Mixed logit mode choice model on panel data: accounting for intra and inter-personal heterogeneity and different correlation over time periods
Content of the presentation

- Background
- Objectives
- Characteristics of the panel data used
- Mixed Logit Model one panel data
  - Analysis of day-to-day and week-to-week variability
  - Systematic and random variability; correlation among individuals over different periods
  - Analysis of identical observations
- Conclusions
Backgrounds

Demand models are usually estimated using cross sectional data, where single-day characteristics of travel behavior are collected for individuals.

The increased need of understanding variability in travel behavior and the parallel advances in travel demand modeling techniques has recently brought the research community to re-discover panel data.

A panel data is essentially a multi-day data where information are collected at a number of successive points in time retaining the same individuals for the entire series of surveys.

Panel data offer major advantages over cross-sectional data as repeated observations from the same individual generally allows to:

- measure with higher precision changes in individual behaviours
- detect variability across days (true variability apart from sample errors)
Backgrounds

Panel data can be classified into two categories:

- **Long survey panels** (the most common in the literature) consist of repeating the same survey (i.e. with the same methodology and design) at “separate” times, for example once or twice a year for a certain number of years or before-and-after an important event.

- **Short survey panels** consist of multi-day data where repeated measurements on the same sample of units are gathered over a “continuous” period of time (e.g. seven or more successive days), but the survey is not repeated in subsequent years.

In particular:

- **Long survey panels** allow to study dynamic effects over waves, such as habit formation, learning and the reaction to important policies.

- **Short survey panels** allow to detect effects such as rhythms of daily life, to explain current behaviours on the basis of the individuals’ history and experience, to account for intrapersonal and interpersonal variability.
Objetives

Having available a six-week panel dataset, gathered in Karlsruhe (Germany) in 1999, we decided to analyze the effect on modal choice of:

1. intrapersonal variability (the day-to-day and week-to-week variation)
2. interpersonal variability (systematic and random heterogeneity over individual preferences and responses)
3. correlation across individuals over three time periods:
   - the single day
   - the single week
   - the days of a wave (all Monday in the wave, and so on)
   - all the trips made by an individual in the panel (all the 6 waves)
4. repeated observations for some individuals
Panel data used

In the Mobidrive survey:
- 160 families and 360 people were interviewed
- an activity diary was used
- all the trips and activities made in the 6 weeks were collected
  (all days of the week, 24 hours a day)

In our analysis we only concentrated in the working days.
The final sample used has:
- 4089 single tours
- 2488 daily schedules
- 674 weekly schedules
- 129 individual schedules
- 56 households schedules
Panel data: informations available

**Alternatives:** car driver, car passenger, public transport, walking, bicycle

**Level-of-services attributes:**
Total travel time (all the times included)
Trip cost (parking cost included)

**Socio-economic characteristics:**
Age, gender, employment status (part-time, full-time), social status, children less than 18 years old, main car users

**Other characteristics:**
Trip purpose: study, work, leisures (and more)
Zone origin of the trip: urban or sub-urban localization
Number of secondary activities performed during each tour (N. of stops)
Number of km travelled by car in a year
Public Transport seasonal ticket
Time available: 24 hours minus the time spent in previous trips and activities.
1. Intrapersonal variability: distribution of the chosen modes over days and weeks

variation among days (Monday, Tuesday, etc.)

- The histograms represent the average value of the trips made by each mode
- Dots and scores are respectively the maximum and the minimum values over days and weeks
- The bold lines represent the standard deviation

Mode choice is fairly stable over days and weeks

The car driver is the most chosen alternative (39%)
1. Intrapersonal variability: Distribution of the trips purpose over days and weeks

The percentage of tours by purpose is quite stable over days but not over weeks. In particular:

- the percentage of tours for shopping is very low the first week of the survey and increases significantly the following weeks
- the same trend, although less pronounced, occurs for the leisure tours
1. Intrapersonal variability: models results

To analyze the variability of the individual behaviour over the sequence-day, modal choice models were estimated with parameters specific for each day and week.

The typical t-test was used:

\[
t_{\text{test}} = \frac{\hat{\beta}_{\text{Fri}} - \hat{\beta}_{\text{Mon}}}{\sqrt{\text{var}(\hat{\beta}_{\text{Fri}} - \hat{\beta}_{\text{Mon}})}}
\]

Models without systematic and random heterogeneity in preferences for parameters and alternatives:

- Travel time and cost are not significantly different across days and weeks.
  
  Only Friday (the marginal utility of travel time) is significantly different from Wednesday and Thursday (t-test > 2.5), but not from Monday and Tuesday (t-test < 1.2).

- ASCs are highly significant.
2. Interpersonal variability: models results

Systematic and random heterogeneity was then included in modelling estimation.

**Systematic heterogeneity:**

- Travel Time parameters vary significantly depending on several SE characteristics (7 parameterised effects were found highly significant).
- The preference specific for each alternative also varies depending on several SE characteristics (16 effects). => ASCs lost completely significance

**More important, it seems that variability is made clearer:**

- Differences among the first 4 days (Monday to Thursday) reduces.
- Differences between Friday and Mon-Tue-Wed-Thu becomes more significant.
- Week 6 appears significantly different from the previous 5 weeks.
2. Interpersonal variability: models results

Systematic and random heterogeneity was then included in modelling estimation.

Random heterogeneity:

- Significant random heterogeneity was found around Travel Time, Cost and Travel Time specific for Friday.

More important, random heterogeneity varies depending on the time period considered for the correlation across individuals.

Workshop on Discrete Choice Models – Lausanne, August 2008
3. Correlation across individuals

Mixed Logit model for panel data where estimated, with the following utility function:

$$U_{aq}^t = \sum_{k=1,...,K} \left( b_k + \mu_{qk}^t + b_k^{SE} SE_q^t \right) X_{kaj}^t + \sum_m \left( b_m + \mu_{qm}^t + b_m^{SE} SE_q^t \right) y_{jm} + \varepsilon_{aq}^t$$

$$t = \{ \text{the same day, the same week, over individuals and households, the day-of-week} \}$$

- $X_{kaj}^t$: Level-of-services attributes
- $SE_q$: Socio-economics variables
- $b_k$, $b_m$: Parameters fix over population and time periods
- $\mu_{qk}$, $\mu_{qm}$: Individual parameters fixed over time periods and randomly distributed
- $y_{jm}$: Index that equals one if $m$ appears in utility function $j$, zero otherwise, allowing for error component
- $\varepsilon_{aq}^t$: Gumbel distributed random terms
3. Correlation across individuals

Correlation were tested for the following periods:

\[ U_{qj}^t = \sum_{k=1,...,K} \left( b_k + \mu_{qk}^t + b_k^{SE} SE_q \right) X_{kqj}^t + \sum_m \left( b_m + \mu_{qm}^t + b_m^{SE} SE_q \right) y_{jm} + \varepsilon_{qj}^t \]

\( t = \{ \text{the same day, the same week, over individuals and households, the day-of-week} \} \)
3. Correlation across individuals: Results

Including correlation among individuals, results improve enormously.

In particular, given the following correlations:

A. Among tours made within each day
B. Among tours made within each week
C. Among trips made within the whole panel by each individual
D. Among trips made only on Monday and so on across the 6 weeks

Models results improve as we extend the correlation from only within each day (A) to the whole panel (C), which does not surprise.

Maybe more interestingly, when correlation among trips made on Monday (and so on) across the 6 weeks (D) is accounted for:

- Results improve more than accounting for the correlation within each week (B)
- Random heterogeneity on Travel Time specific for Friday disappears.
4. Analysis of repeated trips

An important point in the panel data is the presence of habit or persistence in the choice.

We analyzed repeated observations, defined as tours whose characteristics (time, cost, purpose, mode, and so on) are equal to another tour already performed in a previous day (or previous moment of the same day).

It is important to note that repeated behaviors is not strictly the right measure for habits, as repeated choices might be generated by intentions formed repeatedly.

We investigate whether and to which extent modelling results are affected by the presence of repeated observations in the sample. ML models were estimated:

- Including dummy variables that measure if the tour is equal to a tour previously made and reported in the sample.
- Excluding repeated observations.
4. Distribution of repeated trips

The number of the repeated tours in our sample increases after the first days of survey (Monday of week1), but only up to week 3. Interestingly on week 4 and even more markedly on week 5 the number of repeated choices significantly diminished, to increase again on week 6.

Percentages are computed with respect to the total number of repeated trips over the 6 weeks. They are cumulative.
4. Distribution of repeated trips by purpose

Choices are much more persistent for tours whose main activity is work or study.

As expected, instead, the number of repeated tours for shopping and leisure is highly variable over the sequence-day.

Percentages are computed respect to each day; i.e. all the Mondays over the 6 weeks sum 100%.

Note that the number of the repeated trips can be very low (i.e. the 22% of repeated trips on Monday of week 1 corresponds to only 4 trips)

Workshop on Discrete Choice Models – Lausanne, August 2008
4. Mixed Logit with repeated trips: Results

Including dummy variables that measure if the tour was already made.

We found that when a trip is repeated with the same characteristics:

- There is a propensity of using the same mode, if the trip is made by public transport, by bike or walking.
  (N.B. the effect of seasonal ticket is accounted for with a specific variable).
- But the marginal utility of travel time by bike and walking decreases for repeated trips.

The effect of the correlation across individuals over different periods does not vary:

- Travel time and cost show significant random heterogeneity around the mean
- Travel time parameter specific for Friday is not random distributed.
4. Mixed Logit with repeated trips: Results

To analyze to which extent model results depend on the repeated observations present in the panel dataset, all the specifications were estimated excluding the repeated observations.

Comparing models estimated with and without the repeated observations it seems clear that at least a 50% of the improvement in the overall model statistics are due to the presence of repeated observations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>With repeated observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L(\text{max})$</td>
<td>0,09%</td>
<td>3,62%</td>
<td>12,16%</td>
<td>12,96%</td>
<td>24,92%</td>
<td>-6,85%</td>
</tr>
<tr>
<td>$\text{RHO}^2 (C)$</td>
<td>-0,16%</td>
<td>-5,60%</td>
<td>-14,35%</td>
<td>-13,71%</td>
<td>-23,57%</td>
<td>7,48%</td>
</tr>
<tr>
<td>$L(\text{max})/\text{obs.}$</td>
<td>0,09%</td>
<td>3,62%</td>
<td>12,16%</td>
<td>12,96%</td>
<td>24,92%</td>
<td>-6,85%</td>
</tr>
<tr>
<td>Without repeated observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L(\text{max})$</td>
<td>0,09%</td>
<td>1,41%</td>
<td>3,80%</td>
<td>6,08%</td>
<td>13,09%</td>
<td>-3,01%</td>
</tr>
<tr>
<td>$\text{RHO}^2 (C)$</td>
<td>-3,14%</td>
<td>0,70%</td>
<td>-6,34%</td>
<td>-12,21%</td>
<td>-17,63%</td>
<td>4,37%</td>
</tr>
<tr>
<td>$L(\text{max})/\text{obs.}$</td>
<td>0,09%</td>
<td>1,41%</td>
<td>3,80%</td>
<td>6,08%</td>
<td>13,09%</td>
<td>-3,01%</td>
</tr>
</tbody>
</table>
Conclusions

Using a panel data allowed to analyse the effect on model estimation of intra and interpersonal variability, to test several dimensions of correlation across responses and the effect of repeated observations. We found that:

- Model specification is crucial to detect correctly variability over days and weeks. E.g. not accounting for systematic heterogeneity might mask variability over days and weeks.

- However, only Friday is always significantly different, which suggests that data from different days (at least from Monday to Thursday) might be simply pooled.

- Long survey panel data reveal that individual behaviours are repetitive. Almost 40% of the tours made in the 6 weeks are identical to a previous tour.

- However, repeated tours have only the effect of improving the goodness of fit (only because of bigger samples) but do not increase model explanation. (Evidences with simulated data).
Conclusions

Using a panel data allowed to analyse the effect on model estimation of intra and interpersonal variability, to test several dimensions of correlation across responses and the effect of repeated observations. We found that:

- Long survey panels present several dimensions of correlation across individuals, depending on the time period of reference.

- Models improve significantly as the correlation is extended from the single day to the whole panel. As expected.

- However, it seems to have more correlation among the same day (Monday etc.) over the 6 weeks than among days within each week. This effect is in part (50%) due to the presence of repeated observations.
Thanks