Journal of Transport Geography 66 (2018) 330–339

Contents lists available at ScienceDirect

Journal of Transport Geography

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

Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis

Margarita Amaya[1], Ramón Cruzat[2], Marcela A. Munizaga*

Universidad de Chile, Casilla 228-3, Santiago, Chile

ARTICLE INFO

Keywords:
Smartcard
Public transport
Zone of residence
Travel time
Time use

ABSTRACT

Public transport systems with electronic fare collection devices continuously store data related to trips taken by users, which contain valuable information for planning and policy analysis. However, if the card is not personalized, there is no socioeconomic information available, which imposes a limitation on the types of analysis that can be performed. This work presents a simple method to estimate the residence zone of card users, which will allow socioeconomic variables to be estimated, thereby enriching the analytical possibilities. The method, which is based on the observation of morning transactions of frequent users, is applied to a sample of over 2 million cards. The method is evaluated using a sample from the Santiago ODS where users declared their card id and also declared their home address. A sample of 888,970 cards that are observed at least three days in a week and show spatial regularity for the morning transaction is used for zone of residence estimation and analysis of travel patterns and time use. The results show that users who live in the city center or in the wealthier East zone experience lower travel time, spend more time at home and less time at work.

1. Introduction

Cities have the purpose of satisfying the citizens' needs to interact and conduct different activities such as work, study or leisure. Transport systems are designed to allow the required movements within the city efficiently and effectively. However, these goals are not always achieved. Cities are complex systems, with a vast amount of elements, which are sometimes difficult to predict and to govern. Transport systems are also complex, with time-space relations that are evident and absolutely present for the user, but difficult to observe by an external modeler. To be able to understand the time use and travel patterns within a city, it is necessary to observe such behavior. The most common empirical studies of time use and mobility are conducted at a disaggregate (individual) level, with samples obtained from surveys. However, surveys are expensive and difficult to conduct, and there is a trade off between survey length and complexity, and sample size and accuracy (Jara-Díaz and Rosales, 2015). So the information available for conducting such studies is limited.

We now have information available from different technological devices such as mobile phones, which can be observed when they are connected to the antennas. The public transport systems are usually equipped with automatic vehicle location (AVL) devices installed in vehicles and automatic fare collection (AFC) systems. This information allows observing how people move through the city for large segments of the population, such as public transport users or the clients of a particular mobile phone company. However, usually there is no sociodemographic information associated to the movement traces observed, and sociodemographic characteristics are relevant to explain time use and travel behavior. The purpose of this study is to explore the possibility of using AVL and AFC data to make travel pattern and time use analysis. We make the hypothesis that zone of residence is closely related to sociodemographic characteristics. Therefore, a key aspect of the methodology is to estimate the zone of residence of users, which is used as a segmentation variable. The proposed method is applied to Santiago de Chile.

As in many large cities, the public transport system of Santiago, Chile (Transantiago) operates by using an AFC, where users validate when they board buses or enter a metro or bus station. The users are required to wave their cards near a validation device that reads the information and transacts the corresponding charge, allowing the passenger to enter the system. All of the payment transactions are recorded and contain information that is valuable for studying the public transport system. Several systems also have tap-off validation (Park et al., 2008; Ma et al., 2017), which can easily identify the origin and

* Corresponding author.
E-mail address: mamuniza@ing.uchile.cl (M.A. Munizaga).
[1] Currently at Citiplanning.
[2] Currently independent consultant.

http://dx.doi.org/10.1016/j.jtrangeo.2017.10.017
Received 5 May 2017; Received in revised form 12 September 2017; Accepted 24 October 2017
Available online 06 November 2017

destination stops of each trip stage. Others, such as Transantiago, do not require tap-off validation. In those cases, the alighting stop can be estimated by observing the daily sequence of transactions (Trepanier et al., 2007; Munizaga and Palma, 2012). Moreover, methods have been proposed to identify the trip destination, to distinguish transfer stops from activity destinations, and to assign an estimated purpose to each trip (Devillaine et al., 2012; Nassir et al., 2015). These results allow for the collection of valuable information regarding the travel behavior of public transportation users, such as travel patterns and time-use patterns.

From the time use perspective, previous studies have shown that income and other socioeconomic characteristics are relevant variables to explain time assignment. In Santiago it has been observed that users from different residential areas have different characteristics and this explains differences in the time use patterns of users from different residential areas (Jara-Díaz et al., 2013). If we are able to estimate residence zone for a sample of users, and observe the time use revealed by the trips made by them, we should expect to observe similar differences in the time use patterns to those observed by Jara-Díaz et al. (2013) using a sample obtained from the 2001 Santiago OD survey.

In this paper, a method to estimate the residence zones of frequent public transportation users is proposed. It is also applied, and the results are examined to verify if the assumption that justifies the effort does hold. Section 2 contains a brief description of the literature and a description of the context where this study was performed. The proposed method is presented in Section 3 and in Section 4 we present the results. Discussion and conclusions are presented in Section 5.

2. Background

The potential use of smartcard data as a complement of traditional travel survey methods was recognized by Bagchi and White (2005), who highlighted the quantity and temporal nature of this data source, and Utsonomiya et al. (2006), who performed their analysis using the Chicago transit system data. Pelletier et al. (2011) summarised the applications of these data performed until then, describing both the type of system and the uses given to the data. A number of studies have been dedicated to studying travel behavior, Morency et al. (2007) analyzed 10 consecutive months of data and observed weekly transactions to study the effect of holidays, among other things. Utsonomiya et al. (2006) studied the variability in the origin of the first trip of the day, focusing on a subsample of users whose addresses were known, to analyse differences by residence zone. Morency et al. (2007) analyzed 277 consecutive days of data and reached the conclusion that the most common behavior is to travel infrequently. The authors searched for spatial and time patterns using clustering methods. Lee and Hickman (2011) observe travel patterns of frequent public transport users in Minneapolis, US using smartcard data. Ma et al. (2013) also use cluster methods to search for time and space patterns in the massive Beijing, China dataset. Kusakabe and Asakura (2014) propose a data fusion methodology to estimate trip purpose, and use that information to make a frequency analysis over a long period that allows, for example, observing differences during summer time. Langlois et al. (2016) identify activity locations and perform cluster analysis to classify types of users according to the frequency of travel and locations visited. They perform a statistical analysis of the individuals in each cluster and found that they have different socioeconomic characteristics. Ma et al. (2017) use data mining methods to identify commuters, and characterize their travel patterns. As part of the analysis they identify home and work locations. They use a small survey to validate if their model labels the commuters in the sample correctly. The home and work locations are not validated.

Olguín et al. (2009) analyzed travel chains determined from the Santiago 2001 Origin-Destination survey (ODS) to study time-use patterns and determine differences by residence zone, age and gender. The study by Olguín et al. (2009) is interesting, though limited, because a large ODS was conducted only once every ten years and, due to its large cost, only a sample of users were surveyed, typically with one-day surveys. Devillaine et al. (2012) proposed to expand this analysis using massive smartcard data and, after applying a purpose imputation method, time-use profiles were generated. However, the results of this study can only be compared in aggregate because no socioeconomic information is available. Therefore, we cannot explore if the differences by residence zone found by Olguín et al. (2009) are also present in this larger, more reliable and replicable sample.

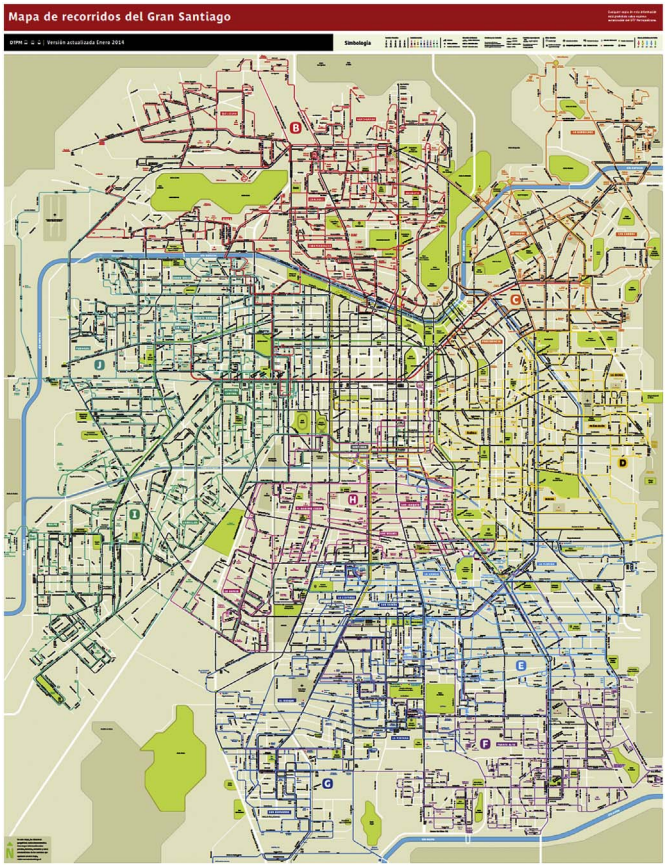

Fig. 1. Transantiago route map.
Source: http://www.transantiago.cl/imagenes/uploads/20160307134044-mapagenerallfebrero2016.pdf.

The public transport system available in Santiago, Chile, named Transantiago, has a wide coverage of the Santiago area and serves 34 of the 37 municipalities that are part of the greater metropolitan area (Fig. 1). The daily number of trips in a working day is 4621 thousands (Muñoz et al., 2015), which represents 27% of the total number of trips, and 41% of motorized trips. The public transport network has over 11,000 bus stops and 108 metro stations. 600 bus routes are operated with over 6500 buses; all equipped with GPS devices (see Gschwender et al., 2016).

Santiago is a spatially segregated city, and important differences in income by residence zone can be observed. Using the six large zones analysis proposed by Jara-Díaz et al. (2013) we built Fig. 2 using data from the 2012 Santiago ODS (Muñoz et al., 2015). Each zone is described in terms of average household characteristics, including household size, income and car ownership. Some mobility indicators are also provided. It can be seen that these large zones are indeed different. The East zone has the highest income, almost three times larger than the average income of the poorer zones, the highest car ownership rate and highest trip rate, but the lowest proportion of trips made by public transport. The Centre zone has the highest public transport market share, and the smallest household size. The rest of the

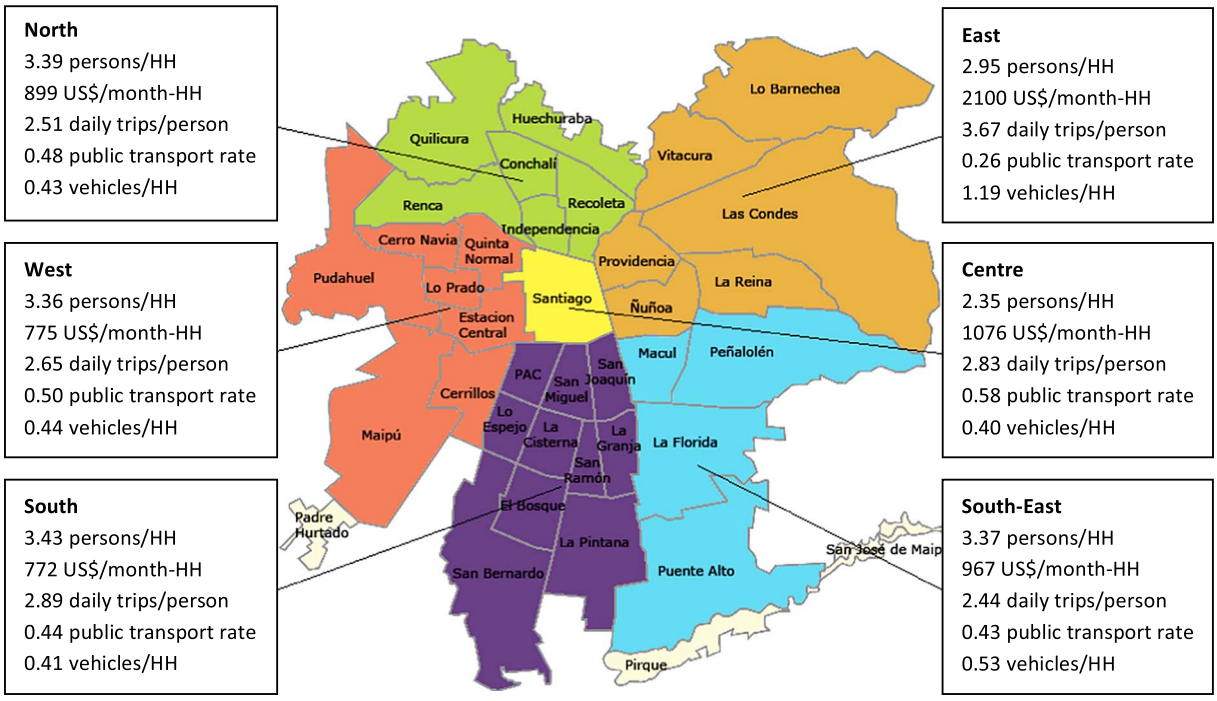

Fig. 2. Description of Santiago in six large zones.

zones are more similar between each other, with some differences in income and car ownership. Based on this information, we propose to use this segmentation to explore differences in travel and time-use patterns of public transport users.

2.1. Alighting estimation method

To be able to build public transport OD matrices from the boarding transactions information, Munizaga and Palma (2012) proposed a method to estimate alighting stops based on the observation of transaction sequences, assuming that all trips are made by public transport using bus, Metro or a combination of both modes. This method identifies the most convenient stop to reach the next boarding location as the one that minimizes a generalized time function, which considers travel time and penalizes walking time, and is valid only when all trips and trip stages are made by public transport on that particular day. Once an alighting stop is identified, the time of alighting can easily be obtained as well. In the case of bus transactions, the time of alighting can be obtained from detailed AVL data by making an interpolation between GPS points. In the case of transactions in Metro, a route within the Metro network must be assumed (minimum time), and then the alighting time can be estimated using information from the operational program.

2.2. Trip identification and trip purpose estimation method

In addition, Devillaine et al. (2012) proposed a method to identify trips, linking trip stages, and to estimate the purpose of those trips. The trip identification is done using the Activity Identification Module (AIM) that considers the sequence of transactions, observing the time lapse between the estimated alighting time of one particular transaction (trip stage) and the boarding time of the next one. Using this procedure, trip stages are linked into trips associated to an activity. Therefore, travel time and transfer time can be estimated. Then, the Purpose Assignment Module (PAM) is used to identify the purpose of the activities, using information about activity duration and sequence along the day. The purposes that can be identified with the available information are Work, Study, Home and Other. Work and Study are the primary activities of the observed cards, with durations of at least two hours during a working day and at least five hours during the weekend. The activity identified as Home corresponds to the laps between the last transaction of the day and the subsequent transaction of the next day. The purpose Other is assigned to any activity that does not comply with the previously mentioned conditions, i.e., activities shorter than two hours during working days and shorter than five hours during the weekend. Additionally, the activity “Travel” can be observed directly from the trip database, as the duration of trips is known.

2.3. Validation

Munizaga et al. (2014) present a validation of these OD matrices obtained from smartcard and GPS data, and trip purpose estimation. They verified the assumptions using exogenous information from surveys, but also using the same database used to make the estimation. The results were very positive, showing that alighting bus stop is correctly estimated in 84.2% of the cases, trip/trip stage identification is correct in 90% of the cases, and the purpose of the trip is correctly estimated in 79% of the cases. They also propose some minor modifications to the original methodology that will contribute to improve those percentages in future applications. In general terms, the conclusion is that this is a reliable database, which represents adequately the travel behavior of the majority of public transport users.

3. Proposed method to estimate the residence zone

We now want to move one step further and observe the time-spatial pattern of frequent users, defined as those who use the public transport system frequently. This definition will include commuters traveling to work and study activities, and users who travel for other purposes. When looking at the aggregate numbers, very reasonable behavioral patterns can be observed: even though there are some cases of users who behave peculiarly, they are a minority. The first assumption we will make is that frequent public transport users use the system for the majority of their trips; therefore, the first transaction in the morning is likely to be near their residence zone. We will leave park and ride, kiss and ride and other combinations with public transport out of the

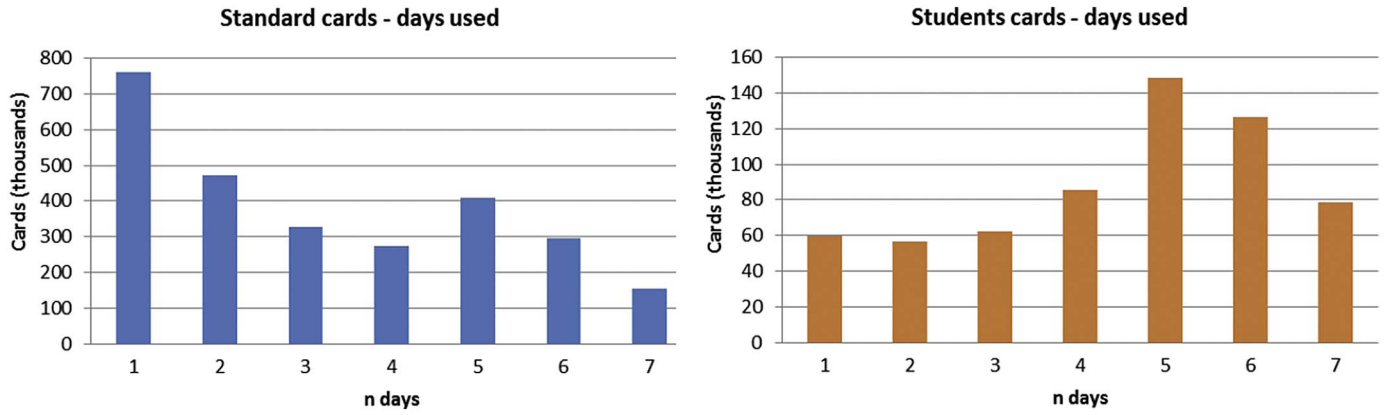

Fig. 3. Card usage histogram for standard and student card users.

analysis (according to the Santiago ODS, they are < 3%). Therefore, we will assume that if all the first-in-the-morning transactions of a particular user occur within a walking distance radius, then the residence of that user is located in that area.

To identify frequent users, we can use the information about transactions observed per user in a week. Fig. 3 shows the histogram of days a card is observed (presents transactions) in a week for standard and student cards. These are the only two different types of card available. Standard card users pay full fare, student card users have discounted fares. Student cards are handed out by the Ministry of Education to primary school students, who have 100% discount, and also to secondary, college and university students, who have 70% discount. As in previous studies, we observe that infrequent users are the most common of the full fare users (one traveling day observed in a particular week). However, the opposite can be observed for students: five traveling days observed in a particular week is the most common behavior. In terms of what can be considered as frequent use of the system, Lee and Hickman (2011) state that cardholders who make two or more transactions per day for 5 consecutive weekdays can be defined as regular users. However, different criteria can be used, depending on the objective of the work. If we consider frequent travelers as those who travel at least five days per week, we obtain approximately 1.2 million observations per week, which represents 33% of standard card users and 57% of student card users. These figures rise to 1.6 million observations (43 and 70% respectively) if four days are considered, and 2 million observations (55 and 88% respectively) in the case of three days required. On the contrary, if 6 days are required, then the number of observations falls below 700,000 (17% of standard cards, 33% of student cards).

For the subsample of frequent users, we observe the spatial distribution of their first morning transactions for the observed days. All transactions occurring between 4 AM and noon are considered to be morning transactions, whereas the first transaction on each particular day is specifically observed. Then, the center of gravity of the coordinates of those first morning transactions is taken as a reference. The distance from the position of the first morning transactions performed on all days (x_i; y_i) to the center of gravity is calculated. If the largest distance d is lower than the walking distance D, then the card is associated with the residence zone corresponding to the zone where those transactions were made (or the majority of them). We illustrate the proposed method in Fig. 4.

Fig. 5 presents an example of application of the residence zone estimation method. In this case, the user walks to different bus stops near their house three days during the observed week. We can observe the position of the morning transactions (yellow pins), but we do not know their home location. The method assumes that all first morning transactions are located within walking distance of the residence of the cardholder. Cards with observed radii greater than a pre-defined threshold are dismissed for this analysis, which can occur if the user makes the first stage of his/her trip by another mode of transport (taxi, park and ride, kiss and ride) or if the user spent the night at a location other than at their house.

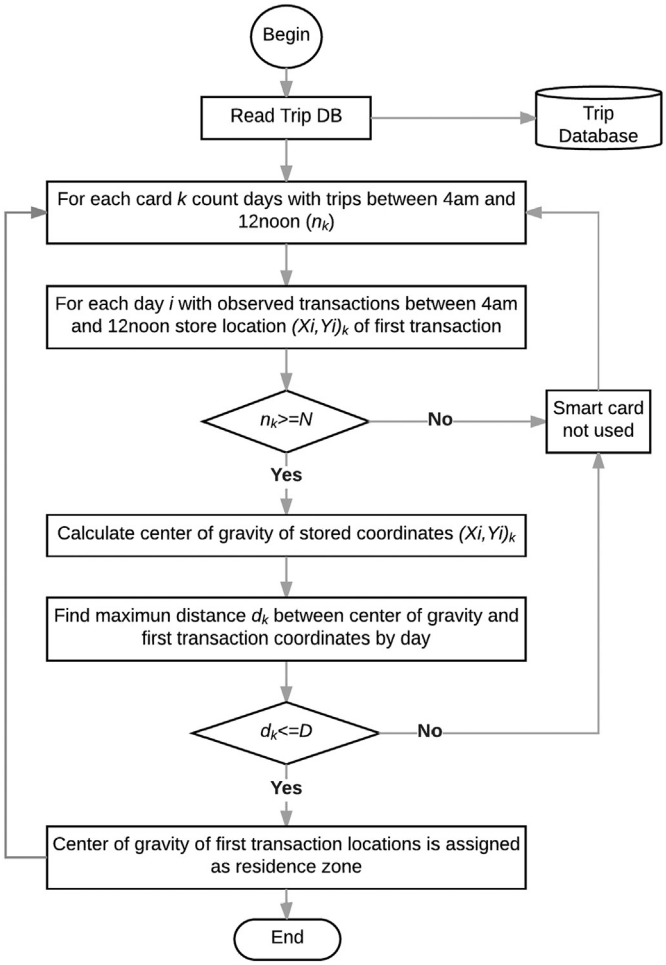

Fig. 4. Residence zone estimation method.

The time window defined to identify the first morning transactions (4 AM-noon) is based on a statistical analysis of the data. The time of day with less activity in public transport usage was identified to be between 3:15–4 AM, and therefore, the natural end of the cycle limit. If we used an earlier time (i.e., midnight), there would be several trip stages observed that correspond to the final trip of the previous day rather than the first trip of the analyzed day. If we used a later time, we would miss several early transactions of users who travel long distances to reach their workplace.

Additionally, there are many cards that register a first trip in the afternoon or evening, which could actually be the first trip of the day;

Fig. 5. Application example of the residence zone estimation method.

however, it is also possible (and likely) that the first trip was performed by a different mode of transport (by taxi or car or on foot) and was, therefore, not observed. For this reason, we excluded cards that only register afternoon or evening trips from the analysis.

4. Application

4.1. Zone of residence estimation

The described method was applied to a database obtained from the week of 14–21 April 2013. From a total of 3,288,464 cards observed that week, 1,294,049 (39%) had a minimum of three days with morning transactions observed in the time window. From these, 68% (888,970) have radii lower than 1000 m, circumscribing all first-in-the-morning transactions. The complete histogram is shown in Fig. 6. It can be observed that most observations are accumulated in the very short distances, showing that a very consistent spatial pattern can be found in a large proportion of frequent users.

A preliminary validation exercise was conducted with a recruited sample of 55 volunteers, most of them students, where 8 of them made at least four morning transactions located within a 500 m radius circle, and in all those cases the estimated zone of residence was correct.

Then, we use ground truth from the Santiago 2012 ODS to make a sensitivity analysis of the zone of residence estimation made with the April 2013 smartcard database. The Santiago ODS has 2422 surveys with observed residence location and declared card id, out of which 358 have zone of residence estimation from the smartcard database. Therefore, those observations can be used for validation of the estimation. Table 1 shows the number of cards where the zone of residence can be estimated for different values of the number of days the card is observed N and the distance threshold D, and also the number of cases where that estimation can be validated with the ODS, and the percentage where the validation shows a correct estimation. It can be seen that the larger number of estimations is obtained for $N = 3$ and $D = 1000$, and the smaller error is obtained for $D = 1000$ and $N = 6$. However, the larger number of correct estimates is obtained with $N = 3$ and $D = 1000$.

Fig. 7 shows the distribution of the distance between estimated zone of residence and location declared in the ODS for $N = 3$ and $D = 1000$. It can be seen that in most cases the zone of residence estimation is quite accurate, with a high concentration of values below 1000 m. The other accumulation point is over 5000 m, which is clearly a matter of error rather than an accuracy problem.

To illustrate this issue, we draw the spatial distribution of these observations with wrong zone of residence estimation. An example of this analysis is shown in Fig. 8, where for each observation (number) the blue pin shows the estimated zone of residence, and the yellow pin shows the declared zone of residence.

Using this type of visualization, and our knowledge of the city, we

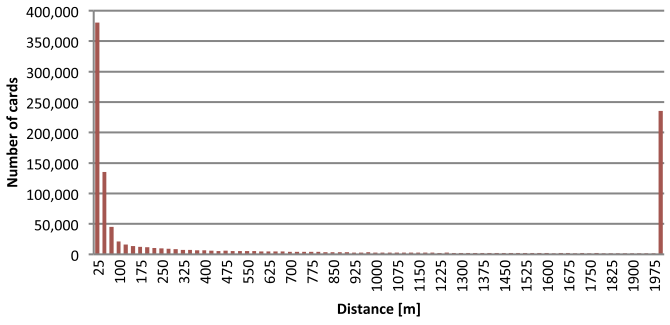

Fig. 6. 3-day users radii histogram.

Table 1
Zone of residence estimation and validation.

N	D = 500 m			D = 1000 m		
	# Estimates	# ODS	Correct	# Estimates	# ODS	Correct
≥ 3	782,756 *23.8%*	226	167 *73.9%*	888,970 *27%*	256	188 *73.4%*
≥ 4	605,874 *18.4%*	156	117 *75.0%*	689,528 *21.0%*	183	136 *74.3%*
≥ 5	417,314 *12.7%*	106	80 *75.5%*	475,471 *14.5%*	122	91 *74.6%*
≥ 6	113,055 *3.4%*	31	25 *80.6%*	133,648 *4.1%*	38	31 *81.6%*

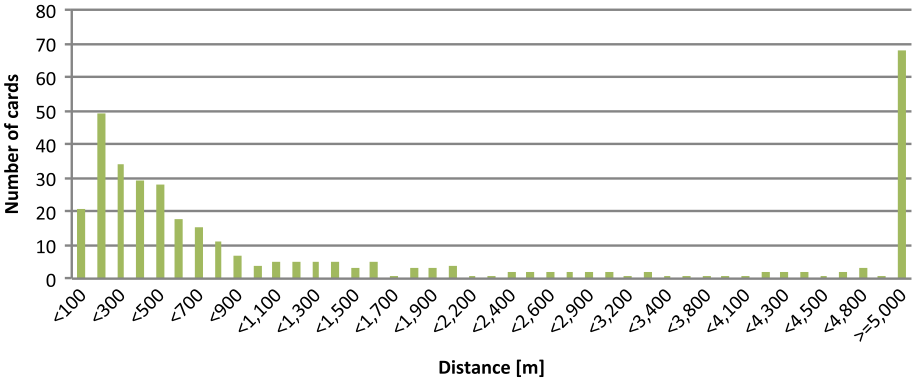

Fig. 7. Histogram of distance between estimated zone of residence and declared zone of residence.

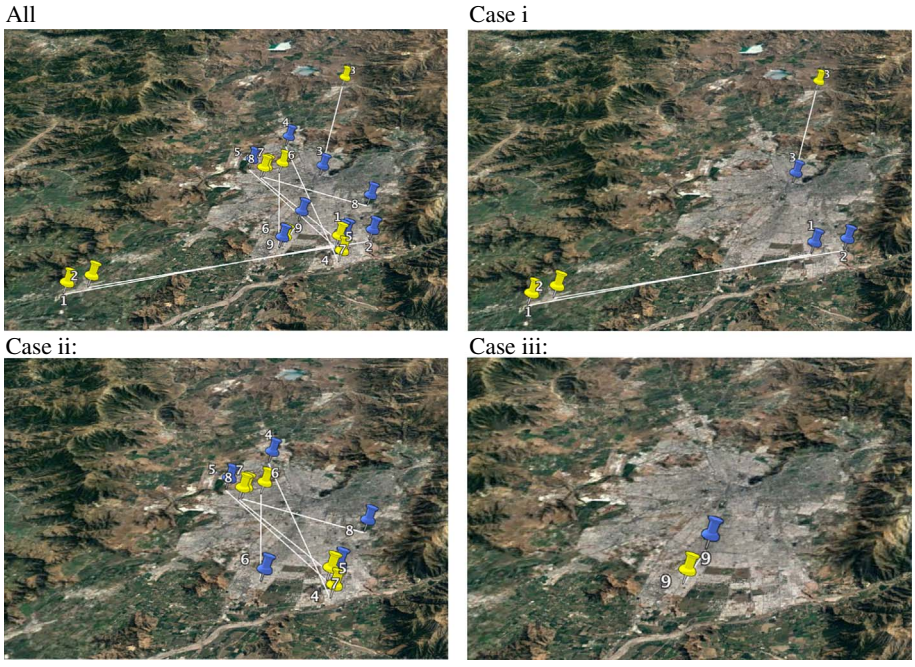

Fig. 8. Visualization of estimated zone of residence and declared zone of residence for observations with wrong estimation (difference over 2000 m).

Table 2
Number of observations available (trips).

Residence zone	Working day		Saturday		Sunday	
	Standard card	Student card	Standard card	Student card	Standard card	Student card
North	46,972	17,645	22,671	7243	8301	2757
West	107,457	39,131	45,551	15,010	16,140	5914
East	34,073	16,307	12,854	5518	5963	2227
Center	19,406	7166	8596	3009	3809	1553
South	73,081	25,419	31,988	10,358	10,403	3877
South-East	84,929	30,317	33,197	11,268	11,427	4467

found three possible sources of error:

Case i) Suburban commuters: Cases where the real zone of residence is located outside the area where public transport is available, and users make a trip stage in a non integrated mode such as suburban bus before their first trip stage in Transantiago (observations 1, 2 and 3 in Fig. 8)

Case ii) Night shift workers: Cases where the real zone of residence is located in the opposite side of the city, usually a residential area, and estimated zone of residence is located in a commercial/industrial area or near a hospital (observations 4–8 in Fig. 8).

Case iii) Park and Ride & Kiss and Ride: Cases where the real zone of residence is located within the area where public transport is available, but users make a trip stage in a non integrated mode, such as car driver or car companion, before their first trip stage in Transantiago (observation 9 in Fig. 8).

In the following sub-sections, we present the travel times and time profiles for the estimated residence zones. However, before the time-use profiles and comparisons, we show the number of observations available per category, which is presented in Table 2. It can be observed that for all zones and days, the number of observations is sufficiently large to confidently calculate average values.

Table 3
In-vehicle travel time and transfer time.

Variable	Residence zone	Week day		Saturday		Sunday	
		Standard card	Student card	Standard card	Student card	Standard card	Student card
In-vehicle travel time [min]	North	36.1	33.2	30.6	29.5	29.8	28.8
	West	38.7	34.3	34.4	31.3	32.0	30.5
	East	29.6	28.0	28.6	24.9	28.8	24.8
	Center	28.8	24.7	26.4	24.4	25.2	24.4
	South	39.9	34.3	35.5	31.5	32.6	29.9
	South-East	37.9	34.4	34.3	31.1	31.6	29.2
Transfer time [min]	North	9.4	8.9	9.7	9.4	10.1	10.2
	West	9.6	9.1	9.7	9.8	10.5	10.1
	East	9.1	8.4	9.5	9.3	10.7	9.4
	Center	9.3	8.7	9.8	10.1	10.2	10.4
	South	9.9	9.4	10.3	10.1	10.4	10.0
	South-East	10.1	9.8	10.4	10.4	10.6	10.2
Total = Travel + transfer time [min]	North	45.5	42.1	40.3	38.9	39.9	39.0
	West	48.3	43.4	44.1	41.1	42.5	40.5
	East	38.7	36.4	38.2	34.1	39.5	34.2
	Center	38.2	33.4	36.2	34.5	35.4	34.8
	South	49.9	43.7	45.8	41.5	43.0	40.0
	South-East	48.0	44.2	44.7	41.4	42.2	39.5

4.2. Statistical description by residence zone

4.2.1. In-vehicle travel time and transfer time

Table 3 shows the average in-vehicle travel time and transfer time for trips made by users with standard and student cards, by day and by residence zone. Differences > 10 min can be observed in the average total values for standard card users on weekdays. The zones with lower travel times were the Center and East zones, with average values under 30. Larger values were observed for the South, West and Southeast zones, reaching figures near 40 min. During the weekend, travel times are smaller, particularly on Sundays. For student cards, the differences between zones are smaller, and the average travel time is also smaller for all zones. These differences can be explained because students have different time patterns and also because of the trip distribution. In Transfer times, the differences between standard card and student card users are smaller, but still student cards tend to have lower values. For standard card users on weekdays, the average transfer time ranged from 9.1 to 10.1 min, depending on the residence zone. On Sundays, the average transfer time was larger, reaching values > 10 min on average in all zones. Adding both effects, the difference in the average total values is 11.2 min between residents of the Center and South zones on working days for standard card users. These results show that there are clear differences between residence zones on average travel time, which justifies this analysis.

From the data set, standard cardholders who travel to/from work Monday to Friday (two trips per day) spend an average of 7.84 h more per month if they live in the South zone of Santiago compared to those that live in the Central zone. This difference reaches to 11.76 h if they make three trips per day. For student cardholders, the difference between the Central and Southeast residents is 7.56 h if they make two trips per day and 11.34 h if they make three trips per day. In the next section, we use the estimated trip purpose, and the time use structure that can be deduced form the information available to make an analysis of how these users distribute their time between different activities.

4.2.2. Time-use profiles

Fig. 9 shows the time-use profile for an average Monday-Thursday working day and for a Friday. This Figure is built using information of cards that have purpose estimation in all the trips observed in a particular day. Before the first trip of the day, we assume that the user (card) is at home. The blue line represents the proportion of users that according to the data are not traveling or performing out of home activities at each time along the day. They are associated to “Home” time use, representing being at home or conducting activities near home that are reached by walking or by other transport modes (different from public transport). The light blue line represents the proportion of users who are traveling (by public transport) per time of day. This information is obtained directly from the database, and associated to “Travel” time use. Between trips, using the methods described in Section 2 we distinguish long activities (work for regular card, study for student cards) from short activities (other). Red, green and purple lines represent the proportions of users at each of these activities per time of day respectively. The aggregate pattern appears reasonable, with over 40% of users at work and approximately 20% at study during working

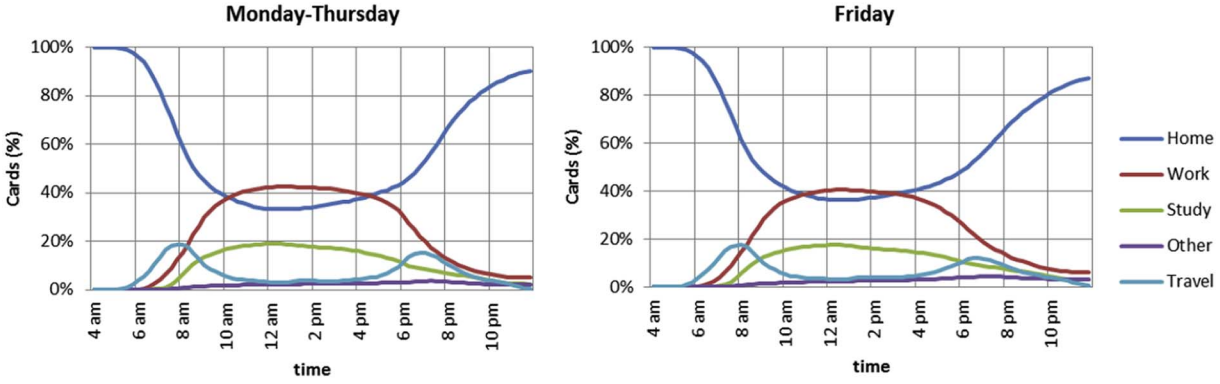

Fig. 9. Time-use profiles.

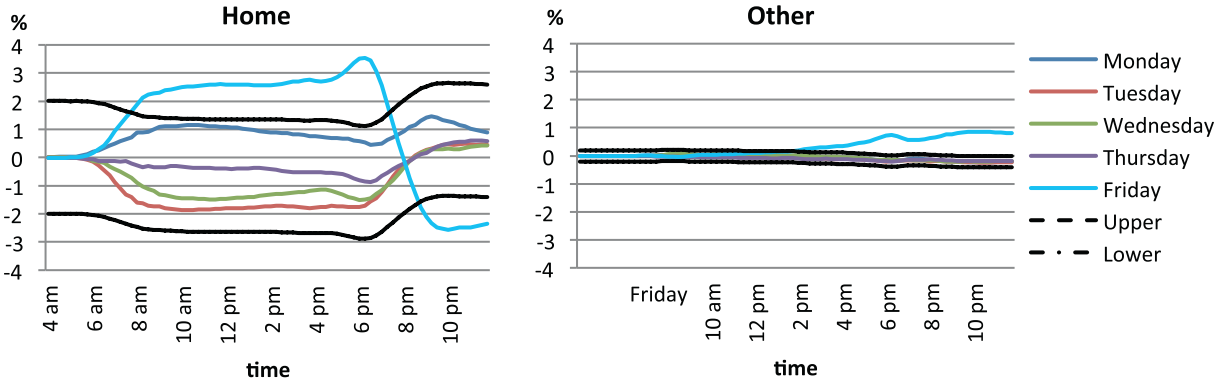

Fig. 10. Comparison of time use by the day of the week.

hours, whereas most people were home during the night, with important travel activity occurring during the morning and evening peak hours and only a small proportion of short (other) activities. Friday is slightly different from the rest of the days.

To analyse the time use structure by zone, we plotted the differences found on each curve, instead of repeating the figure for all zones. Fig. 10 shows the differences between the days of the week for the activities "Home" and "Other". This graph is normalized by the number of observations, calculating the global average first and then calculating the difference of each category with respect to the global average. The results of this calculation are plotted by color to distinguish each day, with a 3% and 0.1% difference with the mean bands indicated by black lines. The differences between time use on Friday and the rest of the week can be observed more clearly in Fig. 10 than in Fig. 9. For Home activity, a similar behavior was observed to occur in all days except Friday; on Friday, a larger portion of users/cards is at home during the day, and out of home at night. Also, we can observe a larger proportion of other activities on Friday afternoon. This difference of behavior by day of the week was not observed by Olguín et al. (2009), but this is reasonable, because the results of Olguín et al. (2009) are based on data from the Santiago 2001 OD survey, made before the last important change on the working hours made by the Chilean regulation (from 48 to 45 h per week). Olguín et al. (2009) found that there was no significant difference between Friday and the rest of the working days in a week, but they recognize that that's probably not true anymore, given the change in the working hours established in 2005 (Código del Trabajo, Chile).

Fig. 11 shows the comparison between zones for Home, Work and Travel activities. It can be observed that the differences are concentrated in the morning and the afternoon. During working hours (9 AM – 6 PM), all curves are similar. Additionally, the East and Center zones are more different from the other zones. A similar result was obtained by Olguín et al. (2009), who compared time-use patterns of different types of users from a small sample that was obtained from a detailed travel survey. Fig. 11 shows that cardholders who live in the East and Central zones tend to stay at home until later in the morning and tend to arrive earlier in the afternoon, which is a result of two different effects: cardholders who live in the East and Center zones arrive at work later and also spend less time traveling during the morning and evening peak hours. The primary difference between Friday and the other working days is that more dispersion is observed on Fridays, with the differentiation between zones beginning earlier.

4.3. Scope and limitations

The proposed method can be applied to data from public transport AFC systems. In systems with tap-in validation only, the estimation of the alighting bus stop and trip identification are required previous steps. The residence zone estimation and time use analysis will be effective for individuals who use public transport for the majority of their trips. Differences by sociodemographic characteristics can be observed in segregated cities. These differences will be captured among individuals from sociodemographic groups that use the public transport system (bus and metro services). We expect an income bias in the sample obtained, as wealthier people do not use public transport as much as poorer people do (this is apparent from Fig. 2). A selectivity bias may also be present in the data obtained for the higher income segments, as higher income travelers will consider more expensive alternatives such as private car or taxi, while lower income users might be captive to public transport. Therefore, higher income users are more likely to choose public transport only when the level of service obtained is sufficiently good compared to the alternative modes (private car and taxi).

5. Conclusions

This paper explores the possibilities for spatial analysis of extensive smartcard databases. Compared to traditional studies based on survey data, smartcard provides significantly bigger sample sizes, and more precision in time and location variables, but no socioeconomic variables. This paper addresses that limitation, which constrains the types of analysis that can be made due to the lack of socioeconomic variables. The socioeconomic information of cardholders is available to researchers only in a few exceptional cases. This is a limitation because income, gender and other socioeconomic variables have an important effect on travel behavior.

The method proposed here allows for a residence zone to be estimated for cardholders who are frequent public transport users, i.e. use the system several days in a week. This is a first step towards socioeconomic characterization because, in a segregated city, such as Santiago, the zone of residence is highly correlated with income and other socioeconomic variables. Therefore, if we can correctly identify the residence zone, we can make assumptions about some socioeconomic characteristics such as income. This approach is only valid in segregated cities. In cities with a highly heterogeneous urban mixture, the assumptions made here would not be reasonable.

The method is validated with exogenous data, showing that over 70% correct estimates can be obtained. Incorrect estimates may be explained by suburban commuters, night shift workers and combined trips (park and ride, kiss and ride). The accuracy of the estimation can be improved increasing the number of days the card is observed.

The application of the method to Santiago showed significant differences in values of the average travel time for travelers from the Center and East zones. Also, the East and Center zones show significantly different time use patterns, compared to the rest of the zones. This type of analysis has been conducted in the past, using time use surveys or detailed origin-destination surveys, which are very expensive and difficult to obtain. The methods presented here can be applied to

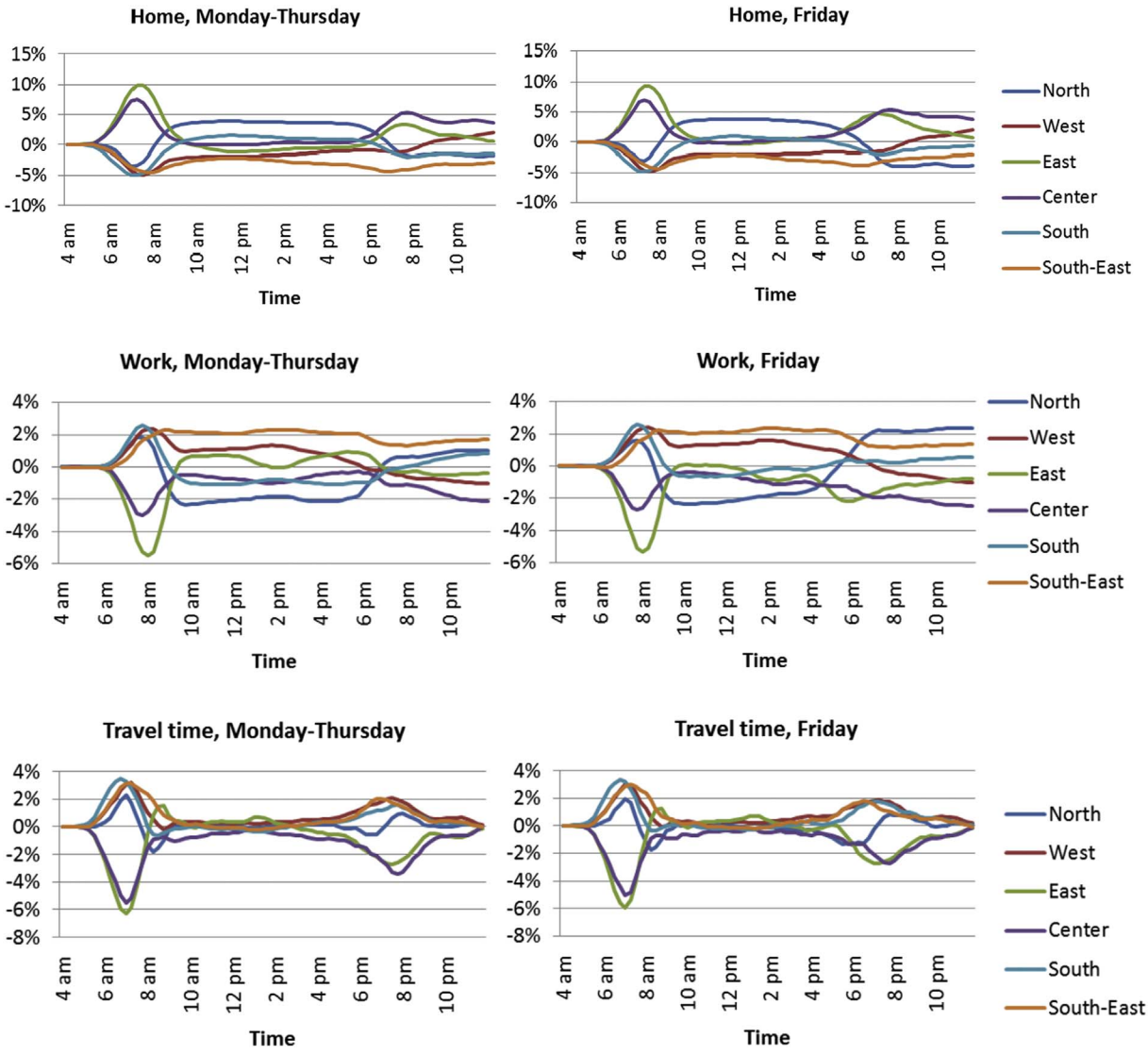

Fig. 11. Time-use comparison by residence zone.

any specific week, allowing for time use variation analyses. Additionally, they can be applied to massive amounts of data, allowing any time-space disaggregation required.

In the long term, this research can be useful to analyse the effects of public policies, making before/after analysis. For example, we will be able to study the effect of important infrastructure investments such as a Metro line that is currently under construction in Santiago de Chile, and evaluate its effects on travel time and the time use profiles of users from different neighborhoods/income segments.

Acknowledgements

Funding: Fondecyt1161589, FondefD10E-1002, Complex Engineering Systems Institute (CONICYT – PIA – FB0816; ICMP-05-004-F). We would also like to thank the collaboration of the Transit Authority DTPM and Joaquín Romero for his help with the English editing.

References

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12, 464–474.
Devillaine, F., Munizaga, M.A., Trepanier, M., 2012. Detection of activities of public transport users by analyzing smart card data. In: Transportation Research Record: Journal of the Transportation Research Board, No. 2276. Transportation Research Board of the National Academies, Washington, D.C., pp. 48–55.
Gschwender, A., Munizaga, M.A., Simonetti, C., 2016. Using smartcard and GPS data for policy and planning: the case of Transantiago. Res. Transp. Econ. 59, 242–249.
Jara-Díaz, S., Rosales-Salas, J., 2015. Understanding time use: daily or weekly data? Transp. Res. A Policy Pract. 76, 38–57.
Jara-Díaz, S.R., Munizaga, M.A., Olguín, J., 2013. The role of gender, age and location in the values of work behind time use patterns in Santiago, Chile. Pap. Reg. Sci. 92 (1), 87–103.
Kusakabe, T., Asakura, Y., 2014. Behavioural data mining of transit smart card data: a data fusion approach. Transp. Res. C Emerg. Technol. 46, 179–191.
Laglois, G., Koutsopoulos, H.N., Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transp. Res. C Emerg. Technol. 64, 1–16.
Lee, S.G., Hickman, M., 2011. Travel pattern analysis using smart card data of regular users. In: Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board of the National Academies, Washington, D.C..
Ma, X., Wu, Y., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. C Emerg. Technol. 36, 1–12.
Ma, X., Liu, C., Wen, H., Wang, Y., Wu, Y.-J., 2017. Understanding commuting patterns using transit smart card data. J. Transp. Geogr. 58, 135–145.
Morency, C., Trépanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transp. Policy 14 (3), 193–203.
Munizaga, M., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. C Emerg. Technol. 24, 9–18.
Munizaga, M.A., Devillaine, F., Navarrete, C., Silva, D., 2014. Validating travel behavior estimated from smartcard data. Transp. Res. C Emerg. Technol. 44, 70–79.
Muñoz, V., Thomas, A., Navarrete, C., Contreras, R., 2015. Encuesta origen-destino de Santiago 2012: Resultados y validaciones. Ing. Transp. 19 (1), 21–36.
Nassir, N., Hickman, M., Ma, Z.L., 2015. Activity detection and transfer identification for public transit fare card data. Transportation 42 (4), 683–705.

Olguín, J., Jara-Díaz, S.R., Munizaga, M.A., 2009. Análisis de patrones de actividades a partir de la EOD 2001. Ing. Transp. 13 (4), 31–38.
Park, J., Kim, D., Lim, Y., 2008. Use of smart card data to define public transit use in Seoul, South Korea. In: Transportation Research Record: Journal of the Transportation Research Board, No. 2063. Transportation Research Board of the National Academies, Washington DC, pp. 3–9.
Pelletier, M., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. C Emerg. Technol. 19 (4), 557–568.
Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. J. Intell. Transp. 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. In: Transportation Research Record: Journal of the Transportation Research Board, No. 1971. Transportation Research Board of the National Academies, Washington DC, pp. 119–126.