Transportation Research Part C 24 (2012) 9–18

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

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

Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile

Marcela A. Munizaga [a,*], Carolina Palma [b,1]

[a] Departamento de Ingeniería Civil, Universidad de Chile, Casilla 228-3, Santiago, Chile
[b] Coordinación Transantiago, Moneda 975, Santiago, Chile

ARTICLE INFO

Article history:
Received 28 September 2011
Received in revised form 19 January 2012
Accepted 20 January 2012

Keywords:
OD matrix
Smartcard data
Automatic vehicle location

ABSTRACT

A high-quality Origin–Destination (OD) matrix is a fundamental prerequisite for any serious transport system analysis. However, it is not always easy to obtain it because OD surveys are expensive and difficult to implement. This is particularly relevant in large cities with congested networks, where detailed zonification and time disaggregation require large sample sizes and complicated survey methods. Therefore, the incorporation of information technology in some public transport systems around the world is an excellent opportunity for passive data collection. In this paper, we present a methodology for estimating a public transport OD matrix from smartcard and GPS data for Santiago, Chile. The proposed method is applied to two 1-week datasets obtained for different time periods. From the data available, we obtain detailed information about the time and position of boarding public transportation and generate an estimation of time and position of alighting for over 80% of the boarding transactions. The results are available at any desired time–space disaggregation. After some post-processing and after incorporating expansion factors to account for unobserved trips, we build public transport OD matrices.

1. Introduction

In the late 1990s, smartcard payment systems were incorporated in some cities, such as Washington (Smartrip) and Tokyo (Suica). This new technology rapidly spread to other cities, and it has become an important part of the present public transport fare collection system. For example, the Oyster card was implemented in London in 2003, offering discount fares (compared with buying single tickets), and it is currently the most popular payment method. In Chicago, the farecard has a very high penetration rate (close to 90%, according to Zhao et al. (2007)). Other examples, with different methods of implementation and different levels of penetration, include San Francisco (Buneman, 1984), Portland (Furth et al., 2006), New York (Barry et al., 2002), The Netherlands (Muller and Furth, 2001; Furth et al., 2006), Changchun (China) (Lianfu et al., 2007) and Quebec (Trépanier et al., 2007; Chapleau and Chu, 2007; Chapleau et al., 2008). In all of these cities, the smartcard is used as one of the payment possibilities. In Santiago (Chile), it is the only available payment system in buses and is by far the most important in the Metro; the overall penetration rate is 97% (Beltrán et al., 2011).

The research challenge of obtaining valuable information from the data generated by smartcard transactions has been taken on by several researchers who recognize its potential to improve public transport planning and operation. For

* Corresponding author. Tel.: +56 2 9784649; fax: +56 2 6894206.
E-mail address: mamuniza@ing.uchile.cl (M.A. Munizaga).
[1] Present address: Cityplanning Consultancy, Chile.

0968-090X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.trc.2012.01.007

example, Chapleau and Chu (2007) propose a method to identify and replace incorrect or suspicious observations from the automatic fare collection system; Trépanier et al. (2007) propose a method to estimate the alighting point of a trip in a system where users only validate when boarding; Lianfu et al. (2007) propose a method to build an OD matrix at bus-stop level, using the data generated in Changchun, China; Zhao et al. (2007) develop a method for inferring rail passenger trip OD matrices from an origin-only automatic fare collection system, where the position of buses are known due to an automatic vehicle location system. For a detailed literature review, see Pelletier et al. (2011). The research efforts have focused on the integration and enrichment of the information available from different passive sources (such as automatic fare collection systems, automatic vehicle location systems and passenger counts), detection and correction of information errors, estimation of alighting or destination point, identification of transfers and generation of OD matrices from the information available.

This paper presents a methodology that increases the dimension and complexity of the public transport system. That is, the methodology is applicable to large-scale multimode public transport systems. The rest of the paper is organized as follows: in the next section, the Transantiago public transport system and the data available are described; in Section 3, we present a statistical analysis of the data; in Section 4, the methodology proposed is presented; Section 5 contains the results of the application and Section 6 contains the conclusions.

2. Background

The data available come from Transantiago, the public transport system that has been available in Santiago, Chile, since February 2007. The system is based on a trunk–feeder structure, where the Metro (underground) is an important component. It has nine feeder operation areas, each serving one part of the city. There are also six trunk operators, which operate larger routes across the city. Trunk operator 1 is the Metro; the rest are bus lines. The payment system is such that each passenger pays a fare when entering the system that allows him/her to make up to three transfers within 2 h. The payment structure is slightly different between buses and the Metro. In the buses, the only payment method is the contactless smartcard, bip!, while in the Metro, it is possible to buy a single ticket or to use the bip! card; however, the percentage of users who buy a single ticket in the Metro is very small (approximately 3%). The fare is also different, slightly higher for the Metro at peak hours; buses have a flat fare. If a passenger uses a bus first and the Metro afterwards, the difference between both fares is charged when entering the Metro. All Metro lines are connected, and internal changes between lines are made without showing the bip! card again.

Santiago, the Chilean capital city, has nearly 6 million inhabitants and has a nonhomogeneous distribution of the population, with clearly wealthier and poorer neighborhoods. The city has a circular shape, with a large proportion of trips heading from the suburbs to the center in the morning, and moving in the opposite direction in the evening. According to the 2001 OD survey (Dictuc, 2003), there were 16 million trips in a working day, and of these, 10 million were motorized trips (38.6% walking or bicycle). The average household size was 3.81 people, and the trip rates were 2.82 trips per person, resulting in 10.76 trips per household. The market share of public transport by then was 53%.

There were severe problems at the beginning of the operation of new the system, but it is now operating normally, with some persisting problems in certain areas. A payment evasion problem has been detected in the buses, geographically biased toward poor neighborhoods located far from the city center. Evasion is almost nonexistent in the Metro. The system contains over 300 bus routes and nearly 6000 buses. It has more than 10,000 bus stops and 85 km of Metro rails. More than 11 million bip! cards have been issued. There are 150 bus stations with very light infrastructure; a fence in most cases, equipped with an extra vehicle payment system where passengers pay when entering the station, which increases the boarding efficiency. These bus stations, called "Zonas paga", operate during peak hours at congested points.

All bip! transactions are recorded in a database that contains information about the operator and the instant when the transaction was made. Each passenger has to make a transaction by putting his/her card close to a payment device when entering a bus, bus station or Metro station. Each payment device has an *id* and is associated with a bus, Metro station or bus station. The information recorded for each transaction includes the card id and type, code of bus or site where the transaction was made, time, date, and amount of money paid. Every week, there are approximately 35 million bip! transactions made with over 3 million bip! cards. The location of the transaction can be obtained directly from a database for Metro and bus stations, because they have fixed location. However, it is not available from this database for bus transactions.

Another database contains geocoded information about all the buses, such as latitude, longitude, time, date and instant speed. Each bus is identified by a plate number and an operator code. The position is available every 30 s. Every week of data contains approximately 80 million GPS observations. The geocoded bus routes and positions of Metro stations, bus stops and bus stations are also known and valuable information. There are timetables associated with bus and Metro services and also with bus stations, but they mainly indicate the operative hours and the frequency of each service (no scheduled services are provided).

On the other hand, bus assignment information is stored in a database that contains the service each bus is giving by time period. The Transantiago authority using the GPS positions of buses throughout the day generates this information. For each service route, they define three points, located at the beginning, middle and end of the route, each with an area of influence given by a 500 m square centered at the point. If a bus is detected going through all three areas of influence, then it is assigned to that service from the instant of the last GPS pulse inside the influence area of the initial point of the service, until the first GPS pulse inside the area of influence of the route ends. Therefore, at the beginning and at the end of routes, there is

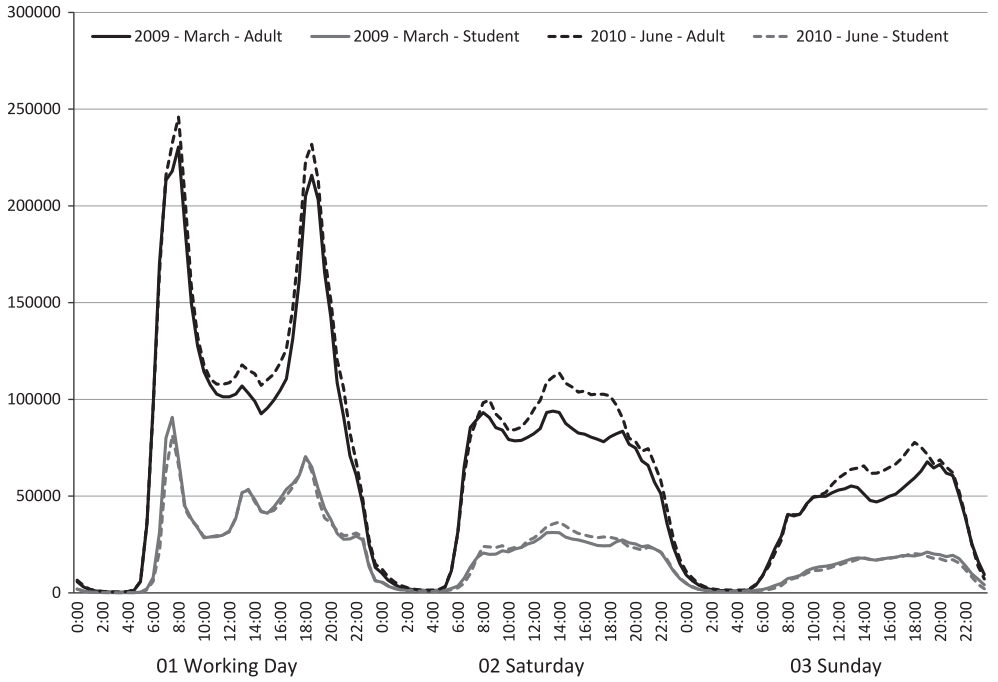

Fig. 1. Time distribution of transactions.

a small loss of information. However, the process has been evaluated by the Transantiago Authority and the operators and has been proven to be reasonably reliable.

3. Statistical description of the data

Every week, over 35 million bip! transactions are recorded. A descriptive statistical analysis based on the 2 weeks used in this study (March 2009; June 2010) is conducted. The number of transactions throughout the week shows similar numbers for working days of approximately six million transactions per day. During the weekend, this number falls to less than four million on Saturday, and less than 2.5 million on Sunday. The relation between total and first-stage transactions is 1.65, which provides an initial idea of the number of segments per trip. In addition, in an average working day, 60% of the transactions are made in buses, 33% in Metro stations and 7% in bus stations.

Fig. 1 shows the time distribution of boarding transactions for Monday through Friday (working day), Saturday and Sunday in the two different weeks. It can be seen that the patterns for the 2 weeks are similar, especially in the case of the working day, where a clear morning peak can be observed at 8:00 for adults and 7:30 for student passengers.[2] Similarly, the evening peak is observed at 18:30 for adults and 18:00 for students. A much smaller but still noticeable midday peak is observed between 13:00 and 14:00. Saturday and Sunday have many fewer transactions and show more differences between time periods. Saturday has a pattern that starts early in the morning (as early as in a working day), but rises only up to 100,000 adults and 25,000 students. The midday peak is more important, and the number of transactions keeps decreasing in the afternoon and evening, arriving at very low values after midnight. The Sunday pattern shows no morning peak, a moderate afternoon peak, and a more important evening peak at 18:00–19:00. This information is valuable for policy measure evaluation. However, more processed information, such as load profiles, is required for planning and design purposes.

By matching the Transactions and Positions databases through bus plates or Metro/Bus-station codes and times, it was possible to identify the position where the transaction was made in 98.5% of the cases for the March 2009 data and in 99.9% of the cases for the June 2010 data. The small percentage of failure to match is due to missing information in the Positions database, because the GPS equipment some times fails to operate. However, a new incentive has been implemented that is paid to the operators depending on the routes served according to the GPS data (Beltrán et al., 2012). Therefore, they have an increasing interest on the GPS equipment to operate correctly. This information can be used to make a spatial analysis of transactions, as shown in Fig. 2 for boarding transactions at bus stops. The aggregate analysis of all routes over time shows that the morning peak is different in different locations. It is remarkably earlier in poorer neighborhoods, which are

[2] Students are identified as those users who have personalized student cards provided by the Chilean government to low income students from primary and secondary schools, and to all university students.

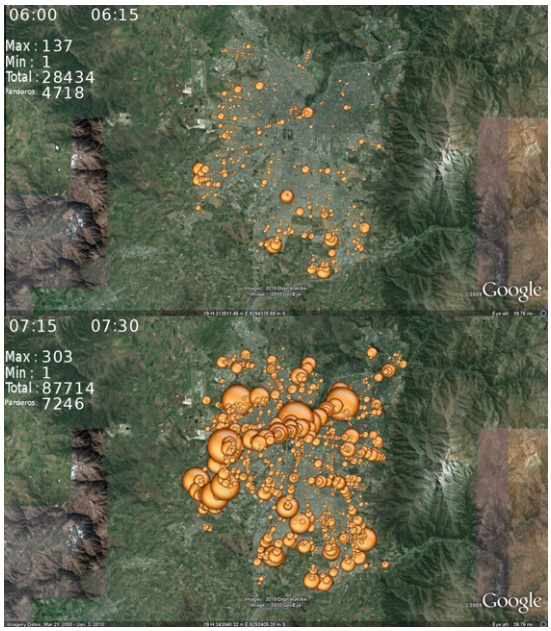

Fig. 2. Boarding transactions at different times.

located in the north, west and south suburbs. The amount of information gathered permits as much time and space disaggregation as required to perform this kind of analysis.

4. Methodology

Our goal is to use this very rich source of information to build an OD matrix of public transport trips. Based on the previous work reported in Section 1 and a preliminary analysis conducted over a small subsample of cards (users), a methodology suitable for large public transport systems such as Transantiago is proposed. Some definitions are required to explain the proposed method. Let us define a trip as a movement from a point of origin to a point of destination (Ortúzar and Willumsen, 2011). Each trip can have one or more segments, which are movements in a particular service (bus or Metro). Origin and destination are the positions where the trip begins and ends, respectively. Boarding and alighting points are the positions where the segment begins and ends, respectively. Given the fare structure of the system, only bus–bus and bus–Metro combinations are recognized as transfers. When a user transfers between Metro lines, the change is not registered by the automatic fare collection system, and therefore, only one trip segment is accounted for.

The main objective of the proposed method is to reconstruct the trip chain of users behind bip! cards by estimating the destination points from the information available. Once this is available, it is possible to analyze behavior, build OD matrices, estimate vehicle load profiles, and complete many other tasks in a simple and direct way. The inputs of the model are three main databases: transactions (boarding) from an automatic fare collection system, vehicle position from the automatic vehicle location system and a geocoded definition of the public transport network. After matching these three databases, it is possible to obtain the position of the transactions and then estimate the alighting point. The estimation procedure is described below. It is different for transactions in buses, bus stations and Metro stations, but in all cases, the result is an estimate of the position-time coordinates of the alighting point. Using this information, our proposed method includes a module to distinguish transfer from destination, identifying trip segments, which is described in Section 4.2. As a result of this procedure, trips and trip segments are obtained for a proportion of the sample. Furthermore, in some of those cases, the method will be able to estimate the alighting point of all boarding transactions of a particular card in a particular day. Those cases are very valuable, because they make it possible to build the public transport trip chain of the person behind a given card. However, there are some cases in which the estimation of the alighting point is not possible.

4.1. Alighting point estimation

Because the system has boarding validation only, it is necessary to estimate the alighting point following the transactions sequence and assuming that the next bip! transaction occurs after alighting. This problem has been partially analyzed before. For example, Barry et al. (2002) presented a method to estimate the alighting station for the New York subway system based on two assumptions: first, that after a trip, users will return to the destination of the previous trip station, and second, that at

the end of a day, users will return to the station where they boarded for the first trip of that same day. Later, Zhao et al. (2007) proposed a method to estimate the alighting point for rail boarding transactions in the Chicago CTA system, focusing on rail boardings followed by a bus boarding transaction. They assigned the nearest station to the next boarding bus stop within a 400 m radius as the alighting station (trip segment destination). To apply this method, they made the same assumptions as Barry et al. (2002) but also assumed that the maximum walking distance is 400 m, or 5 min. Trépanier et al. (2007) developed an object-oriented method to estimate the alighting bus stop in the bus system of Gatineau STO. They also took the Barry et al. (2002) assumptions and used the distance to the next boarding as the main criteria to define the alighting bus stop but incorporated the possibility of looking at the next day, even observing weekly travel patterns to complete missing information. Zhao et al. (2007) reported a 71% success rate in estimating alighting stations for rail boardings, while Trépanier et al. (2007) obtained a 66% success for the bus-only Gatineau system.

We propose a method to estimate the alighting station in a multimodal public transport system, where boarding transactions are observed in a complex network in which users travel using the Metro and buses and sometimes validate their trip in a bus station instead of doing so directly on a bus. We will assume that each card corresponds to a user, so card and user will be used indistinctively. Even though the basic principles are the same as those assumed by Barry et al. (2002), Zhao et al. (2007) and Trépanier et al. (2007), we need to incorporate additional constraints and modify the objective function. The basic idea, as shown in Fig. 3, is to follow the trip chain of a card and identify the alighting position (bus or Metro station) by looking at the position and time of the next boarding. This is only possible when both the current and proceeding transactions have position information, which is taken from the automatic vehicle location database. In the case of the last transaction of the day, as in previous works, we assume that its destination is close to the point where the first trip of the day began, finishing the daily trip cycle for that particular user. If there is only one trip per card, no inference is possible with single day information.

4.2. Buses

In a complex network, such as the one in Santiago, the aforementioned methodology of identifying the point of the previous trip route closest in terms of physical distance to the position of the next boarding cannot be applied directly because in many cases, erroneous points will be identified. An example of this is when a bus route uses the same street in both directions and a person takes a bus for a few blocks to travel to a destination and then takes the same for the return trip. In this case, the next boarding point will be located in the opposite sidewalk (of the same road). Therefore the minimum distance criteria will identify exactly that location as the closest in terms of physical distance. However, the bus already passed very near the next boarding in the initial direction, and it is unlikely that a user would take a long tour just to avoid crossing a street. To overcome this difficulty, we propose using a generalized time instead of distance as the function to be minimized. The alighting point is searched along the trajectory of the bus, known from the GPS database.

The position-time alighting estimate (x_a, y_a, t_a) is that of the bus trajectory that minimizes the generalized time distance with the next boarding position-time. Generalized time (Tg_i) is defined in Eq. (1) as the time associated with position i t_i, plus a variable that represents an estimate of walking time between boarding and alighting, multiplied by a penalization factor. The walking time is estimated as the distance between position i and the position of the next boarding, which is identified by

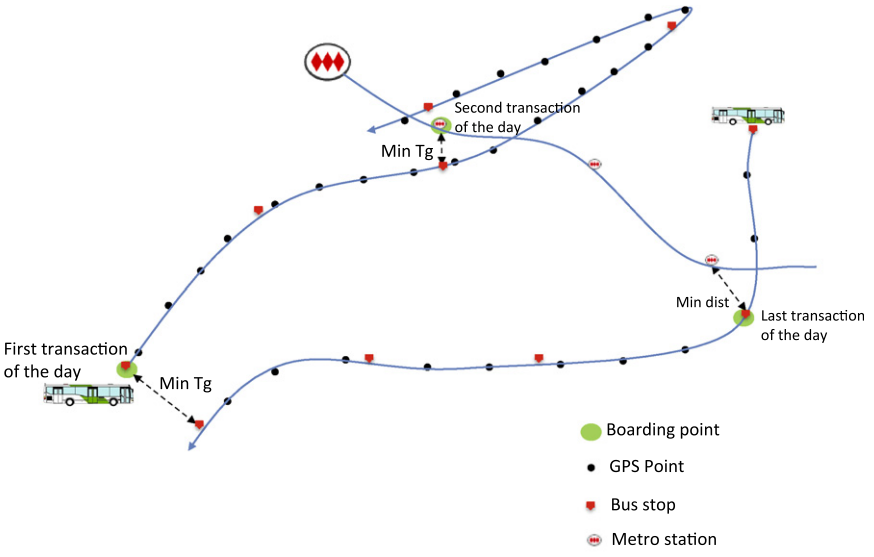

Fig. 3. Alighting estimation model.

a subindex post, $d_{i\text{-}post}$, divided by the average walking speed s_w. The penalization factor, f_w, is obtained from discrete choice models as the disutility of *walking-time* over that of *in-vehicle-travel-time*. The travel time associated with position i can be obtained as t_i minus boarding time.

$$Tg_i = t_i + f_w \cdot \frac{d_{i\text{-}post}}{s_w} \tag{1}$$

The search is then conducted over all i positions of the bus trajectory that are within walking distance (d) of the next transaction position. Therefore, the optimization problem can be written as:

$$\begin{aligned} &\min_i \ Tg_i \\ &\text{s.t.} \\ &d_{\text{i-post}} < d \end{aligned} \tag{2}$$

This will identify a case in which the bus is sufficiently close to the destination to alight and walk, avoiding the aforementioned problem of two-way routes, where the minimum distance point can be very inconvenient in time. This situation is illustrated in Fig. 4, where a passenger boards a line that goes from left to right. The route of that bus goes up to a certain point to the right, and then returns to the left. If the route goes in both directions along the same street, or even if they are not the same street, but are streets that are close to each other, a passenger whose destination is the point designated with *X* in Fig. 4 will not remain in the bus along the entire route and alight exactly at the closest point to his/her next boarding. The passenger will be more likely to alight at the more convenient *i* point because of travel and walking time. This example illustrates an extreme case. However, there are many other cases where the minimum generalized time point must be used to avoid a potential bias.

To implement the method in an efficient and feasible way, a time window is defined for the search in the bus trajectory from the instant when the user boards the bus. This is a parameter of the model, which can be set at different levels for trunk and feeder routes, depending on the characteristics of both types of services. If this constraint becomes active, the limit is doubled because the closest point is then likely to be further away along the bus trajectory.

Another parameter required by the model is the distance assumed as walking distance *d*. The distance a person is willing to walk probably depends on the type of person, type of city, weather, gradient and other factors. For this study, this parameter was set to 1000 m.

If no solution is found for Eq. (2) within the maximum distance threshold, then it is assumed that there is a missing trip or segment, which is most likely to be in another transport mode or using another bip! card. In that case, that trip is labeled as "no alighting point estimation".

4.3. Metro

In the case of Metro trip segments, the boarding and alighting points are Metro stations. The boarding station is known directly from the data, and the alighting station is estimated as that closer (in distance) to the next boarding, within a circumference defined by the walking distance, *d*. If there is no station within that distance, it is assumed that there is a missing piece of information, and the alighting point cannot be estimated.

For those cases where an alighting Metro station is found within the *d* radius, the instant when that alighting occurred must be estimated. Because only the boarding station is known, a Dijkstra (1959) shortest path procedure is implemented to estimate the route followed by the user to go from the boarding to the alighting station. The travel time between stations, waiting time at stations, and walking time inside the station are parameters of this procedure. The values of those parameters are obtained from the operating plan of the Metro. The total travel time is calculated as the summation of the corresponding components.

4.4. Bus station

Probably the most difficult case is that of individuals who board at a bus station where the bip! transaction is made when entering the station, and the user can then board any of the buses from routes that use that bus station as a bus stop. Therefore, in this case, there is an additional problem to be solved: to assign a bus to each transaction made at the bus station. Once a bus has been assigned, the aforementioned procedure for buses can be applied.

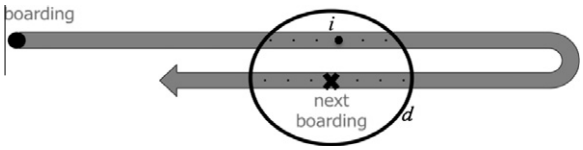

Fig. 4. Search procedure.

First, all of the routes that use that bus station and have at least one bus stop within walking distance from the position of the next bip! transaction are identified. If only one route matching that situation is found, then it is assumed that the user will probably board the first bus of that route that passes through the bus station after the bip! transaction is made. If there is no route that has at least one bus stop within walking distance of the next boarding point, it is not possible to estimate the alighting point. Finally, if there are two or more such routes, an assumption has to be made on which bus the user boards. To do this, the common bus lines concept proposed by Chriqui and Robillard (1975) is applied. The user is assumed to choose a set of routes X that minimize his/her expected time ET from arrival at the bus station to arrival at the bus stop or station of the next boarding transaction. Eq. (3) shows the formulation of the problem and the expression for the expected time ET that depends on average route frequencies f_l, a parameter k that represents the expected waiting time as a fraction of the time interval between buses and a generalized time Tg_l. We take k to equal 1/2, which is valid for homogeneous arrivals.

$$\text{Let } X = \{x_1, x_2, \ldots x_n\} \qquad x_l = \begin{cases} 1 & \text{if } l \in \text{CL} \\ 0 & \text{otherwise} \end{cases}$$
$$\min ET = \left\{ \frac{k + \sum_l Tg_l \cdot f_l \cdot x_l}{\sum_l f_l \cdot x_l} \right\} \tag{3}$$

Once the set of common lines is found, the user is assumed to take the first bus observed after his/her arrival at the bus stop, from any of the common lines, which make up an optimum set. Then, the previously described procedure used to estimate the bus alighting point is used.

4.5. Postprocessing

After the estimation of alighting position and time, we are able to calculate travel times, time between alighting and next boarding, load profiles and other relevant variables. However, to build an OD matrix, it is necessary to differentiate trips from trip segments, identifying the trip destination as the place where a person goes to conduct an activity. We implemented a simple rule, and assume that if a person (card) stays for longer than 30 min in a particular point, then it is a destination. Otherwise, it is a transfer point. This method will fail to identify very short activities and very long waiting times. However, additional information can be included to improve this estimate. For example, if two transactions in a row are made in the Metro or in the same bus route, a destination is assumed between the two, regardless of the time interval, because it is very unlikely that someone would go out of the Metro network during a trip, unless he/she has something to do at that location. Information about the route frequency has not yet been included, but we are exploring that possibility. In the cases where the alighting point of the previous transaction could not be estimated, we assume that if the time elapsed to the next transaction is over 2 h, then this next transaction is the first segment of a new trip.

Because not all trips are identified with this procedure, it is necessary to build expansion factors to account for unobserved trips. There are three cases that require different treatment: (i) trips associated with a bip! transaction where the origin is known but the destination could not be estimated, (ii) trips detected through a bip! transaction for which neither the origin nor the destination are known and (iii) trips not detected through bip! transactions (fare evasion or elusion).

For the first case, we shall assume that the distribution of trips with unknown destination is the same as that of other trips with the same origin. Therefore, we build an expansion factor associated with trips of a particular origin or time period that accounts for all those trips with an unknown destination, as f_{it} defined in:

$$f_{it} = \frac{\sum_j Trips_{ijt}}{\sum_{j \neq null} Trips_{ijt}}, \tag{4}$$

where i is the trip origin, j is the destination according to the zonification of the studied zone (in our case 800 ESTRAUS[3] zones) and t is the time period. A similar factor could be applied to trip segments to build a trip segments OD matrix.

To account for trips detected through a bip! transaction for which neither the origin nor the destination are known, we build a general expansion factor, similar to the previous one, but including only the time disaggregation, as shown in:

$$f_t = \frac{\sum_{ij} Trips_{ijt}}{\sum_{\substack{i \neq null \\ j \neq null}} Trips_{ijt} \cdot f_{ijt}}, \tag{5}$$

A similar treatment should be applied to account for evasion. However, high-quality additional information is required to be able to do that. Unfortunately, the effect of fare evasion and or avoidance is not homogeneous, and apart from underestimating the OD matrix, could induce some biases. For example, in the case of Transantiago, there are two types of fare evaders: the casual evader, who usually does not validate in the first stage of the trip at the feeder service because of the lack of charge points, but in the next stage of the trip, usually in the Metro, charges his/her card and validates and the hard evader,

[3] ESTRAUS is the strategic transport model for Santiago, and its zonification is used for most data collection and modeling initiatives.

who has the strong intention to evade and will not validate in any of the trip segments. Both types of evasion described imply different biases. The first type of evasion described will contribute to underestimation of trips in feeder services, especially those with origins in zones with low commercial activity, and to overestimate trips originating at Metro stations. The latter would probably imply an underestimation of trips in services that operate in high evasion areas. As mentioned before, evasion measures show an uneven spatial distribution of evasion rates.

4.6. Computational implementation

To process these large databases, all the information is stored in PostgresSQL 8.3 files. Several indexes are used to improve the consultation speed. The code to process the data is built in C++, using libpqxx as the interface with the database. The results are visualized on Google Earth using an API for the KML format. All the stages of the procedure described above are implemented as described, with parameters that can be varied by the user, such as walking distance, time thresholds and walking speed, among others.

The overall implementation of the alighting estimation procedure was very time consuming. However, the problem is separable, so it can be distributed to various computers running simultaneously.

5. Application and results

The method described was applied to 36 million observations corresponding to 1 week of operation of the system in March 2009 and to 38 million observations obtained in 1 week in June 2010. In both time periods, the method was able to estimate the alighting point for over 80% of the boarding transactions. In Table 1, we present the success percentages and the distribution of reasons for nonestimation. It can be seen that the most common problem is that no possible alighting point is found within one kilometer of the position of the next boarding (too far). This is probably due to the fact that sometimes people take a different method of transport, such as taxis, car shares, bicycles or walking, for some of the trip segments or because they evade. The second most important reason for not being able to estimate the alighting point is the case of cards that are observed only once in a particular day (single transaction). Data errors are due to problems in the GPS database and the complementary information (for example route path, location of bus-stations); they were an important share of the transactions in March 2009, but the errors decreased to less than 2% in June 2010. This is because both sources of data errors have decreased. The quality of GPS information has improved because it affects the finances of the operators, and the Transantiago authorities have made an effort to improve the quality of the complementary information to obtain more reliable operation indicators. Finally, we also report as nonestimation some obviously incorrect estimates for which the model predicts alighting at the same point of boarding. This could be due to a missing transaction in a public transport mode (evasion or data error) or a missing trip segment in another nonintegrated transport mode.

In Table 2, we present the success rates by type of transaction. It is not surprising that the higher success rates are obtained for Metro transactions, and the lower ones are obtained for bus station transactions, because the latter require more assumptions and are subject to more uncertainty. Bus transactions are in the middle. Even though the figures are not shown in Table 2, we have also found that the success rates are higher for working days and peak hours, and lower for weekend days and off peak hours, and higher for students than adult users. Nevertheless, the global figures are very high compared with those reported by Zhao et al. (2007) and Trépanier et al. (2007). We believe this is partially due to a more sophisticated method but also to the size and quality of the databases used as input.

Given the success in the alighting point estimation, in many cases, we are able to build trip structures, using the 30-min threshold and the simple rules proposed in Section 4.2 to distinguish transfers within a trip from a destination where an activity is to be conducted. After applying these rules, we build the expansion factors f_{it} and f_t defined in Eqs. (4) and (5) to distribute trips without a known origin and/or destination. The average f_{it} over all periods in the entire week and over all origins i is 1.6 for March 2009 and 1.4 for June 2010. The variation coefficients are 0.66 and 0.13, respectively. Note that this correction factor accounts for trips for which any of the trip segments are missing or incomplete. To improve these values, data mining techniques could be used, looking at information for the entire week to make an inference for a particular day when one of the trip segments was not observed.

In Tables 3 and 4, we present OD matrices of trips registered through the automatic fare collection system at an aggregate level for March 2009 and June 2010, respectively. Both matrices were built for an average working day, using the same

Table 1
Percentage success in alighting estimation.

	March 2009	June 2010
% Success in alighting estimation	80.77	83.01
Nonestimation reasons distribution %		
Too far	7.3	7.6
Single transaction	5.2	5.4
Data error	4.3	1.6
Wrong estimate (same location)	2.43	2.39

Table 2
Percentage success by type of transaction.

	March 2009		June 2010	
	Transactions	% Success	Transactions	% Success
Bus	22,546,517	81.7	23,614,012	84.5
Metro	11,383,047	89.0	12,470,020	89.4
Bus station	2127,170	67.1	2034,917	72.2

Table 3
Aggregated OD matrix March 2009.

	North	West	East	Center	South	South-East	D_j
North	157,950	36,389	51,489	78,906	22,988	18,087	**365,810**
West	34,164	294,670	116,217	162,561	37,041	30,525	**675,177**
East	49,382	112,436	317,606	173,157	70,812	150,056	**873,450**
Center	74,593	167,516	160,132	171,932	103,127	96,399	**773,700**
South	22,222	34,877	73,977	104,116	189,216	54,216	**478,624**
South-East	18,379	30,450	158,839	97,234	55,614	250,057	**610,572**
O_i	**356,690**	**676,338**	**878,261**	**787,906**	**478,798**	**599,339**	**3777,333**

Table 4
Aggregated OD matrix June 2010.

	North	West	East	Center	South	South–East	D_j
North	176,291	38,976	57,360	86,022	24,671	20,565	**403,885**
West	37,568	313,822	131,083	176,402	38,491	33,571	**730,937**
East	52,693	120,714	349,172	187,507	71,893	160,822	**942,799**
Center	84,344	178,518	168,284	176,849	106,518	100,024	**814,538**
South	24,658	37,350	83,124	113,291	194,636	56,133	**509,192**
South–East	20,115	33,326	170,041	104,565	56,563	255,868	**640,478**
O_i	**395,668**	**722,707**	**959,064**	**844,636**	**492,772**	**626,982**	**4041,830**

aggregate zonification used to present aggregate results of the last OD survey in Santiago (Dictuc, 2003). Even though these results cannot be directly compared with those of the matrix generated with the OD survey, we observe that the structure of the matrices obtained is reasonable, with more trips in the diagonal (short trips) and important underestimation of trips in the poorer zones, probably due to fare evasion.

Even though significant changes to the public transport system, including extensions to the Metro network, were implemented between March 2009 and June 2010, the trip structures in both time periods are remarkably similar. The difference in the total number of trips in March 2009 and June 2010 could be due to a seasonal effect.

Because the origin and destination information is available as a time–space coordinate, these matrices are available at any desired disaggregation level[4]. However, the expansion factors are only reliable up to the level of disaggregation used to build them (800 ESTRAUS zones). Also, it must be noted that these matrices include only paid trips. Additional information is required to account for fare evasion.

6. Discussion and conclusions

We have proposed and implemented a method to obtain an Origin–Destination matrix from smartcard data for a multimodal public transport system based on the estimated alighting stop for trip segments and have proposed a few simple rules to identify activity stops. The main contributions of this paper are: moving from a trip segment perspective to a trip Origin–Destination estimation, and using generalized time rather than physical distance to estimate alighting point for transfers and destinations.

Even though some information is missing, such as evasion and trip segments in nonintegrated transport modes, and validation with exogenous data needs to be conducted, the magnitude of the achieved success rates suggests that this new source of information has a great potential and can actually replace an important part of the large Origin–Destination surveys. However, this is only an initial step, and more research is required to obtain a reliable origin destination matrix from the data available. This methodology has to be validated, using a control sample to verify the results and the validity of the

[4] Video of boarding (Yellow) and alighting (Green) OD flows at bus stop level is provided as supplementary material.

assumptions. In addition, some additional information is required regarding fare evasion to account for unobserved (unpaid) trips. The simple rules used to identify activities can be replaced by more sophisticated rules that include additional information, such as land use at the stop and frequency of the service used by the traveler. Finally, independent measurements of passenger flows could be used to correct for unobserved sources of bias.

The matrices obtained with this methodology are different in nature from those obtained with traditional Origin–Destination surveys, where the sample size is limited by budget constraints, and usually cannot adequately show the disaggregate distribution of trips. The proposed methodology can recover the trip distribution structure quite well, as the "sample size" is over 80%. However, the total number of trips have to be validated. Alighting stop estimation information has already been used in the exploratory analysis for the redesign of Transatiago services to build load profiles of bus routes.

We identify three clear lines for further research: (1) validation, using a control sample and complimentary measurements, (2) identification of activity stops with more sophisticated methods and (3) use of data mining techniques to complete missing information.

Acknowledgements

Funding: Fondecyt 1090204, FONDEF D10I-1002, ISCI (ICM P-05-004-F, CONICYT FBO16). We thank the support of Transantiago and the collaboration of Mauricio Zúñiga and Daniel Fischer. Previous versions of this paper were presented at WCTR 2010 and TRB 2011.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.trc.2012.01.007.

References

Barry, J.J., Newhouser, R., Rahbee, A., Sayeda, S., 2002. Origin and destination estimation in New York City with automated fare system data. Transportation Research Record 1817, 183–187.

Beltrán, P., Cortes, C., Gschwender, A., Ibarra, R., Munizaga, M., Ortega, M., Palma, C., Zuñiga, M., 2011. Obtención de información valiosa a partir de datos de Transantiago, XV Congreso Chileno de Ingeniería de Transporte.

Beltrán, P., Gschwender, A., Palma, C., 2012. The impact of compliance measures on the operation of a bus system: the case of Transantiago. Research in Transport Economics (RETREC), submitted for publication.

Buneman, K., 1984. Automatic and passenger-based transit performance measures. Transportation Research Record 992, 23–28.

Chapleau, R., Chu, K.K., 2007. Modeling transit travel patterns from location-stamped smart card data using a disaggregate approach. In: Presented at the 11th World Conference on Transportation Research, June 24–28 2007, Berkeley, CA.

Chapleau, R., Trépanier, M., Chu, K.K., 2008. The ultimate survey for transit planning: Complete information with smart card data and GIS. In: Presented at the 8th International Conference on International Steering Committee for Travel Survey Conferences, Lac d'Annecy, France.

Chriqui, C., Robillard, P., 1975. Common bus line. Transportation Science 9, 115–121.

DICTUC, 2003. Actualización de encuestas Origen Destino de viajes, V Etapa. Informe Final a Sectra, Santiago.

Dijkstra, E.W., 1959. Note on two problems in connection with graphs (spanning tree, shortest path). Numerical Mathematics 1 (3), 269–271.

Furth, P.G., Hemily, B.J., Muller, T.H.J., Strathman, J.G., 2006. Uses of Archived AVL-APC Data to Improve Transit Performance and Management: Transportation Research Board, TCRP Report No. 113.

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 Communications, Networking and Mobile Computing WiCom 2007, 3147–3150.

Muller, T.H.J., Furth, P.G., 2001. Trip time analyzes: key to transit service quality. Transportation Research Record 1760, 10–19.

Ortúzar, J.de D., Willumsen, L.G., 2011. Modelling Transport, fourth ed. Wiley, Chichester.

Pelletier, M.P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transportation Research Part C 19, 557–568.

Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems 11, 1–14.

Zhao, J., Rahbee, A., Wilson, N., 2007. Estimating a rail passenger trip origin–destination matrix using automatic data collection systems. Computer-Aided Civil and Infrastructure Engineering 22, 376–387.