Transportation Research Part C 44 (2014) 70–79

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Validating travel behavior estimated from smartcard data

Marcela Munizaga *, Flavio Devillaine, Claudio Navarrete, Diego Silva

Departamento de Ingeniería Civil, Universidad de Chile, Blanco Encalada 2002, Santiago, Chile

ARTICLE INFO

Article history:
Received 11 December 2013
Received in revised form 11 March 2014
Accepted 12 March 2014

Keywords:
Validation
Smartcard data
OD matrices
Public transport

ABSTRACT

In this paper, we present a validation of public transport origin–destination (OD) matrices obtained from smartcard and GPS data. These matrices are very valuable for management and planning but have not been validated until now. In this work, we verify the assumptions and results of the method using three sources of information: the same database used to make the estimations, a Metro OD survey in which the card numbers are registered for a group of users, and a sample of volunteers. The results are very positive, as the percentages of correct estimation are approximately 90% in all cases.

1. Introduction

The acquisition of travel information from smartcard data is a growing trend. Many researchers have foreseen the opportunity to obtain high-quality information at a very low cost and have therefore developed tools to obtain valuable information generated as a side-product from original data on the operation of transit systems. Bagchi and White (2005) and Utsunomiya et al. (2006) presented visionary discussions of the potential of smartcard data. Pelletier et al. (2011) performed a complete review of the different applications explored to date. Among the different types of applications reported and envisioned, the use of rules-based processing to obtain OD matrices is a recurrent practice. Some examples of this application include those described by Barry et al. (2002), who used MetroCard data from New York to estimate station-to-station OD flows; Lianfu et al. (2007), who proposed a numerical method to build an OD matrix at the bus-stop level, using data from Changchun, China; Zhao et al. (2007), who inferred rail passenger trip OD matrices from an origin-only automatic fare collection system, in which the position of buses is known due to an automatic vehicle location system; Trépanier et al. (2007), who focused on the destination estimation method and applied it to data from Gatinueau, Canada; Wang et al. (2011), who developed an application for Oyster card data from London; and Munizaga and Palma (2012), who proposed modifications to previous methods to adapt them to the large and complex network of Transantiago (Chile).

In the majority of these cases, the estimation of alighting stops is crucial to the reliability of the results obtained, but many other assumptions have to be made to obtain valuable results. However, most authors recognize that the assumptions made in the process of estimating destination, route and activities from raw smartcard data need to be validated. A few attempts at validation have been made. For example, Barry et al. (2009) proposed validating the data obtained from smartcards by conducting a comparison with other sources of information, such as exit and entrance counts at subway stations and bus ride check data (counts of boarding and alighting flows as well as overall loads). However, additional information is not always available with the coverage and quality required to conduct a validation process. Farzin (2008) compared the aggregate results of an origin–destination (OD) matrix obtained from approximately 5% of the total trips (due to a lack of GPS

* Corresponding author. Tel.: +56 229784649
E-mail address: mamuniza@ing.uchile.cl (M. Munizaga).

http://dx.doi.org/10.1016/j.trc.2014.03.008
0968-090X/

equipment on many buses) with the results from an OD survey conducted almost 10 years before the smartcard data were obtained. The results were not conclusive due to the lack of representativeness for both matrices. Wang et al. (2011) also compared data with boarding/alighting counts and obtained positive results at an aggregate level. Devillaine et al. (2013) analyzed the potential bias of an OD matrix obtained from smartcard data and, with that focus, explored the possibility of endogenous validation, analyzing the available data and the results of the methods applied in processing the data. The authors also attempted to validate their results with exogenous data, but the survey that they utilized corresponded to a different week than the available smartcard data.

The objective of this paper is to propose and apply a series of methods to validate the assumptions made in the set of methodologies used to input boarding positions, alighting stops, routes chosen, activities at the destination and purpose assignment for the case of Transantiago (a public transport system in Santiago, Chile). An OD matrix was constructed using approximately 80% of the boarding transactions, including over 20 million observed trips in a week (Munizaga and Palma, 2012; Devillaine et al., 2012). Three sources of information are available for validation:

- 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, of which a small percentage provided their smartcard ID for validation purposes; and
- personal interviews from a small sample of volunteers who provided travel and personal information.

Analyses of the information used to make the estimations and of the estimations themselves (endogenous validation) can be useful to detect errors and thus improve the methodology. The large sample of OD surveys contains information that can be 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 users who provided their card ID number can be used to explore the stronger and most crucial assumption of the model: the estimation of alighting stop. The sample of volunteers also allows for the exploration of more ambitious estimation such as the identification of trip destinations/transfers and trip purpose. In the remainder of this section, we synthesize the most relevant aspects of the methodology to be validated (details can be found in the cited references). Section 2 describes the endogenous validation, Section 3 describes the exogenous validation with the Metro OD survey data, Section 4 describes the validation based on volunteers and Section 5 concludes.

1.1. Description of the methodology to be validated

Munizaga and Palma (2012) proposed a method for observing card transactions in public transport systems and for estimating travel sequences using information from transaction sequences. Only boarding transactions were observed because the payment system does not require validation when alighting. The assumptions that Munizaga and Palma (2012) made to estimate a trip matrix were as follows:

- 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 most convenient to reach the next boarding location. This station would be the nearest in the case of the Metro and the stop that minimizes the generalized time (weighted function of 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 (1 km) are considered. Within the Metro network, a deterministic route choice (minimum travel time) is assumed.
- Trips are defined as sequences of trip stages with less than 30 min between the end of one stage and the beginning of the next, without consecutive validations in the Metro or on the same bus route.

Using these assumptions, the alighting stop was estimated for over 80% of the boarding transactions, and OD matrices were built. The resulting OD matrices appear reasonable, being much denser than the OD matrices obtained from surveys. However, Munizaga and Palma (2012) recognized that more sophisticated methods for identifying trips and trip stages were required. The correct identification of trips and trip stages is crucial to obtain reliable OD matrices. Origins and destinations are locations at which the needs of the users are to be satisfied through engagement in activities, while transfers are simply a consequence of the interaction between the transit network and those needs. Using the results of Munizaga and Palma (2012), Devillaine et al. (2012) proposed a method for estimating the location, duration and purpose of activities. Given a daily sequence of transactions, the time intervals between estimated trips are regarded as work if the time elapsed between the estimated alighting from one trip and the time of boarding for the next is more than 5 h in the case of regular card users. For student cards, if the time elapsed between the estimated alighting and the next boarding is more than 2 h, study activity is assumed for that lapse. If the time elapsed between the estimated alighting and the next boarding was less than 2 h, we assumed an activity categorized as "Other" both for regular users and students. Home was assumed to be the destination after the estimated alighting for the last transaction of the day and the first boarding transaction of the next day. Following the analysis of Devillaine et al. (2013) regarding the validity of these assumptions, we explore the use of endogenous validation and propose new rules in Section 2. These new rules are tested using the exogenous validation sample in Section 4.

2. Endogenous validation

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

2.1. Alighting stop estimation

2.1.1. Walking distance

Munizaga and Palma (2012) proposed the use of 1 km as a limit for walking distance, i.e., the search for a position-time alighting estimate was conducted within that threshold. If the position of the next boarding was farther away from the bus or Metro route, then a missing trip stage was assumed, possibly due to the use of another transport mode (taxi, for example) or to fare evasion. This case was coded as "Too far" by Munizaga and Palma (2012) and was the most relevant cause of failure (over 7% of total boarding transactions). We explored the sensibility of this parameter, analyzing cases in which the estimated alighting stop was farther than 1000 m from the next boarding. Fig. 1 shows the spatial distribution of alighting estimation failures due to these criteria (the distance between the estimated alighting and next boarding is too far to be considered a reasonable connection). We observed that the distribution of such cases is not homogeneous in the city. Busier neighborhoods (such as the CBD, for example) have a much higher concentration of connections over 1 km. Analyzing these cases in more detail using Google Earth to visualize locations with a high concentration of connections considered distant, we found that commercial neighborhoods in which the transit network is dense sometimes lead one to arrive through one transit corridor in the morning and leave from a different corridor in the evening, usually implying distances over 1 km between morning alighting and evening boarding. This result suggests the use of different parameters for commercial and residential areas and calibration of the d-parameter for each case; however, a large and reliable exogenous database would be required to analyze this phenomenon in greater detail.

2.1.2. First transaction of the day

The methodology proposed by Munizaga and Palma (2012) used the first transaction of the day to estimate the 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). 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 Fig. 2, we found that at midnight there is some activity corresponding to the end of the cycle of the previous day, rather than the beginning of the new day's cycle. This finding is quite relevant because if the first/last trips of the day are not correctly identified, then some of the assumptions fail, and, for example, the two-stage trip required to return home for a person who boards the first stage of the trip before

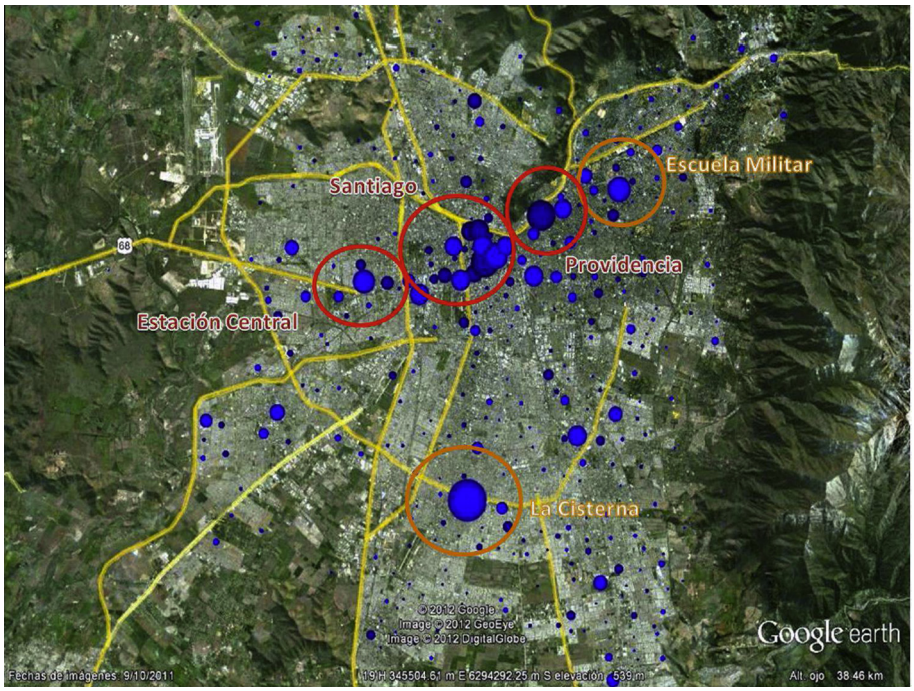

Fig. 1. Spatial distribution of connections considered "too far".

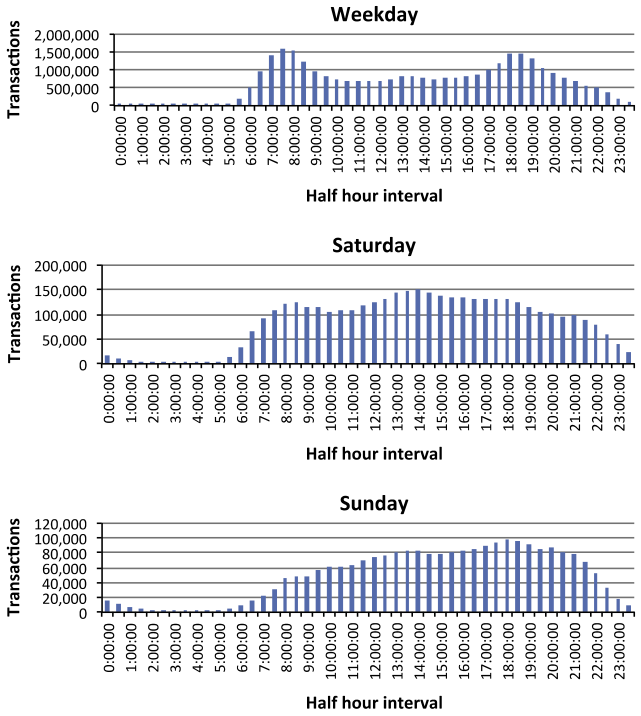

Fig. 2. Time distribution of transactions.

midnight and the second stage of it after midnight will be regarded as two short trips, one on each day. Therefore, we will not be able to find a reasonable match between the first and the last trip of a day. Therefore, we suggest changing the arbitrary transition period between one day and the next from midnight to the time when the lowest activity is observed, i.e., the number of boarding transactions is the smallest. In our data, that time is 4:00 AM. This modification will help reduce errors.

2.1.3. Single transaction

The second most relevant cause of failure in the estimation method for alighting 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 Fig. 3), we found that an important proportion of these transactions occur in the evening and the afternoon and therefore suggest the possibility of estimating alighting using information from the first transaction of the next day, if available. In our application, this estimation was possible for 7% of the single transactions.

2.2. Trip stage identification

Munizaga and Palma (2012) proposed the use of two simple rules to identify a destination at which an activity has been conducted. The first rule is the 30-min rule. This criterion means that if the alighting time of a certain transaction and the time of the next consecutive transaction of the same user have a difference of 30 min or more, that frame is labeled as an activity. Consequently, the stage before such a 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

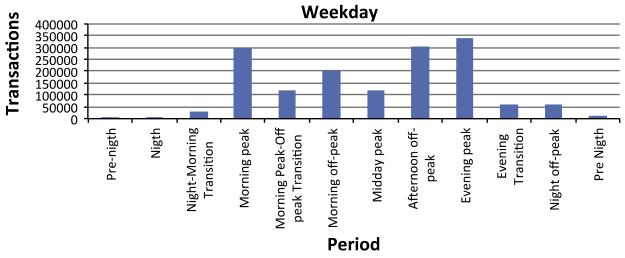

Fig. 3. Time distribution of single transactions.

are made on the same bus route (even if the two consecutive transactions are on the same route in opposite directions) or at Metro stations, regardless of the time frame between the first alighting and the boarding of the next transaction. The basis of this criterion is that the only reason a user would alight from a certain bus and then board a vehicle on the same route is to perform a certain activity that provides utility to the user to compensate for the time (and possibly fare) loss.

To verify the validity of these assumptions, we analyzed the relationship between OD on-route distance and Euclidean distance. We found many reasonable cases in which the on-route distance was slightly larger than the Euclidean distance, and we also found some less reasonable cases in which the on-route distance was much larger than the Euclidean distance. Fig. 4 illustrates why a very large ratio can be suspicious. In this case, it is unlikely that a user would make a three-leg bus trip (Route 1, Route 2 and Route 3) to go from A to B. It is most likely that this case represents two trips rather than one, and some activity is conducted at certain intermediate stops.

We propose the use of a trip-cutting (or activity-estimating) criterion, utilizing as inputs the on-route traveled distance as well as the estimated OD Euclidean distance. This approach takes into account an observable phenomenon that is inaccurately estimated using the current methodology: the fact that several mass transit users modify their daily main-purpose trips to chain short-term activities (such as paying bills, minor shopping, and running errands), often saving time and trips even though this chain usually means taking a detour from the usual route or modal choices. This detour is often sufficiently meaningful that a distance filter can be used to capture the cases involved. We define f_d as follows:

$$f_d = \frac{d_{on-route}}{d_{euclidean}} \tag{1}$$

We observed cases in which $f_d < 1$, indicating a distance estimation error because the estimated on-route distance cannot be smaller than the Euclidean distance. These cases, though existent, are also rare and are primarily caused by errors in bus stop coordinates or occasional bus GPS malfunctions. In the current study, the tolerance is set at $f_d < 0.98$ to account for distance and coordinate approximations, bus GPS precision and their overall propagation. As for the threshold for identifying trips that involve chained short activities that must be separated, after a statistical analysis of the data, we propose to use $f_d < 2$ as the threshold. Fig. 5 shows that the vast majority of trips are below the level of 2. To validate this assumption, further information is required.

Every trip that yields an f_d value above the threshold must be split, and the different trips in this case must be identified and properly separated. This separation can be easily performed when the original trip has two stages because each stage 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 as trivial because there are several ways to divide the original trip. For these cases, we propose the method described below.

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

$$DL_{sol_i} = -\sum_{trip\ j\ \in\ sol_i} |f_{d_j} - f_{d_j}^{real}| \tag{2}$$

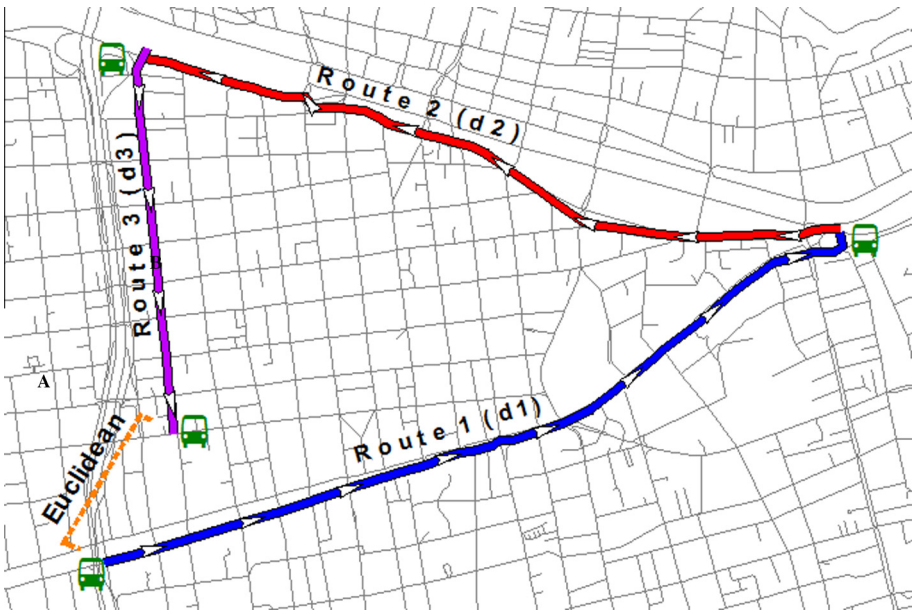

Fig. 4. An example of a three-stage trip Euclidean distance and on-route distance.

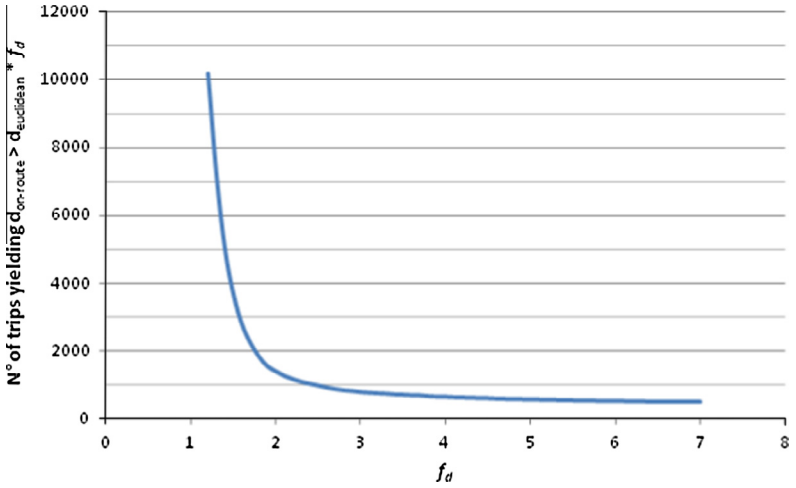

Fig. 5. Number of trips below f_d vs. f_d.

where f_{d_j} is the ratio f_d of trip j and is defined as

$$f_{d_j}^{real} = \frac{1}{\#N_j} \cdot \sum_{trip\ k \in N_j} f_{dk} \tag{3}$$

where N_j is the set of trips with a number of stages equal to j that complies with

$$0.98 \leqslant f_g \leqslant 2 \tag{4}$$

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

The second methodological improvement is related to the use of the observed frequency of buses to refine activity detection. This criterion is based on the fact that if a user does not board the bus on his selected route when he has the chance to do so, then he is most likely engaged in an activity somewhere nearby rather than waiting unnecessarily at the bus stop. Therefore, we propose to use GPS data to verify whether 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 must be introduced because there may be other reasons for not 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 at that stop. We propose a threshold of three buses passing through the bus stop as a condition to consider that the user was engaged in an activity nearby rather than waiting at the bus stop.

3. Exogenous validation with OD Metro surveys

3.1. Sample description

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

The information gathered included the time and location when/where the survey was taken, user type (regular, student, elderly), age, trip origin (zone, nearest street intersection), bip! card ID, mode used to access the Metro (walking, bicycle, car, bus route), Metro trip stage destination (line, station), station used to transfer between lines, final trip destination (zone, nearest street intersection), egress mode (walking, bicycle, car, bus route), trip purpose and income. After applying some basic consistency filters to the validation database, we obtained 1350 trip stages conducted in buses and on the 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 others for which the card existed, but the card did not register any transactions in that week. As a result, 684 cards could be used for the validation process. However, in some cases in which the card was found and had

transactions for the week of the survey, 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. Overall, the final number of surveys usable for the validation process was 601.

Although the survey allowed for the declaration of only one access stage and only one egress stage per trip, we could obtain further information from the transactions database; thus, we identified up to three access and egress stages, as illustrated in Fig. 6. Only one of the pre-metro and one of the post-metro bus stages will appear in the survey database; however, this information may be useful for validation purposes.

In Table 1, we report the number of stages found in the smartcard database that can be associated with the surveyed trips. We also report the percentage of those stages in which the Munizaga and Palma (2012) methodology could estimate alighting. The percentage of success in the alighting estimation is slightly higher than the figures reported by Munizaga and Palma (2012) for the entire database, most likely because this sample is not representative; the sample is clearly biased towards the Metro, for which the estimation percentage is higher. When comparing by type of transaction (Metro, bus or bus station), the figures are similar.

3.2. Validation of boarding stop

The boarding location was validated through a comparison of the location declared in the survey and the estimation made by the Munizaga and Palma (2012) methodology after matching the transaction information (time, validator ID) with the location information available from buses' GPS and recorded data for fixed position validators (Metro and bus stations). This process involves some assumptions and data processing that need to be validated. For Metro transactions, the respondents declared boarding at a Metro station, and the same information is recorded in the bip! database. Therefore, we do not expect any difference in this case, and the coincidence is indeed 100%. In the case of bus and bus-station transactions, we found some cases in which the coincidence was exact, i.e., the bus stop or bus station at which the passenger boarded was nearest to the intersection of streets declared in the sample. We also found some cases in which the bus stop identified from the methodology was near the street intersection declared in the sample but was not the nearest. This case usually arose when the declared intersection was more visible or well known compared to the exact one. We believe that the surveyor could have induced this type of response. Both of these cases were considered as correct estimations, even though some of them are more precise. A few errors were found for a more complex bus station (Estación de Intercambio Modal La Cisterna), where we detected some problems with the GPS data primarily due to the underground bus operations at this location. The overall percentage of correct estimations of boarding location is 98.9%.

3.3. Validation of alighting stop

Validation of the alighting stop is crucial for the proposed method because the rest of the methodology relies on the estimation of the alighting stop. Only transactions with alighting stop estimations and valid survey responses can be validated; in this case, there was 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 the trip sequence to be reconstituted. There were a few cases (1.8%) in which the alighting stop was different from the declared stop because in the next stage, which was declared, the user did not validate the bip! card or changed their destination. These cases could be due to a survey error, given that the last trip stages are stated rather than revealed. There was one observation (0.2%) in which the method failed due to the parameters used to weight walking versus in-vehicle travel time.

Fig. 6. Illustration of surveyed and non-surveyed trip stages.

Table 1
Description of the validation sample.

	Pre-metro stage 3[a]	Pre-metro stage 2[a]	Pre-metro stage 1 (declared)	Metro stage	Post-metro stage 1 (declared)	Post-metro stage 2[a]	Post-metro stage 3[a]	Total
Total	2	18	96	596	158	52	4	926
With alighting estimation	1	14	88	520	121	40	3	787
	50.0%	77.8%	91.7%	87.3%	76.6%	76.9%	75.0%	85.0%

[a] Not declared.

3.4. Validation of route choice

The Metro OD survey also contained information about the route choice within the Metro network. The Munizaga and Palma model assigns routes to Metro transactions assuming a deterministic minimum cost choice. The available information allows us to explore whether it is reasonable to assume deterministic choice or whether we should move towards a stochastic model. For this analysis, all the observations in the Metro survey can be used, as card ID information is not required to compare the chosen route with the route that the model assigns to that OD pair, which is unique. The Metro network has 5 lines and 108 stations, 16 of which are transfer stations. There are 11,540 OD pairs in the network, but over 20% of them correspond to cases in which the origin and destination are on the same line; hence, the deterministic assignment is correct. Focusing on the OD pairs that require a transfer, i.e., the origin and destination stations are on different lines, we have a sample of 135,316 trips observed in 5039 OD pairs. In 2876 (57%) of those trips, only one route was observed in the survey sample; in 1,716 cases (34%), two routes were observed; and in the remaining 447 (9%), three or four routes were observed. In the cases in which two routes were observed in the survey, we determined whether there was clear dominance of one route over the other using a measure of statistical dispersion (Gini coefficient), and we observed that 68.5% of the OD pairs in which two routes were chosen showed clear dominance of one over the other. Overall, we observed three types of cases:

(1) OD pairs in which all of the users in the sample chose the same route (67%).
(2) OD pairs in which one route was clearly dominant whereas the other(s) had very little demand (18%).
(3) OD pairs in which users clearly chose more than one route (15%).

Cases 1 and 2 occurred most frequently; this is a positive finding because it demonstrates that there are many OD pairs for which the deterministic choice assumption holds. Additionally, in the vast majority of those cases, the route assigned by the model is the same route that the users chose. However, there are a few cases in which the model assigns a different route. This assignment clearly represents an implementation error, which can be corrected by modifying the model parameters (travel times and transfer penalties). Nevertheless, there are a significant number of cases corresponding to case 3, where the current approach is not adequate, and in these situations, a stochastic route choice model must be implemented. This is a relevant subject for further research, as two new Metro lines will be added to the network in the near future, increasing the transfer options. We are currently developing a stochastic model to be applied to those OD pairs for which the deterministic model is not suitable.

4. Exogenous validation with volunteers

To complete this validation analysis, we recruited a sample of 53 volunteers, primarily composed of students, who were shown the estimations of the method for a particular week (prior to the interview) and were given the information available in the database from their personal smartcard use. Then, they were asked to validate the results of the model. During the week analyzed, the volunteers made a total of 885 transactions, corresponding to 586 trips. The advantage of this sample is that it allows validation of not only the alighting stop estimation and trip/trip stage cut but also the validation of more complex derivations such as purpose assignment.

The validation of the trip/trip stage identification showed that the procedure correctly identified the trips in 527 cases (90.0%) and failed in 56. There were also 3 cases that could not be validated because the volunteer did not remember the trip. The main reasons explaining why the method failed to correctly identify the trip/trip stages are described in Table 2.

Among the 527 correctly identified trips, the purpose of the trip was estimated in 448 because the remaining 79 did not have an alighting stop estimation. Among those 448 cases, the method correctly estimated the purpose in 352 (79%). The main reasons explaining why the purpose assignment method failed are described in Table 3.

Some of the errors reported in Tables 2 and 3 can be corrected by implementing improvements, modifying parameters and incorporating automatic checking processes. The parameters can be modified using the information available in the main database but also by using the information available from exogenous validation samples. For example, we re-defined the boundary between one day and the following from midnight to 4:00 AM based on the observation of the large database

Table 2
Causes of failure in the trip/trip stage identification procedure.

Due to passenger and bus crowding at the bus stop, the user boarded the 4th bus after arrival. The method used a threshold of 3 buses of the same route while the passenger is waiting at the bus stop	12
An intermediate trip stage did not have an alighting estimation; therefore, it was automatically coded as a trip stage	9
Incorrect cut due to distance relation criteria	8
Extremely short activity, impossible to detect	8
Error propagation because a previous trip was not correctly identified	8
Implementation errors	7
Waiting time over 30 min due to a large interval between buses or overcrowded buses that could not be boarded	4

Table 3
Causes of failure in the purpose assignment method.

"Other" activity with a duration of over 2 h	26
Stop-by at the house between trips (coded as "Other" instead of "Home")	16
Work activity with a duration of less than 2 h (typically work trips rather than trips to work)	14
Work activity of a student card holder	12
Study activity with a duration of less than 2 h (weekday)	10
"Other" activity conducted at the end of the day	7
Study activity using a regular card	5
Study activity conducted at the end of the day	3
Study activity with a duration of less than 5 h (weekend day)	1
Trip to Home at 2 AM coded as Work (the new day begins right after midnight)	1
Single trip (in a particular day) coded as "Home"	1

but also based on the surveys. If we move that limit too much to the right, trips that are actually the first activities of one day will be regarded as the last of the previous day. This parameter calibration process can be improved with a larger validation sample. Other errors can be corrected by processing the available information and adapting the methodology to accommodate particular situations. For example, load profile information can be useful to implement the non-boarded bus criteria depending on bus occupancy rates. In addition, locations of zones of residence could be used as a criterion to distinguish "Home" from "Other", "Work" or "Study" activities, depending on where these activities occur. Both zone of residence and load profiles can be estimated from the available data. However, there are errors that cannot be detected or corrected without exogenous information, such as, for example, errors due to a missing trip stage or to the use of a regular card for study activities or a student card for work activities. However, the percentage of correct estimations is very high in all stages of the process, showing that this information is indeed reliable.

5. Conclusions

We have explored the reliability of the method proposed by Munizaga and Palma (2012) to estimate public transport 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 certain percentage of cases; and a group of volunteers who agreed to participate in a validation exercise.

All stages of this validation analysis proved to be valuable. After this process, we were able to identify implementation errors that are easy to correct. We were able to identify some methodological improvements that will enhance the quality of the results, and this study also gives an initial assessment of the reliability of the results, which is very positive. The most relevant methodological improvements include the introduction of flexibility in the maximum walking distance parameter (d), a re-definition of the time limit between one day and the next as 4:00 AM, the possibility of using single transactions matching with information from the following day, and a new method for identifying trip stages to distinguish transfers from activities using a criterion on route distance versus Euclidean distance and include the observed frequency of buses at bus stops.

In terms of validation, we found that the boarding stop is correctly estimated in 98.9% of the cases, the alighting bus stop is correctly estimated in 84.2% of the cases, the trip/trip stage identification is correct in 90% of the cases, and the purpose is correctly estimated in 79% of the cases. These values are extraordinarily positive and demonstrate that the generated information is highly reliable. In the near future, we will obtain a large validation sample that is representative of the population from the Santiago 2012–2013 OD survey, which will allow us to validate and estimate these percentages according to various segments of the population (income, type of user, zone of residence). This sample will also allow us to calibrate some of the parameters to minimize errors. In summary, these results, which are already quite good, can be further improved by implementing the suggested methodological modifications.

Finally, it is worth mentioning that these results have been transferred to the industry, and are currently being used for planning purposes, both by the operators and by the transport authority. This can also be considered a proof of value or qualitative validation.

Acknowledgments

Funding: Fondo Nacional de Desarrollo Científico y Tecnológico (1120288), Fondef D10I-1002, ISCI (ICM P-05-004-F, CONICYT FBO16). We especially appreciate the collaboration of Directorio de Transporte Público Metropolitano.

References

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12 (5), 464–474.
Barry, J.J., Newhouser, R., Rahbee, A., Sayeda, S., 2002. Origin and destination estimation in New York City with automated fare system data. Transp. Res. Rec. 1817, 183–187.

Barry, J., Freiner, R., Slavin, H., 2009. Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transp. Res. Rec. 2112, 53–61.
Devillaine, F., Munizaga, M.A., Trepanier, M., 2012. Detection activities of public transport users by analyzing smart card data. Transp. Res. Rec. 2276, 48–55.
Devillaine, F., Munizaga, M.A., Palma, C., Zúñiga, M., 2013. Towards a reliable Origin-Destination matrix from massive amounts of Smartcard and GPS data: application to Santiago In: Zmud, J., Lee-Gosselin, M., Munizaga, M.A., Carrasco, J.A. (Eds.). Transport Survey Methods; Best Practice for Decision Making, Emerald, pp. 695–710.
Farzin, J., 2008. Constructing an automated bus Origin-Destination matrix using farecard and global positioning system data in Sao Paulo, Brazil. Transp. Res. Rec. 2072, 30–37.
Lianfu, Z., Shuzhi, Z., Yonggang, Z., Ziyin, Z., 2007. Study on the method of constructing bus stops OD matrix based on IC card data. Wireless Commun. Netw. Mobile Comput. WiCom 2007, 3147–3150.
Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive Smart card data from Santiago, Chile. Transport. Res. 24C (12), 9–18.
Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transport. Res. Part C 19 (4), 557–568.
Trépanier, M., Chapleau, R., Tranchant, N., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. J. Intell. Transport. Syst. 11 (1), 1–14.
Utsunomiya, M., Attanucci, J., Wilson, N., 2006. Potential uses of transit smart card registration and transaction data to improve transit planning. Transp. Res. Rec. 1971, 119–126.
Wang, W., Attanucci, J.P., Wilson, N.H.M., 2011. Bus passenger origin-destination estimation and related analyses using Automated Data Collection Systems. J. Public Transport. 14 (4), 131–150.
Zhao, J., Rahbee, A., Wilson, N., 2007. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Comput. Aided Civil Infrastruct. Eng. 22, 376–387.