Validating travel behavior estimated from smartcard data
Marcela Munizaga*, Flavio Devillaine, Claudio Navarrete, Diego Silva
Departamento de Ingeniería Civil,
Universidad de Chile, Santiago, Chile
*Corresponding Author
e-mail: mamuniza@ing.uchile.cl

Word count: 4393 words, 3 tables, 4 figures
Total: 6143

1. Introduction

To obtain travel information from smartcard data is a growing tendency. Many researchers have foreseen the opportunity of obtaining high quality information at very low cost, and therefore have developed tools to obtain valuable information from the original data generated as a side-product of the operation of transit systems. However, most authors recognize that the process of estimating destination, route and activities from smartcard data involves several assumptions, and the results need to be validated. We have conducted such validation for the set of methodologies used to input boarding position, alighting stop and route chosen for the case of Transantiago (public transport system in Santiago, Chile). To do so, we have used three sources of information: information from the same database used to make the estimations (endogenous validation), information from a detailed origin-destination survey applied to a sample of 300,000 metro users, where a small percentage of them provided their smartcard id for validation purposes, and personal interviews to a small sample of volunteers who provided all their travel and personal information. The endogenous validation was mainly used to detect errors and improve the methodology. The large sample of Origin-Destination (OD) surveys was used to compare the route chosen by the users within the Metro network with the route assigned by the model (minimum cost). The small sample of metro users who answer the OD survey and also provided their card id, and the sample of volunteers were used to validate the more strong assumptions of the model: alighting estimation and purpose assignment. The results of this exploratory validation analysis are very promising. The endogenous validation showed very reasonable estimations, and helped to detect some errors that can be easily corrected with slight modifications to the previous methodology. Both the OD survey and the personal interviews showed that the estimation of the boarding site is very precise, and that the estimation of alighting stop is reasonably good in estimating the exact alighting site, and reach levels over 95% of accuracy at a more aggregate level.

Description of the methodology to be validated
Munizaga and Palma (2012) propose a method to observe card transactions in the public transport system, and estimate the travel sequence using information of the transactions sequence. Only boarding transactions are observed, because the payment
system does not require validation when alighting. The assumptions that Munizaga and Palma (2012) make to estimate a trip matrix are:

- Trip stages begin at the time/location when/where the validations occur.
- The end of a trip stage can be found at the stop or station more convenient to reach the next boarding location. This would be the nearest in the case of Metro and the one that minimizes generalized time (weighted function of in vehicle travel time and estimated walking time) for bus and bus station transactions. In the last case, common lines are considered to identify possible routes. In all cases, only stops within walking distance (1km) are considered. Within the Metro network, deterministic route choice (min travel time) is assumed.
- Trips are defined as sequences of trip stages with less than 30 minutes between the end of one stage and the beginning of the next one, and without consecutive validations in Metro or in the same bus route.

Using these assumptions, alighting stop is estimated for over 80% of the boarding transactions, and OD matrices are built. The resulting OD matrices look reasonable, being much more dense than those obtained from surveys. However, Munizaga and Palma (2012) recognize that more sophisticated methods to identify trips and trip stages are required.

The correct identification of trips and trip stages is crucial to obtain reliable OD matrices. Origins and destinations are places where certain needs are to be satisfied through the engagement on activities, while transfers are just a consequence of the interaction between the transit network and the user’s needs. Following the analysis made by Devillaine et al. (2013), we explore this using the endogenous validation and propose new rules. These new rules are tested using the exogenous validation sample.

Using the results from Munizaga and Palma (2012), Devillaine et al. (2012) propose a method to estimate location, duration and purpose of activities. This is also applied to Gatineau, using the results of Trépanier et al (2007).

### 3. Endogenous validation

Following Devillaine et al. (2013) we pursue the idea of conducting endogenous validation, i.e. analyzing the data to verify assumptions, and also to detect anomalous behavior, with the ultimate objective of proposing methodological improvements.

#### 3.1. Alighting stop estimation

##### 3.1.1 Walking distance

Munizaga and Palma (2012) propose to use 1,000m as a limit for walking distance, i.e. the search for a position-time alighting estimate is conducted within that threshold. If the position of the next boarding is further away from the bus or metro route, then it is assumed that there is a missing trip stage, which can be due to the use of another transport mode (taxi for example) or to fare evasion. This case is coded as “Too far” by Munizaga and Palma (2012) and is the most relevant failure cause (over
7% of total boarding transactions). We explored the sensibility of that parameter, analyzing cases where the estimated alighting stop was further than 1,000m from the next boarding. Surprisingly, we found that the distribution of such cases was not homogeneous in the city. More busy neighborhoods (like the CBD for example) have much more concentration of over 1,000m connections. Analyzing the cases in more detail, we found that commercial neighborhoods, where the transit network is dense, sometimes allow arriving through one transit corridor in the morning, and leaving from a different one in the evening, and that usually implies distances over 1,000m between morning alighting and evening boarding. We suggest to use different parameters for commercial and residential areas, and to calibrate the d-parameter for each case.

3.1.2. First transaction of the day
The methodology proposed by Munizaga and Palma (2012) uses the first transaction of the day to estimate alighting stop for the last boarding, assuming that there is a cycle, that begins and ends at the same point. Presumably, the cardholder’s home address. If this procedure fails, then the method looks at the first transaction of the next day. By default, days begin at 0:00 and end at 23:59. However, looking at the time distribution of transactions shown in Figure 1, we found that at midnight there is some activity that is due to the end of the previous day cycle, rather than due to the beginning of the new day cycle. Therefore, we suggest redefining the point where one day ends and the next one begins as that where the number of boarding transactions is the smallest. In our data, that would be 4:00AM.
3.1.3. Single transaction
The second most relevant cause of failure in the alighting estimation method is the presence of cards that are observed only once in a particular day, accounting for 5% of the total transactions. Looking at the time distribution of single transactions (shown in Figure 2), we found that they are more common in the evening and the afternoon, and therefore suggest estimating alighting using information of the first transaction of the next day, if available. In our application this was possible for 7% of the single transactions.
3.2. Trip stage identification

Munizaga and Palma (2012) propose to use two simple rules to identify a destination where an activity has been conducted. The first one is the thirty-minute rule. This criterion means that if the alighting time of a certain transaction and the time of the next consecutive transaction of that same user have a difference of 30 minutes or more, that frame is labeled as an activity; consequently, the stage before such frame is separated from the stage after it, constituting two different trips. The second criterion suggested by Munizaga and Palma (2012) is activated if two consecutive transactions of the same user are made on the same bus route (even if it’s on the same route and opposite directions) or at metro stations, regardless of the time frame between the alighting of the first and the boarding of the next transaction. The basis of this criterion is that the only reason a user would alight a certain bus and then board a vehicle of the same route is to perform a certain activity that reports utility to the user to compensate for the time (and possibly fare) loss.

To verify the validity of these assumptions, we analyzed the relation between Origin-Destination on-route distance and Euclidean distance. We found many reasonable cases, where the on-route distance was slightly larger than the Euclidean distance, and we also found some less reasonable cases where the on-route distance was much larger than the Euclidean distance. Figure 3 illustrates why a very large ratio can be suspicious. In this case, most likely there are two trips rather than one.
We propose to use a trip-cutting (or activity estimating) criterion utilizing as inputs the on-route traveled distance as well as the estimated OD Euclidean distance. This takes into account an observable phenomenon, which under the current methodology is being misestimated: the fact that several mass transit users modify their daily main-purpose trips to chain short-term activities (such as paying bills, minor shopping, running errands, etc), often saving time and trips, even though this usually means taking a detour from your usual route or modal choices. This detour is often meaningful enough so a distance filter can be used to capture the involved cases.

Defining:

\[ f_d = \frac{d_{on\text{-}route}}{d_{euclidean}} \]  

A first observation is that cases where \( f_d < 1 \) present a distance estimation error, since estimated on-route distance results to be smaller than Euclidean distance. These cases though existent are also rare, and mainly caused by errors in bus stop coordinates or occasional bus GPS malfunctions. In the current study, the tolerance is set on \( f_d < 0.98 \), to account for distance and coordinate approximations, bus GPS precision and their overall propagation. As for the threshold to identify trips that are involving chained short activities and need to be separated, after an statistical analysis of the data, we propose to use \( f_d > 2 \) as threshold. Figure 4 shows that the vast majority of trips are below that level. To validate this assumption, further information is required.
Every trip that yields $f_d$ above the threshold has to be split, and the different trips within it need to be identified and properly separated. This can be easily done when the original trip has two stages, since each of them is re-coded as a trip and the transfer between them is re-coded as an activity. The case of trips with three or more stages is not so trivial, since there are several ways to cut the original trip. For these cases, we propose to use the method described below.

Consider a trip with two or more stages, yielding distances such that $f_d > 2$. Let $C$ be the set of ways to divide the trip (solutions). The distance likelihood of solution $i \in C$ is defined as

$$DL_{sol_i} = -\sum_{\text{trip } j \in sol_i} |f_{dj} - f_{dj}^{real}|$$  \hspace{1cm} (2)

Where $f_{dj}$ is the ratio $f_d$ of trip $j$, and $f_{dj}^{real}$ is defined as:

$$f_{dj}^{real} = \frac{1}{#N_j} \sum_{\text{trip } k \in N_j} f_{dk}$$  \hspace{1cm} (3)

Where $N_j$ is the set of trips with equal amount of stages as $j$, and that comply with

$$0.98 \leq f_d \leq 2$$  \hspace{1cm} (4)

By definition, the distance likelihood of a solution is negative. Consequently, given a certain trip complying with $f_d > 2$, the solution whose distance likelihood is the maximum, will be the solution that minimizes the distances ratio differences compared to the trips that are considered to be correctly estimated (trips complying with equation 4).

**Bus non-boarding criterion**

The second methodological improvement is related to using information of bus scheduling to refine activities detection. This criterion is based on acknowledging the
fact that if a user does not board the bus on his/her route selection having the chance to do so, then s/he is probably engaged in an activity somewhere nearby rather than waiting unnecessarily at the bus stop. Therefore, we propose to use GPS data to verify if buses of the same route taken by the user were observed at that bus stop while the user was supposedly waiting for the bus. Some flexibility has to be introduced, because there may be other reasons for non-boarding, for example, sometimes, certain buses that are at passenger capacity level skip a bus stop if none of the passengers on board requests to alight on it. We propose to set a threshold of three buses passing though the bus stop as a condition to consider that the user was conducting an activity nearby rather that waiting at the bus stop.

4. Exogenous validation with OD Metro surveys

Sample description
Each year, Metro de Santiago conducts demand studies that include OD surveys. In 2010, they decided to include, for the first time, a question about the smartcard id. They did this as an exploratory experiment; therefore the question was included for a subsample only. Taking care of confidentiality issues, the data was made available to us to validate the assumptions made by our methods. To conduct the validation process, we applied all our methods to smartcard+GPS data obtained in the same week when the survey was applied. The data available included all the details of the trip the person was undergoing at that moment. The survey was made on a Metro station, therefore all trips had at least one stage in Metro, and the information about the previous stages was revealed (the stage had already occurred) while the information about stages after the survey was stated (the stage was about to occur).

The information gathered included time and location when/where the survey was taken, type of user (regular, student, elderly), age, trip origin (zone, nearest streets intersection), bip! card id, mode used to access metro (walk, bicycle, car, bus route), metro trip stage destination (Line, station), station used to transfer between lines, final trip destination (zone, nearest streets intersection), egress mode walk, bicycle, car, bus route), trip purpose, income. After applying some basic consistency filters to the validation database, we obtained 1,350 trip stages made in buses and Metro, corresponding to 882 trips registered in 882 surveys. When crossing with the transaction database, we found that some ids did not exist in the database, and other, even though the card exists, it did not register any transaction on that week. Therefore, we ended up with 684 cards that could be used for the validation process. However, there were some cases where even though the card was found, and it had transactions in the week when the survey was made, it was not possible to find the transactions corresponding to the survey, i.e. there was no transaction at the declared Metro station in the declared time period. The final number of survey usable for the validation process is 601.

As the survey allowed declaring only one access trip stage and only one egress trip stage per trip, and we have further information from the transactions database, we identified up to three access and egress stages. Even though only one of each will appear in the survey (per trip), the additional information can be useful for the
validation process. In Table 1 we report the number of stages found on the smartcard database, identifying also those where the Munizaga and Palma (2012) methodology was able to estimate alighting.

**Table 1: Description of the validation sample**

<table>
<thead>
<tr>
<th></th>
<th>Pre-metro stage 3*</th>
<th>Pre-metro stage 2*</th>
<th>Declared pre-metro stage</th>
<th>Metro stage</th>
<th>Declared post-metro stage</th>
<th>Post-metro stage 2*</th>
<th>Post-metro stage 3*</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>With alighting estimation</td>
<td>1</td>
<td>14</td>
<td>88</td>
<td>520</td>
<td>12</td>
<td>40</td>
<td>3</td>
<td>787</td>
</tr>
<tr>
<td>Without alighting estimation</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>76</td>
<td>37</td>
<td>12</td>
<td>1</td>
<td>139</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>18</td>
<td>96</td>
<td>596</td>
<td>158</td>
<td>52</td>
<td>4</td>
<td>926</td>
</tr>
<tr>
<td>Percentage with estimation</td>
<td>50.0</td>
<td>77.8</td>
<td>91.7</td>
<td>87.3</td>
<td>76.6</td>
<td>76.9</td>
<td>75.0</td>
<td>85.0</td>
</tr>
</tbody>
</table>

*: Not declared

The percentage of success in estimating alighting is slightly higher than the figures reported in Munizaga and Palma (2012) for the whole database. This is probably due to the fact that this sample is not representative; it is clearly biased towards Metro, where the estimation percentage is higher. When comparing by type of transaction (metro, bus or bus station), the figures are similar.

**Validation of boarding stop**

The validation of boarding location was made through comparison of the location declared in the survey and the estimation made by the Munizaga and Palma (2012) methodology. For Metro transactions, the respondents declared boarding Metro station, and the same information is recorded in the bip! database. Not surprisingly, the coincidence in this case is 100%. In the case of bus and bus-station transactions, we found some cases where the coincidence was exact, i.e. the bus stop or bus station where the passenger boarded was the nearest to the intersection of streets declared in the sample. We also found some cases where bus stop identified from the methodology was near the streets intersection declared in the sample, but was not the nearest. This usually happened when the declared intersection was more “known”. We believe that the surveyor could have induced this type of response. Both these cases were considered as correct estimations, even though some of them are more precise. A few errors where found in the more complex bus station (Estación de Intercambio Modal La Cisterna) where we have detected some problems with the GPS data; this is mainly due to the underground bus operations at this location. The overall percentage of correct estimation of boarding location is 98.9.

**Validation of alighting stop**

The validation of alighting stop is crucial for the proposed method, because all the rest of the methodology relies on the alighting stop estimation. Only transactions with alighting stop estimation and with valid survey responses can be validated, this is a total of 715 boarding transactions. Being consistent with the definition of correct estimation for the case of boarding location, we found a total of 602 cases of correct estimation (84.2%). The main source of error (12.5%) is the use of non-integrated
modes such as taxi or car share, and fare evasion, where the sequence of transactions does not allow reconstituting the trip sequence. There are a few cases (1.8%) where the alighting stop is different from the declared one because in the next stage, which was declared, the user did not validate the biP! card, or changed the destination. This could be due to a survey error, given that the last trip stages are stated rather than revealed. There is one observation (0.2%) where the method fails apparently because of the parameters used to weight walking versus in vehicle travel time.

*Validation of route choice*

The Metro OD survey also contained information about the route choice within Metro. The Munizaga and Palma model assigns route to Metro transactions assuming a deterministic minimum cost choice. The information available allows us to explore if it is reasonable to assume deterministic choice, or if we need to move towards a stochastic model. For this analysis, we could use all the observations in the sample where there were options to reach the destination station from the origin station; the card ID information was not required, because the chosen route can be compared with the route that the model assigns to that origin-destination (OD) pair, which is unique. This analysis was made with over 130,000 observations, and the results were very interesting. Looking at observed route choices, we found three cases:

1. OD pairs where all users in the sample chose the same route.
2. OD pairs where one route was clearly dominant, and the other(s) have very little demand.
3. OD pairs where clearly users chose more than one route.

Surprisingly, cases 1 and 2 were the most frequent, which is positive because it shows that there are many OD pairs where the deterministic choice assumption holds. Also, in the vast majority of those cases the route assigned by the model is the same that the users chose; however, there are a few where the model assigns a different route. This is clearly an implementation error, which can be corrected by modifying the model parameters (travel times and transfer penalties). On the other side, there is a significant amount of cases 3, where the current approach is not adequate, and a stochastic route choice model has to be implemented. We are currently analyzing modeling options.

4. **Exogenous validation with volunteers**

To improve this validation analysis, we recruited a sample of 53 volunteers, most of them students, who were shown the method’s estimations for a particular week (previous to the interview date), given the information available in the database from their personal use smartcard. Then, they were asked to validate the model’s results. During the week analyzed, the volunteers made 885 transactions, corresponding to 586 trips. The validation of the trip/trip stage identification showed that the procedure identified the trip correctly in 527 cases (90.0%), and failed in 56. There were also 3 cases where this validation could not be made because the volunteer didn’t remember
the trip. The main reasons why the method failed to identify trip/trip stages correctly are described in Table 2.

**Table 2: Causes of failure in the trip/trip stages identification procedure**

<table>
<thead>
<tr>
<th>Cause</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Because of passengers and buses crowding at the bus-stop, the user boarded 4th bus after arrival. The method cut if three buses of the same route are observed while the passenger is waiting at the bus-stop.</td>
<td>12</td>
</tr>
<tr>
<td>An intermediate trip stage didn’t have alighting estimation, therefore it was automatically coded as a trip stage.</td>
<td>9</td>
</tr>
<tr>
<td>Incorrect cut due to distance relation criteria</td>
<td>8</td>
</tr>
<tr>
<td>Extremely short activity, impossible to detect.</td>
<td>8</td>
</tr>
<tr>
<td>Error propagation because previous trip was not correctly identified.</td>
<td>8</td>
</tr>
<tr>
<td>Implementation errors</td>
<td>7</td>
</tr>
<tr>
<td>Waiting time over 30 min due to a large interval between buses or overcrowded buses that could not be boarded</td>
<td>4</td>
</tr>
</tbody>
</table>

From the 527 trips correctly identified, the method used to estimate the purpose of the trip could estimate purpose in 448, because the remaining 79 did not have alighting stop estimation. From those 448, the method estimated the purpose correctly in 352(79%). The main reasons why the purpose assignment method failed are described in Table 3.

**Table 3: causes of failure in the purpose assignment method**

<table>
<thead>
<tr>
<th>Cause</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Other” activity of duration over two hours</td>
<td>26</td>
</tr>
<tr>
<td>Stop-by at the house between trips (coded as “Other” instead of “Home”)</td>
<td>16</td>
</tr>
<tr>
<td>Work activity of duration below two hours (typically work trips rather than trips to work)</td>
<td>14</td>
</tr>
<tr>
<td>Work activity of a student card holder</td>
<td>12</td>
</tr>
<tr>
<td>Study activity of duration below two hours (weekday)</td>
<td>10</td>
</tr>
<tr>
<td>“Other” activity conducted at the end of the day</td>
<td>7</td>
</tr>
<tr>
<td>Study activity using a regular card</td>
<td>5</td>
</tr>
<tr>
<td>Study activity conducted at the end of the day</td>
<td>3</td>
</tr>
<tr>
<td>Study activity of duration below 5 hours (weekend day)</td>
<td>1</td>
</tr>
<tr>
<td>Trip to Home at 2AM coded as Work (the new day begins right after midnight)</td>
<td>1</td>
</tr>
<tr>
<td>Single trip (in a particular day) coded as “Home”</td>
<td>1</td>
</tr>
</tbody>
</table>

Some of the errors reported in Tables 2 and 3 can be corrected with implementation improvements, modifying parameters and implementing automatic checking processes. An example of this would be re-defining the frontier between one day and the following from midnight to 4AM. There are some other that can be corrected by processing information available from the data, and adapting the methodology to accommodate particular situations. For example, load profiles information can be useful to implement the non-boarded bus criteria depending on bus occupancy rates. Also, location of the zone of residence could be used as a criterion to distinguish “Home” from “Other”, “Work” or “Study” activities, depending on where they occur. Both Zone of residence and Load profiles can be estimated from the available data. On the other side, there are errors that cannot be detected or corrected without exogenous information, like for example those due to a missing trip stage, or to the
use of regular card for study activities or student card for work activities. However, the percentage of correct estimations is very high in all stages of the process, proving that this information is indeed reliable.

6. Conclusions

We have explored the reliability of the Munizaga and Palma (2012) proposed method to estimate public transport Origin-Destination (OD) flows and their level of service using smartcard and GPS data. The validity of the main assumptions was analyzed using three sources of information: the same data used to develop the method, a detailed OD survey applied to Metro users, including the card id in a percentage of them, and a group of volunteers who agreed to participate in a validation exercise. All parts of this validation analysis proved to be valuable, after this process we were able to identify implementation errors, which are easy to correct; we were able to identify some methodological improvements that will contribute to improve the quality of the results, and also, this gives us an initial assessment of the reliability of the results, which is very positive. As these are not representative samples of the population of public transport users, we cannot claim that these are definitive figures, but these first numbers are quite impressive.

Acknowledgements

Funding: Fondecyt (1120288), Fondef D10I-1002, ISCI (ICM P-05-004-F, CONICYT FBO16). We specially thank the collaboration of Transantiago and Metro.

References


