of Time PT CHF/h</td>
<td>-30</td>
<td>2</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Value of Time PT access CHF/h</td>
<td>819</td>
<td>15</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>TT PT / TT Car</td>
<td>-</td>
<td>-4.46</td>
<td>12.33</td>
<td>4.07</td>
</tr>
<tr>
<td>TT Walk / Access time PT</td>
<td>-</td>
<td>0.19</td>
<td>1.83</td>
<td>1.19</td>
</tr>
<tr>
<td>Transfer / TT PT min</td>
<td>-220.43</td>
<td>107.00</td>
<td>31.28</td>
<td>32.92</td>
</tr>
<tr>
<td>Interval / TT PT</td>
<td>-</td>
<td>0.96</td>
<td>7.00</td>
<td>3.47</td>
</tr>
<tr>
<td>Access time / TT PT</td>
<td>-</td>
<td>-27.10</td>
<td>7.67</td>
<td>3.02</td>
</tr>
</tbody>
</table>

Source: Schmujtz (2015)
Do we have a MAUP-like problem for DCM?

Source: Schmujtz (2015)
Do we get the time horizon right?
What should we do?
Next steps

- Become more systematic
  - Test for choice set size effects
  - Test for the stability of the estimates \textit{wrt} choice set
  - Test for the stability \textit{wrt} imputation of the attribute values

- Check for the right unit of analysis

- Check for the right set of explanatory variables
Questions?

www.ivt.ethz.ch
References

Jäggi, B. (Forthcoming) Decision modeling on the household level for energy, fleet choice and expenditure, Dissertation, ETH Zürich, Zürich.


Schuessler, N. (2010) Accounting for similarities between alternatives in discrete choice models based on high-resolution observations of transport behaviour, ETH Zürich, Zürich.
