Analysing Multi-day Travel Behaviours with Public Transit Smart Card Transaction Data

Ka Kee Chu and Robert Chapleau, École Polytechnique de Montréal, Canada

Abstract submitted to the 12th International Conference on Travel Behaviour Research

Background
Until recently, analyses on multi-day travel behaviour have long relied on self-reported travel surveys and activity diaries which suffer from data quality problems in the following aspects:

- Short timeframe: self-reported methods require significant participation from the respondents. Therefore, the number of days in a multi-day survey or diary is limited.
- Respondent burden and fatigue: several researches show that reporting rates and accuracy decreases as the survey progresses. In addition, respondents may not be able to complete the whole process.
- Sampling: the commitment from the data collecting party as well as the respondent makes a large sample size prohibitively labour intensive, costly and impractical.
- Lack of temporal and spatial details: self-reported methods are based on recollection of the respondents. The imprecision in time (rounding) and location (not knowing the exact address) in their reporting increases the uncertainties of the findings. Research often shows that significant number of trips are omitted.
- Coding: discrepancy exists in coding due to interviewees or respondents who have different understanding of concepts. Coding and transcribing into digital format are time-consuming and prone to errors, resulting in invalid trip declarations.

Automatic vehicle location (AVL) technologies and electronic fare collecting (EFC) systems, increasingly common among public transit agencies, offer automated data collection on transit trips. Systems with personalized fare media such as the smart cards integrated with global positioning system (GPS) provide a setting to examine multi-day travel behaviour through its boarding transaction records which have the following properties:

- Long timeframe: the systems collect data as long as they are operational.
- Complete geographic coverage: all trips made within the transit network are included.
- Disaggregate data: each transaction is individually recorded along with detailed information on the card, operations and trip.
- Personalized fare media: some types of fare, for example reduced fare, requires the use of a personalized smart card which limits its use to the registered cardholder.
- Passive data collection: no extra task from the part of respondent other than maintaining the fare validation routine. No recalling is required.
- Large scope: trips made by the population of smart cardholders, which usually form the majority of public transit users, are “surveyed”.

• High fidelity on temporal and spatial details: boarding transaction time is recorded to the minute and boarding location to the stop-level.
• Standardized coding: data fields in transaction record follow pre-defined values in the databases.
• Integration with operations data: route, run and vehicle numbers are some of the operations details recorded in each transaction.

Given that, automatically collected data have their own shortcomings:
• Erroneous values in transactions resulting from mis-assigned trip and equipment failure require a complex procedure of data validation using public transit concepts and spatial-temporal logics, as well as imputation based on repetition of planned service and cardholders’ historic boarding pattern.
• Similar to other travel survey done with automatic devices, certain information cannot be gathered by the system without respondent input. This includes trip origin and destination as well as trip purpose. Inference is possible under certain conditions.
• In case of entry-only fare validation, the alighting stop of a trip segment, needs to be estimated through a data enrichment process.

These issues are discussed in previous works by the authors and an elaborated post-processing method is proposed. The resulting data are used to analyse travel behaviour of transit users.

Methodology
Since the data are in a disaggregate form and are organized by object, travel behaviour can be examined by individual, by sub-groups or by the system as a whole. One month of boarding transaction data (about 760,000 records) from a mid-size transit agency are used in the study. They go through the validation and data enrichment processes which include alighting stop estimation, transfer identification and itinerary reconstruction. The current preoccupation is to infer trip details, namely trip origin and destination, trip purpose, activity location and duration (figure 1).
One experiment is to associate students to an educational institution. Students are derived from reduced-fare smart cards and the exercise involves integrating external spatial data into the geographic information system (GIS). A comprehensive list of educational institutions of the city is obtained and for each eligible card, an education institution is assigned based on the multi-day boarding pattern. The boarding locations (excluding transfer) are aggregated by time of day in order to reveal clearer patterns. The institution closest in distance to the most frequently used stop during the afternoon peak will be assigned to the card.

With a major activity space sorted out, it would be possible to infer the residential location to the nearest stop based on other boarding transactions. Other than travel length and trip duration, details such as trip origin, destination, purpose, frequency and regularity for most trips can be estimated. Experiments to associate trip generator to other user groups are also sought (figure 2).
The demand of subgroups such as a student population, users of a particular trip generator or from individual borough, as well as the variation in demand, can be analysed as a unit as any unit can be freely defined with disaggregate data (figure 3).

Figure 2 Intensity of transactions aggregated by half-hour showing the association of boarding activity with trip generators at different time of the day

Figure 3 The day-to-day variation of demand by time of day (shown in different colours) of students from a particular educational institution.

A reference demand of an average weekday can be synthesised from one month of data or more. The variation with respect to this reference is valuable for analysing day-to-day, weekly, monthly and seasonal travel patterns.
Conclusion
This paper shows that public transit smart card data are suitable to analyse multi-day travel behaviour and the complexity and rigour of the methodology required to perform reliable analyses with the data. Without a data post-processing methodology for validation and enrichment, analysis results would be flawed. The paper also demonstrates techniques to infer trip details useful to travel behavioural studies. Demand synthesised from an extended time period would serve as a reference to analyse different scales of variability in demand.