PUBLIC TRANSPORT OD MATRIX ESTIMATION FROM SMART CARD PAYMENT SYSTEM DATA

Marcela Munizaga, Universidad de Chile, Casilla 228-3, Santiago, Chile
mamuniza@ing.uchile.cl

Carolina Palma, Transantiago, Moneda 975, Santiago, Chile
carolina.palma@transantiago.cl

Pamela Mora, Universidad de Chile, Casilla 228-3, Santiago, Chile
pmora@ing.uchile.cl

ABSTRACT

Many cities in the world have incorporated information technology to their public transport systems, and continue advancing along these lines. Santiago, Chile is not an exception, and the new public transport system Transantiago has introduced GPS bus location and smartcard payment systems. However, this particular case has some characteristics that make it particularly interesting for passive data collection: most payments are made by smartcard, therefore a very high percentage of boarding transactions are recorded in a huge Transactions database and all buses are equipped with GPS device that generates an even bigger Positions database. These characteristics, which are not yet present in other transport systems, will most likely be the standard in the future. This represents an excellent opportunity for public transport planners, as cards (users) can be followed through the system to identify their travel patterns. This source of information has so much space-time detail that once processed would permit analyzing not only mean attributes at any desired level of time-space disaggregation, but also variance and regularity of behaviour. This paper describes the data and some of the potential applications, and shows some preliminary results. A method is proposed to estimate boarding and alighting bus stops, and to estimate travel time and time assigned to activities between trips. The method is applied to a sample of the database. Some preliminary results and the success rates are shown.

Keywords: public transport, OD matrix, passive data collection

INTRODUCTION

In the late 90s smartcard payment systems were incorporated in some cities such as Washington (Smartrip) and Tokyo (Suica). This new technology rapidly spread to other cities, and nowadays it has become an important part of the public transport fare collection system. For example the Oyster card was implemented in London in 2003, with discount fares (compared with buying single tickets) and it is currently the most popular payment method. In Chicago, (Zhao et al, 2007) the Chicago card has a very high penetration rate. Other examples, with different ways of implementation and different levels of penetration are: San Francisco (Buneman, 1984), Portland (Furth et al., 2006), New York (Barry et al., 2002), Netherlands (Muller and Furth, 2001; Furth et al., 2006), Changchun (China) (Lianfu et al., 2007) and Quebec (Gatineau, Quebec, Canadá) (Trépanier et al., 2007; Chapleau and Chu, 2007; Chapleau and Chu, 2007; Chapleau et al., 2007; Chapleau et al., 2008). In all 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 by far the most important in the Metro (99%); therefore, the penetration rate is very near 100%.

The research challenge of obtaining valuable information from the data generated by smartcard transactions has been taken by several researchers, who recognize its potential to improve public transport planning and operation. The MADITUC research group has developed several methods to obtain the information and to improve its quality. Chapleau and Chu (2007) propose a method to identify and replace incorrect or suspicious observations from the automatic fare collection system. Trepanier 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 Origin-Destination (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 Origin-Destination (OD) matrices from an origin only automatic fare collection system, where the position of the buses in known through an Automatic Vehicle location system.

The research efforts have focused in the integration and enrichment of the information available from different passive sources (such as automatic fare collection systems, automatic vehicle location systems, passenger counts), detection and correction of information errors, estimation of alighting or destination point, identification of transfers, generation of origin destination matrices from the information available.

This paper presents a methodology that goes a step further in terms of the dimension and complexity of the public transport system, as it is applicable to a large scale, multimode public transport system. 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 the methodology proposed is presented, section 4 contains the results of the preliminary application. Section 5 concludes.

THE DATA

The data available comes from Transantiago, the public transport system 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 almost without intersection. There are also six trunk operation areas, which are larger and have intersection between each other and with the feeder areas. Trunk operator 1 is the Metro, the other five are bus lines. The payment system is such that each passenger pays a fare when entering the system and that allows him/her to make up to three combinations within a two hours time window. The payment structure is slightly different in the buses and metro parts of the system. 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!; however, the percentage of users who buy single ticket in Metro is very small (estimated in around 1%). The fare is also different, being slightly higher for Metro at peak hours; buses have flat fare. If a passenger uses a bus first and metro afterwards, the difference between both fares will be charged when entering the metro system. All the metro lines are connected, and changes between them are made without showing the bip! card again, but if someone exits the metro and re-enters, will be charged the full fare again.

As a very general description of Santiago, it is the capital city of Chile, it is divided into 34 districts, it has nearly 6 million inhabitants, and the distribution of the population is not homogeneous; there are clearly identifiable wealthier and poorer neighbourhoods. The city has a circular shape, with a large proportion of trips going from the suburbs to the centre in the morning, and from the centre to the suburbs in the evening. According to the 2001 Origin Destination survey, there were 16 million trips in a working day, and from them 10 million were motorized trips (38.6% of trips were walking or bicycle trips). The average household size was 3.81, and the trip rates were 2.82 trips per person, 10.76 trips per household. The market share of public transport was by then 53%.

There were severe problems at the beginning of the operation of the system, but most of them have been solved, and the system is now operating normally, although some problems persist in certain areas. An evasion problem has been detected in the buses, biased towards certain geographical areas, mainly poor neighbourhoods located far from the city centre. Evasion is almost inexistent in Metro. The system contains over 300 bus routes, and nearly 6,000 buses operating daily. 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 (basically a fence) equipped with extra vehicle payment system where passengers pay when entering the station, to increase the boarding efficiency. These bus-stations, called “Zonas paga”, operate mainly 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 (put his/her card close to a payment device) when entering a bus, a bus station or a metro station. Each payment device has an id and is associated to a bus, a metro station or a bus station. The information recorded for each transaction includes the card id and its type, time and date when the transaction was made, bus or site where the transaction was made, and the amount of money paid. Every week there are around 35 million bip! transactions, made by over 3 million bip! cards in nearly 6,000 buses.

Another database contains geo-coded information of all the buses, such as latitude and longitude, time and date, and instant speed. Each bus is identified with its plate number, and also the operator to whom it belongs. In most of the observations, the position is available every 30 seconds. In some cases this period is longer, and shorter in others. Every week of data contains around 80 million GPS observations. Figure 1 shows the position of the buses at an instant. The most important corridors can be clearly observed.

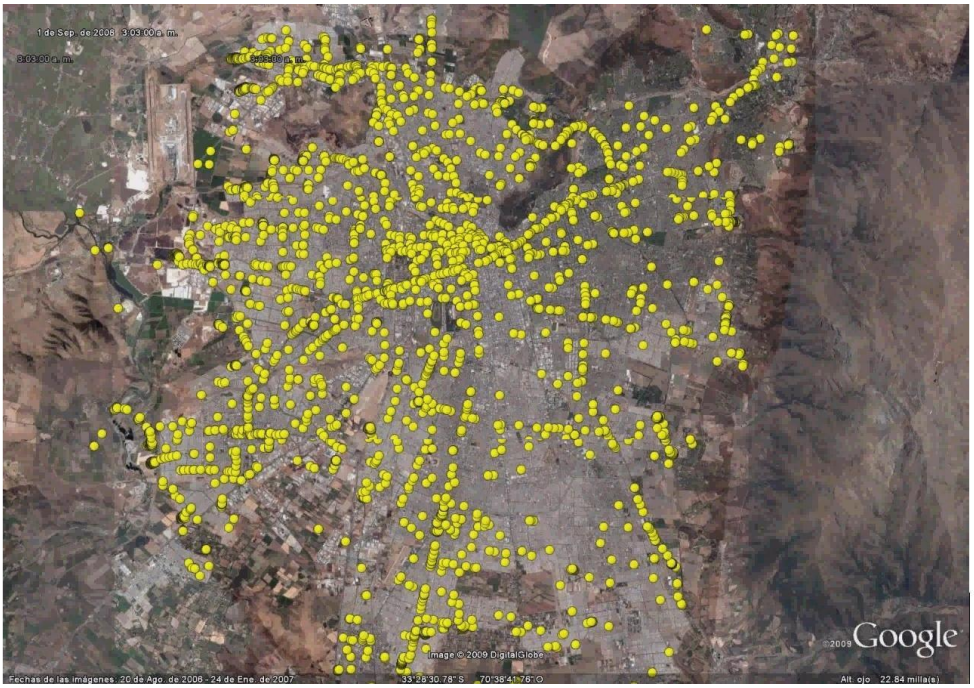

Figure 1 - Position of buses

The geocoded bus routes, and position of Metro stations, bus-stops and bus stations is also known and valuable information. There are timetables associated to bus and metro services, and also for 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 information about the service each bus is giving in a certain period. The Transantiago Authority using a triangulation procedure generated this information. They defined three points of each service route, if a bus passes through the three points in a certain period, then it is assigned to the service. This process has been evaluated both by the Transantiago Authority and by the operators and proved to be reasonably reliable.

Statistical description of the data

A descriptive statistical analysis is conducted for one particular week (1-7 September 2008) from the database. Looking at the transactions information, it can be observed that 44% of the bip! transactions (boarding) are made in buses from trunk operators, 36% in metro stations, and 20% in buses from feeder operators. From the bus transactions, the vast majority are made directly in the bus and less than 10% are made in bus-stations. The number of transactions along the week show very similar numbers for working days of around six million transactions per day, during the weekend this number falls to less than

four million on Saturday, and less than two million on Sunday. The relation between paid and zero cost transactions is 1.63, which gives an initial idea of the number of stages per trip.

Figure 2 shows the time distribution of boarding transactions along a working day by mode. Morning peak and evening peak can be clearly observed close to 8 AM and 7 PM respectively. A much smaller but still noticeable midday peak is observed between 1 and 2 PM. Note that the metro peak and bus peak occur at different times (bus peak earlier, specially in the morning). Saturday and Sunday (not shown here) have much less transactions. Saturday has a pattern that starts early in the morning (as early as in a working day), but rises only up to 60,000 transactions per hour during the afternoon. Sunday shows very little activity, with less than 40,000 transactions per hour along all day and no clear peak at any time.

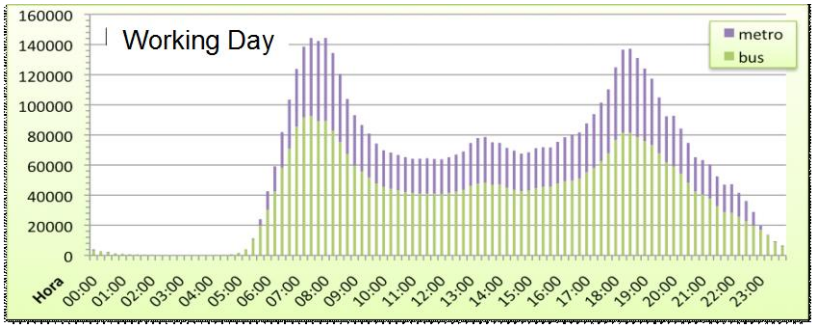

Figure 2 - Time distribution of transactions along a working day

Figure 3 shows the boarding profile by type of user. It can be seen that time profiles are different for students and adult passengers. The morning peak and evening peak are slightly earlier for students. Also, the evening peak is less pronounced for them. This information is valuable for policy measure evaluation; however, more precise information such as load profiles is required for planning and design purposes.

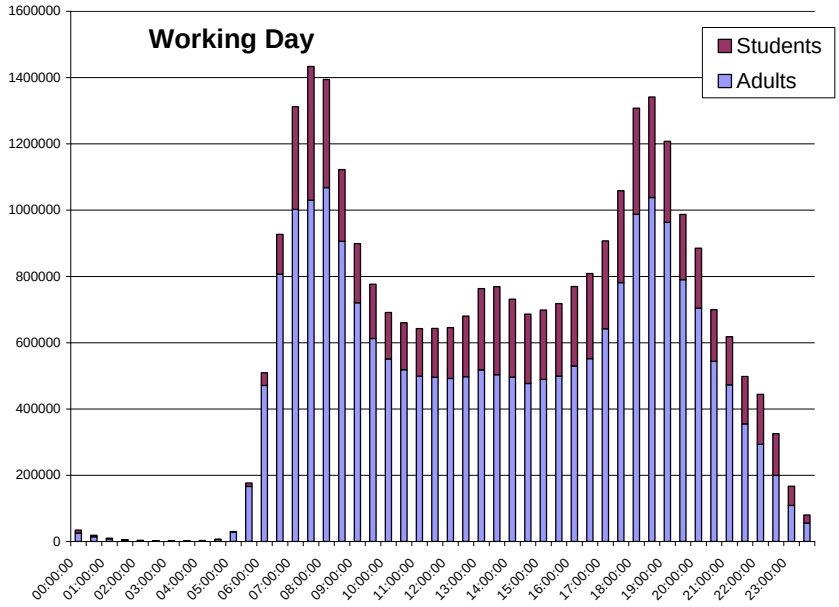

Figure 3 - Boarding transactions by type of user

Matching the Transactions database with the Positions database through bus plate or Metro/Bus-station code and time, it is possible to identify the position where the transaction was made in 98.5% of the cases. This information can be used to make a spatial analysis of transactions, as shown in Figure 4 for boarding transactions at bus stops of any route. The aggregate analysis of all routes, over time clearly shows that morning peak time is different in different locations, remarkably earlier in poorer neighbourhoods. The amount of information gathered permits as much time and space disaggregation as required to perform this kind of analysis.

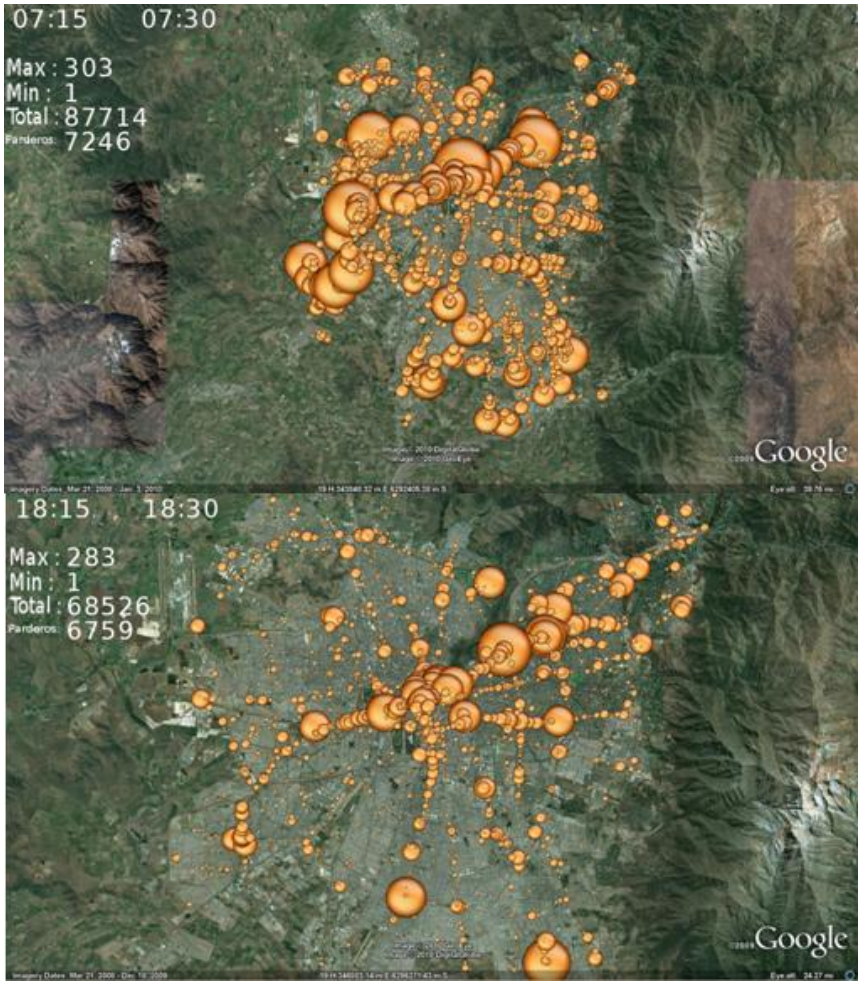

Figure 4 - Boarding transactions at different times

METHODOLOGY

Based on the work by Trepanier et al (2007) 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, 1994). Each trip can have one or more stages, 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 stage begins and ends, respectively.

Proposed method

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 behaviour, build origin-destination matrices, estimate vehicle load profiles, and many other tasks in a simple and direct way. The proposed model has several components, as shown in Figure 5. The inputs of the model are three main databases: transactions (boarding) from automatic fare collection system, vehicles position from the automatic vehicle location system and 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 alighting point. The estimation procedure is described below. It it different for transactions in buses, bus stations and Metro stations, but in all three cases the result of it is an estimate of the position-time coordinates of the alighting point. Then, using this information, our proposed method includes a module to distinguish transfer from destination, identifying trip stages. As a result of this procedure trips and trip stages 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 allow building the public transport trip chain of the person behind that card. On the other hand, there are some cases where the estimation of the alighting point is not possible. Special interest was placed into those cases also.

Alighting point estimation

To estimate the alighting point it is assumed that the next bip! transaction is posterior to the alighting. Following Trepanier et al (2007) it is also assumed that the alighting bus stop is close to that of the next boarding. This is only possible to apply when both the current transaction and the next one have position information (from the AVL database). In the case of the last transaction of the day, it is assumed 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 (card). If there is only one trip per card, no imputation is possible with single day information. The model is shown in Figure 6, where the three possible cases are illustrated: next transaction in a bus, a metro station or a bus station; as said before, depending on that, the estimation procedures are different.

In a complex network such as the one in Santiago the Trepanier et al (2007) methodology of identifying the point of the previous trip route closest (distance) to the position of the next boarding cannot be applied directly, because in many cases an erroneous point will be identified. An example of this is when a bus route uses the same street in both directions, a point from the return direction might be the closest, but the bus already passed very near the next boarding in the initial direction. To overcome this difficulty, it is proposed to use generalized time instead of distance, as the function to be minimized.

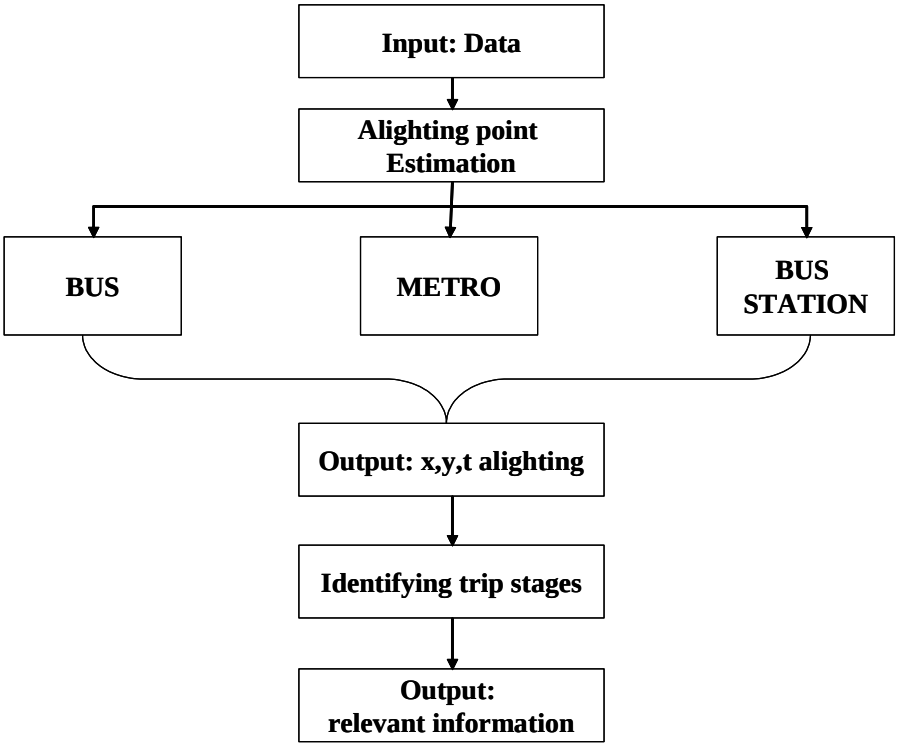

Figure 5 - proposed method structure

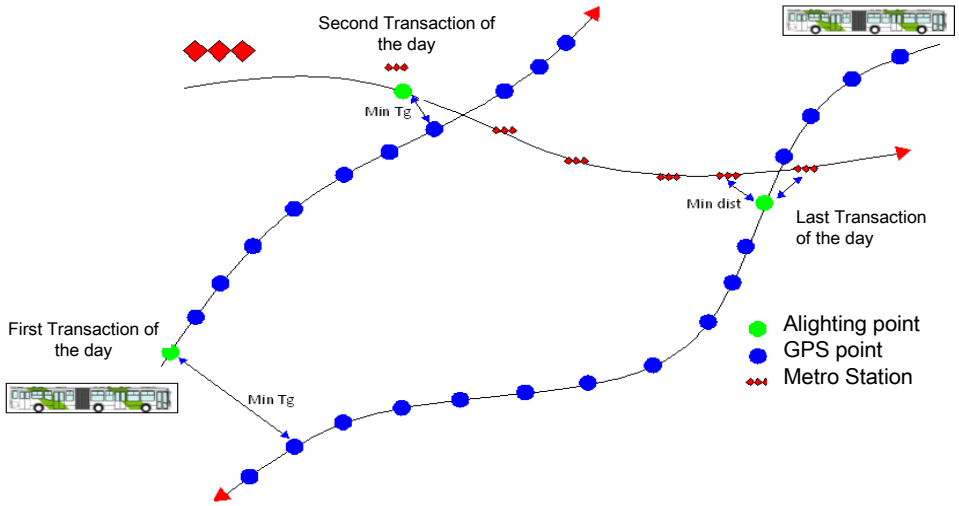

Figure 6 - Alighting estimation model.

If the previous trip or trip stage is in a bus, the alighting point is searched along the trajectory of that particular bus, known from the GPS database. The position-time alighting estimate (x_a,y_a,t_a) is the position-time of the bus trajectory that minimizes the generalized time distance with the next boarding time-position. In equation (1) generalized time (Tg_i) is defined as the time associated to position i t_i, plus the distance between position i and the next position identified by sub index post: $d_{i\text{-}post}$ divided by the average walking speed s_w and multiplied by

a weight factor f_w representing the disutility of walking time as a proportion of in vehicle travel time.

$$Tg_i = t_i + f_w \cdot \frac{d_{i-post}}{s_w} \quad (1)$$

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

$$\text{Min } Tg_i \quad (2)$$
$$s.t.$$
$$d_{i\text{-}post} < d$$

This will identify a case where 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 terms of time. This situation is illustrated in Figure 7, 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 ways along the same street, or even if they are close (but not the same) streets, a passenger whose destination is the point designated with an X in Figure 7 will not remain in the bus along the whole route to alight exactly at the closest point of his/her next boarding, s/he will rather alight at the more convenient *i* point considering travel and walking time.

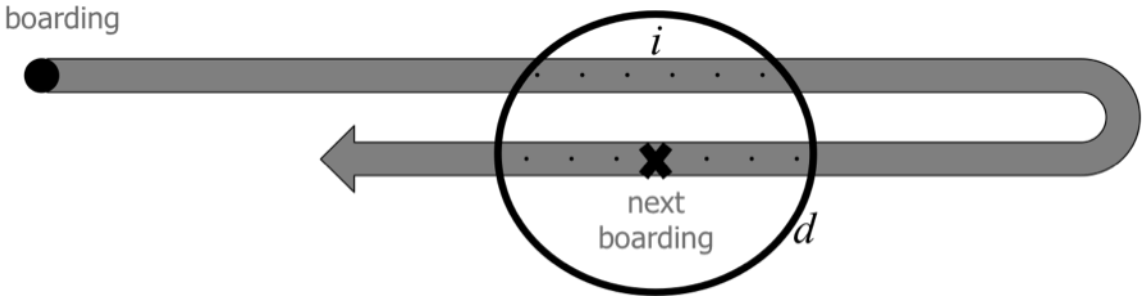

Figure 7 - search procedure illustration

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 constrain becomes active, the limit is doubled, because then the closest point is likely to be further away along the bus trajectory.

Another parameter required by the model is the distance that can be assumed as walking distance d. It probably depends on the type of person, the type of city, the weather and many other factors. This parameter was initially set to 1,000 metres.

If no solution is found for equation (2) within the maximum distance threshold, then it is assumed that there is a missing trip or stage, probably in another transport mode or using another bip! card. In that case that trip is labelled as one where the alighting point cannot be estimated.

Metro

In the case of Metro stages, 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 part of information and the alighting point cannot be estimated.

For those cases where an alighting metro station is found within the d ratio, the instant when that alighting occurred must be estimated. As 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, detention time at stations, and walking time inside the station are parameters of this procedure. The total travel time in Metro is the sum of the corresponding components.

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 into any of the buses from routes that use that bus station as 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 buses procedure can be applied.

As a first stage all those 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 in 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 is boarded by the user. 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 that minimize his/her expected travel time, and board the first bus of that optimum set. Observed frequencies are used to implement this procedure.

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. The previously described procedure used to estimate the bus alighting point is used.

Identifying trip stages

Once the alighting point is estimated, the travel time can be calculated as the time elapsed from boarding to alighting. Furthermore, the time difference between estimated alighting and the next boarding transaction could also be calculated. Using this information, some attempts have been made to separate destination from boarding points. In this initial application, a destination was defined as any point where the person (card) stays for longer than 45 min.

With this procedure we obtain full information of a percentage of the trips that can be used to obtain origin-destination matrices, load profiles, and other relevant variables.

APPLICATION AND PRELIMINARY RESULTS

The described method was applied to a sample of 63,221 observations, randomly obtained from a one-week database (September 1-7, 2008). From that sample, the method was able to estimate the alighting point in the vast majority, as shown in Figure 8. The small percentage labelled as *Incorrectly estimated* corresponds to cases where the estimated alighting point is the same of the boarding point. There are a few reasons why this can happen; one example is when a person takes a service at a particular site to go somewhere, then come back to the same point by any other mean (walking, car, taxi) and again takes a service at the same site. There might be other errors in the 82% of cases where the method was able to estimate alighting point, but it is not possible to identify them without exogenous information.

Figure 9 shows the main reasons for not being able to estimate the alighting point in the remaining 15% (9,705 observations). It can be seen that nearly 42% of the failures are due to the missing trip/stage case, where the next transaction position is too far away from the boarding route to assume a transfer walk. The second most important reason is the GPS data, which is not available for either the initial or posterior bip! observation in around 25% of the cases. Another 17% is due to cards that have only one transaction in a particular day. The remaining 16% are due to errors in the complementary data, such as position of bus stations, route definition and service assignment.

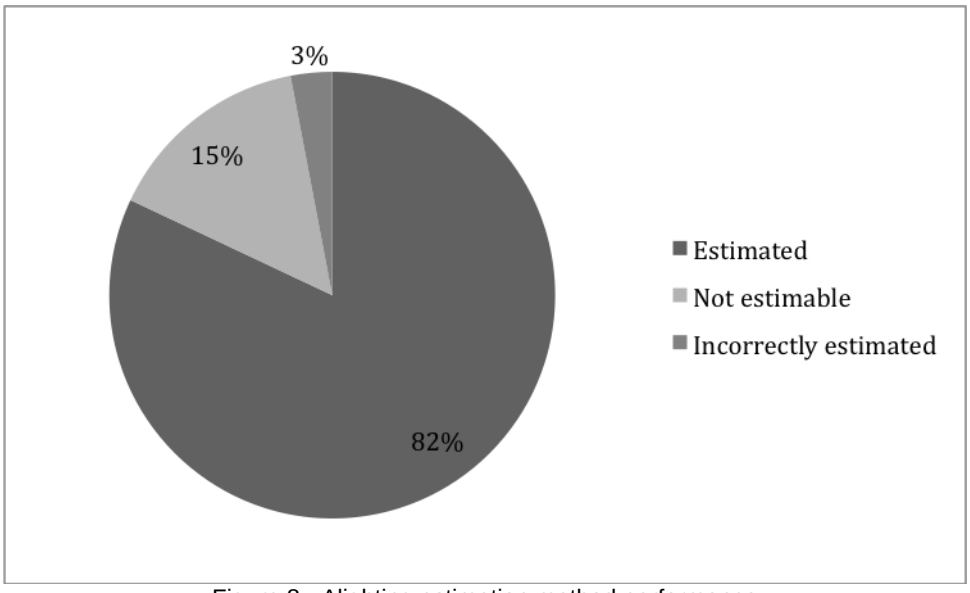

Figure 8 - Alighting estimation method performance

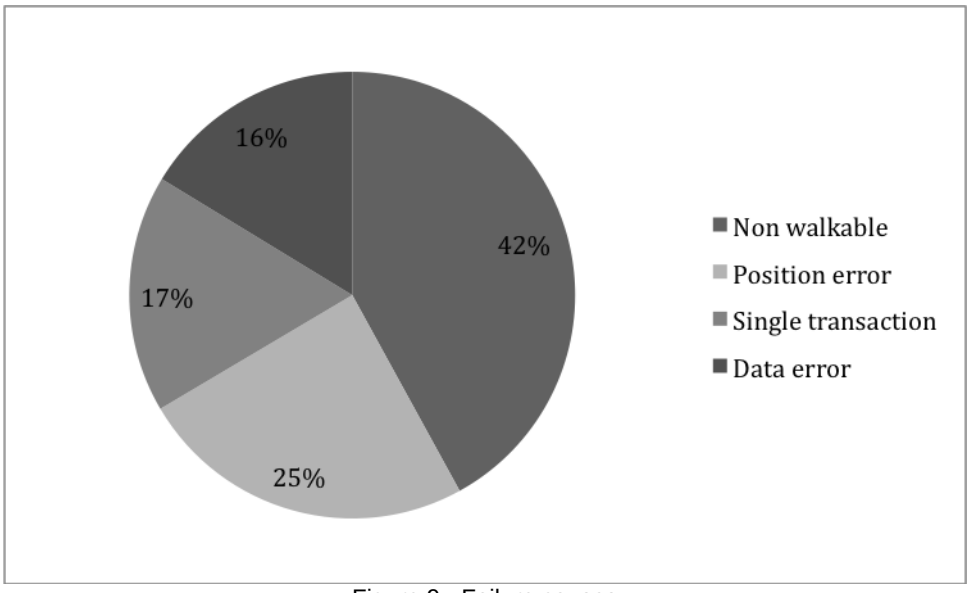

Figure 9 - Failure causes

Applying this procedure, the alighting position and time was obtained for 51,759 boarding transactions, fully defining that amount of trip stages. Then, the simplified rule to separate transfers from destinations was applied. If the time between estimated alighting and next boarding was more than 45 minutes, it was assumed that there was an activity conducted by the user at that point, and that position was identified as a trip destination. In this way, it is possible to separate trips from trip stages. Figure 10 shows the destinations by municipal district for working day, Saturday and Sunday to illustrate the spatial distribution of the destinations. Trip destinations are clearly concentrated in the more commercial zones of the city.

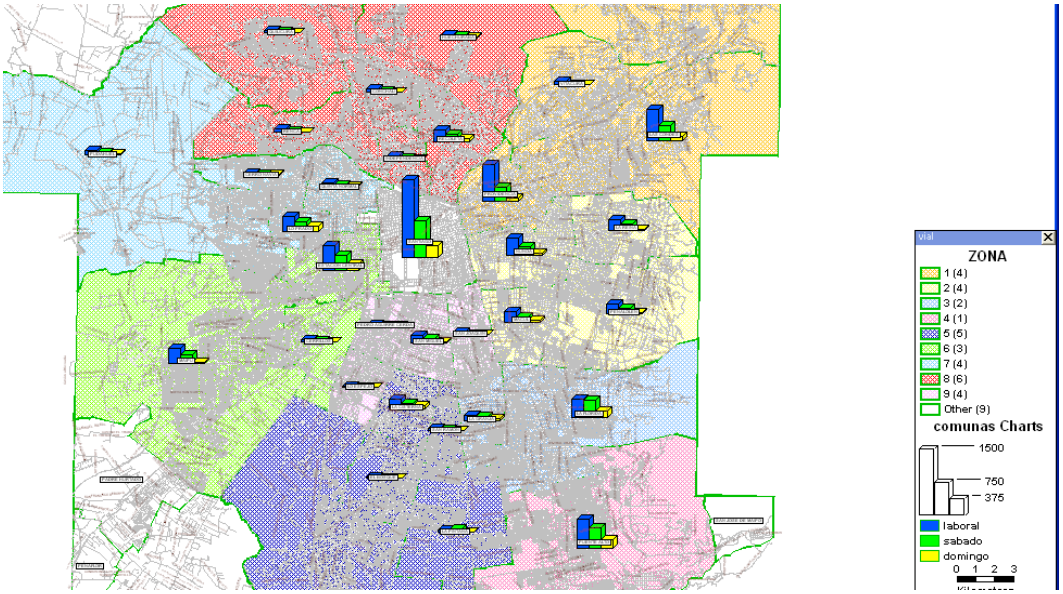

Figure 10 - Destination by Transantiago zone

Figure 11 contains the trips per day histogram for working days, saturday and sunday. Note that two trips per day is the most common figure, especially during working days. This is consistent with information obtained from surveys.

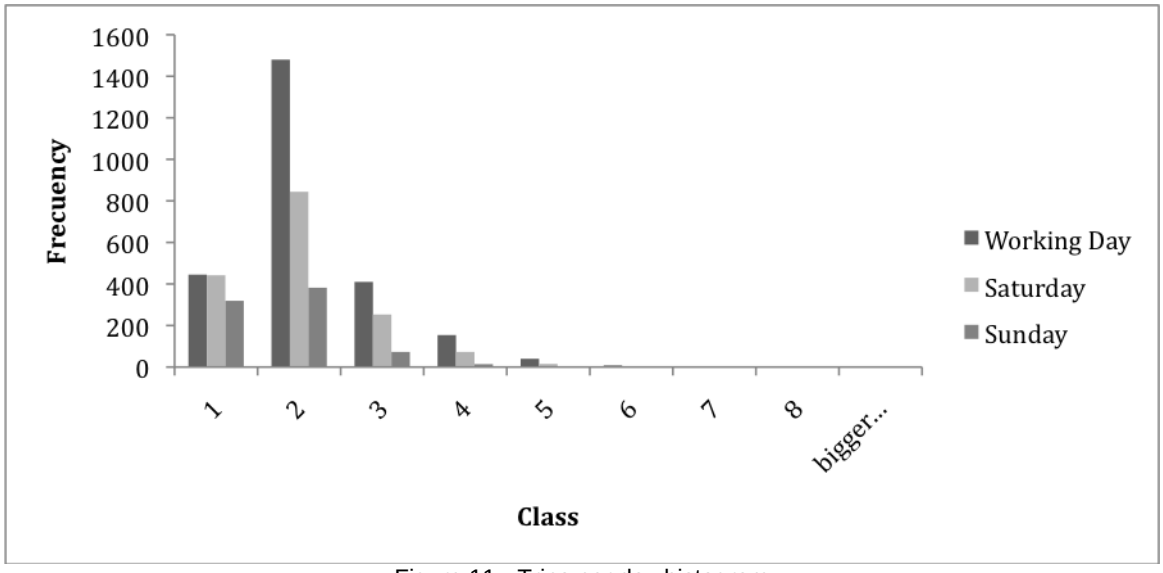

Figure 11 - Trips per day histogram

Figure 12 shows the number of stages per trip. It can be seen that the majority are one stage trips, but there is a large proportion of two stages trips also. Trips of three or more stages are less common.

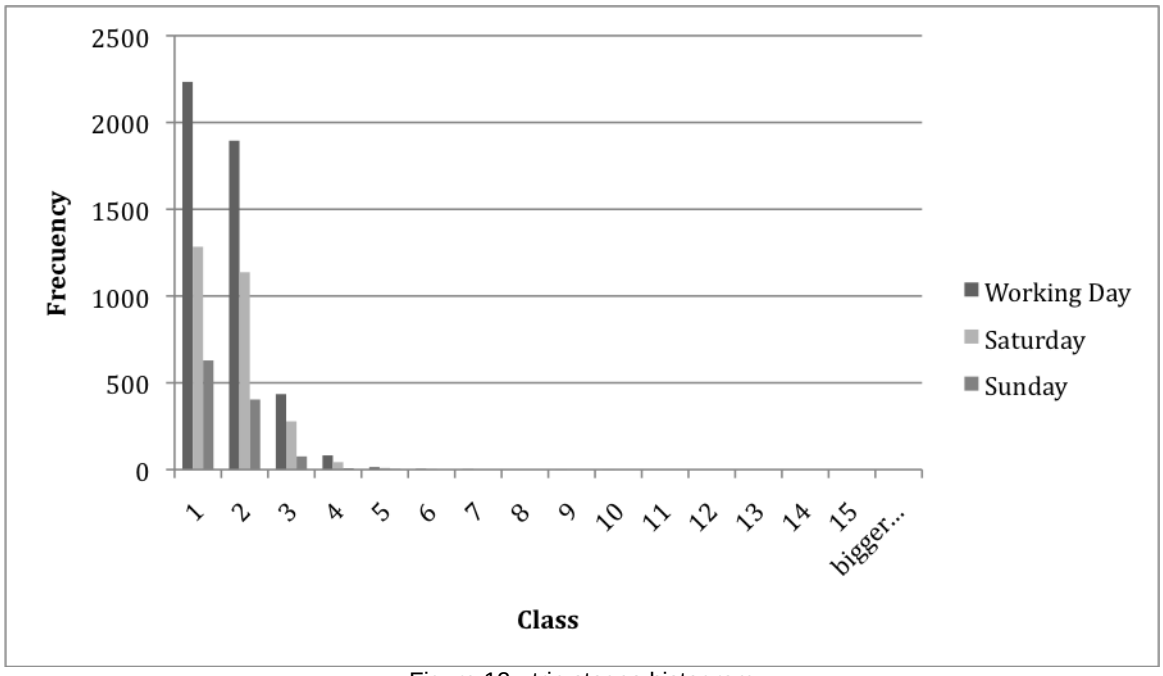

Figure 12 - trip stages histogram

Finally, Table 1 presents the estimated Origin-Destination matrix at an aggregate level. The structure of this matrix is similar to that of the last OD survey in Santiago. However, to compare the figures, this matrix must be expanded with the appropriate correction factors.

Table I - Origin-Destination matrix for the subsample

O/D	North	West	East	Center	South	South-East	O_i
North	552	145	195	410	115	125	1542
West	122	1093	660	983	125	196	3179
East	208	562	1557	1126	404	912	4769
Center	374	824	961	889	509	748	4305
South	124	150	428	612	476	264	2054
South-East	117	177	972	754	217	1261	3498
D_j	1497	2951	4773	4774	1846	3506	19347

Even if the method was applied to the whole sample, there will be some errors that will require external information to build correction factors. One of these errors is due to fare evasion, which, according to some preliminary measures, is not homogeneously distributed in the city. Another source of bias might be the use of other transport modes for some of the trips. For example, users of higher income have more options than poorer users that are captive to public transport. The next stage of this research is to define a methodology to obtain the complementary information necessary to build correction factors.

CONCLUSIONS

A method to obtain information from automatically generated data in a large and complex public transport system such as Transantiago has been presented. The alighting estimation method is quite robust. Its preliminary application shows promising results, as it was possible to estimate alighting position-time in 82% of the boarding transactions. This information can be analyzed with as much time-space disaggregation as required. We believe this will change the way public transport planning is conducted. Santiago is a privileged case study, as smartcard penetration is almost 100%, and all buses are equipped with GPS device; however, this will probably be the standard in many cities in the near future. The size of the databases is a challenge, as some processing can take several hours to run in powerful computers; therefore, only simplified processes can be done in real time. There are some limitations of the method that are unlikely to be overcome without additional information. One of them is the case where only a single transaction is observed for a particular card in a particular day. Another limitation is that there is no information available about non-integrated modes such as shared taxi, taxi and car. If users take one of these to reach the metro or bus network, then there is a missing piece of information that will induce an error in the estimation of the trip chain of those users.

In terms of further research, first of all the method and its results have to be validated, which is something we are already working on. Secondly, the module that distinguishes transfers from destinations must be refined. There is information available that can help to this process. For example, frequency of bus services and land use at the position. A 20 min stop in a commercial zone with very frequent bus services is probably a short activity, while a 20 min stop in a zone without commercial activities and infrequent bus services might be a very bad connection. In the future, we plan to validate the information obtained using a control sample; it is very important to contrast the results of our method with the users behind those cards. The other important issue for further research are expansion factors to build an Origin-Destination matrix correcting for potential biases. If the control sample is large enough, it can be used also to build expansion factors.

ACKNOWLEDGEMENTS

This research was partially funded by Fondecyt grant 1090204, PBCT Redes Urbanas and the Millennium Institute Complex Engineering Systems (ICM P-05-004-F, CONICYT FBO16). We specially thank the collaboration of Mauricio Zúñiga and Flavio Devillaine.

REFERENCES

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

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

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

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

Chriqui, C. and Robillard, P. (1975) Common bus line. Transportation Science 9, 115-121.

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

Furth, P.G., Hemily, B.J., Muller, T.H.J. and 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. and 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. and Furth, P.G. (2001) Trip time analyzes: Key to transit service quality. Transportation Research Record 1760, 10-19.

Ortúzar, J. de D. and Willumsen, L.G. (1994) Modelling Transport, Second edition. Wiley, Chichester.

Trépanier, M., Tranchant, N. and 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. and 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.