Transportation Research Part A 141 (2020) 307–322

Contents lists available at ScienceDirect

Transportation Research Part A

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

Fare evasion correction for smartcard-based origin-destination matrices

Marcela A. Munizaga[a,b,*], Antonio Gschwender[a,b,c], Nestor Gallegos[a,1]

[a] Departamento de Ingeniería Civil, Universidad de Chile, Chile
[b] Instituto Sistemas Complejos de Ingeniería (ISCI), Chile
[c] Directorio de Transporte Público Metropolitano DTPM, Chile

ARTICLE INFO

Keywords:
Fare evasion
OD matrices
Public transport

ABSTRACT

Origin-Destination matrices obtained from smartcard data are very valuable because they contain vast amounts of information and can be obtained at a very low cost. However, they can only account for trips paid by smartcard. Trips paid by other means, as well as non-paid trips, must be incorporated using additional information. This paper discusses the biases that are introduced due to fare evasion and presents a sequential method to estimate correction factors due to partial evasion in some trip stages, as well as total fare evasion in all trip stages, using external information regarding trips not registered in the smartcard database. We apply this method to the case of Santiago, Chile, where partial evasion (during a bus trip stage prior to a Metro trip stage) and total evasion (during all bus-only trip stages) are relevant fare evasion situations. Information from the Santiago Metro Origin-Destination survey and from external fare evasion measurements is used. The results indicate a 5% partial evasion rate for bus trip stages prior to Metro trip stages, and a 37% total fare evasion rate for bus-only trips. This paper is a contribution towards establishing new methods to feasibly obtain OD matrices through the adequate merging of automatically-collected data with complementary traditional measurements and survey instruments.

1. Introduction

Automatically-collected passive data have emerged as a very important source of useful low-cost information, but the methodologies developed for this purpose must address inherent data limitations. These passive data limitations include partial trips registered (in the case of coexistence of smartcard and paper tickets), the absence of socioeconomic information (when smartcards are not personalized, or when personal information is confidential), and difficulties when deducing the transportation mode used (in the case of smartphone data). New procedures must be developed to address these issues and extract “the best of both worlds”, i.e., cheap and massive databases alongside specific (and expensive) traditional measurements and surveys to provide complementary missing information. Within this context, in this paper we present a method to improve the estimation of public transport OD matrices from passive data, using complimentary measurements and very specific surveys to take fare evasion into account.

Fare evasion is a common problem in public transport systems. Defined as not paying for a ride or paying too little, fare evasion can

* Corresponding author at: Departamento de Ingeniería Civil, Universidad de Chile, Chile.
E-mail address: mamuniza@ing.uchile.cl (M.A. Munizaga).
[1] Currently at Servicio de Impuestos Internos.

https://doi.org/10.1016/j.tra.2020.09.008
Received 27 October 2017; Received in revised form 4 September 2020; Accepted 14 September 2020
Available online 8 October 2020
0965-8564/

Table 1
Average monthly household income per municipality.

Municipality	Average household income [USD/month]	Municipality	Average household income [USD/month]
Vitacura	6111	Lo Espejo	1498
Las Condes	5150	Estación Central	1497
La Reina	4448	Renca	1483
Providencia	4437	El Bosque	1454
Ñuñoa	4414	Lo Prado	1411
Lo Barnechea	3713	Puente Alto	1397
Santiago	2693	San Ramón	1379
San Miguel	2337	Pedro Aguirre Cerda	1339
Macul	2195	Pudahuel	1313
Quilicura	2145	San Bernardo	1298
La Cisterna	2103	Conchalí	1279
La Florida	1949	La Pintana	1262
Maipú	1938	La Granja	1206
San Joaquín	1833	Independencia	1201
Cerrillos	1663	Cerro Navia	1195
Quinta Normal	1603	Recoleta	1133

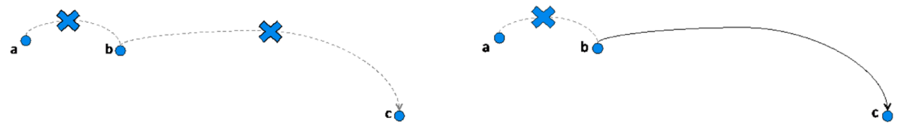
Source: CASEN (2013).

be an attractive form of fraud for passengers (Smith and Clarke, 2000). Some of the factors that contribute to crime in public transport in general and to fare evasion in particular are overcrowding, lack of supervision, and anonymity. Due to its impact on the revenue stream and the overall financial sustainability of public transport systems, fare evasion has been a concern for many years (see, for example, Boyd et al. 1989; Dauby and Kovacs, 2007; Lee, 2011). Aside from the impacts on system finances and moral questions regarding users and operators, there is a consequence of missing data. The payment system is one of the most important sources of demand information. In systems with automatic fare collection (AFC), paid trips are recorded in databases that are then used to perform different types of demand analysis. Non-paid trips are not included in these databases.

Different cities have undertaken analyses on fare evasion, including Melbourne (Delbosc and Currie, 2016), New York (Reddy et al. 2011), and Cagliari (Barabino et al., 2015). The focus of these studies is usually on understanding fare evasion behaviour to enlighten fare evasion reduction policies. Some authors have developed sophisticated methods to observe and understand the behaviour of fare evaders. Dai et al. (2017) conduct a field experiment to explore the relation between fare evasion and attitudes towards dishonesty, finding a correlation between dishonesty observed in lab experiments and fare evasion. Delbosc and Currie (2016) make a distinction between accidental, unintentional, and deliberate evaders, depending on whether they intended to pay but could not, or whether they decided not to pay. Using a cluster analysis method, they identify three groups: deliberate evaders, unintentional evaders, and never evaders. The behaviour of these three groups is very different; the authors also note some socioeconomic differences. These findings have relevant policy implications, as different enforcement strategies should be used for each of these groups. Delbosc and Currie (2019) perform a deep review of literature on fare evasion, classifying the literature in three broad groups based on the perspective of the work: a conventional transit system perspective, a profiling perspective (examining ethical issues), and a customer motivation perspective.

As part of the conventional transit system perspective, research has been conducted on fare evasion control and fare inspection optimization. For example, Yin et al. (2012) propose a method for scheduling patrols for fare inspection considering the potential massive scale of fare evasion behaviour. Barabino et al. (2014) propose an economic framework to establish the optimum number of inspectors, and providing a case study from Italy. Correa et al. (2017) propose a bilevel programming model, where the network operator determines probabilities for inspecting passengers at different locations, and the fare-evading passengers respond by optimizing their routes given the inspection probabilities and travel times.

The public transport system in Santiago, Chile, known as Transantiago, has a dramatic fare evasion problem, reaching levels close to 30% in buses (DTPM, 2016). This behaviour is so concerning to the authorities that the Chilean Ministry of Transport has made combatting fare evasion a priority. Factor Estratégico (2010) analyses the fare evasion behaviour of Transantiago users from the customer motivation perspective and identifies different types of fare evaders, similar to those identified by Delbosc and Currie (2016). The intentional (or deliberate) evader is a traveller who has decided to use the system without paying for it. This user will try to evade payment during all trip stages, or will prefer to make one-stage trips, not paying for that stage. This behaviour is observed mainly in bus users. The circumstantial (or unintentional) evader is a user who regularly pays for his or her trips, but is occasionally not able to. According to this study, the main reasons for circumstantial evasion are the lack of charging points in residential areas, which is a shortcoming of the system, and carelessness of passengers who fail to charge their cards prior to travelling. This circumstantial evasion may be total (in the case of one-stage trips), but will, in many cases, be partial. A case observed with some frequency during field observations is that of users who board a feeder bus with insufficient funds on their card. They usually tell the driver that they do not have enough funds and that they will recharge at the metro station - and they usually do so. Delbosc and Currie (2019) provide figures showing that over 80% of fare evaders in Santiago did not pay because they could not find a fare loading point prior to beginning their first trip stage. This observation is consistent with the results of surveys such as the Metro ODS (Ipsos, 2013) and the Santiago ODS (Muñoz et al., 2016), which show a lower proportion of Metro-only trips and a higher proportion of bus-Metro trips than what is observed in the matrices obtained from the fare payment system. As every Metro station has fare loading points, this type of evasion



Complete trip fare evasion, all trip stages   Partial fare evasion, one trip stage only

Fig. 1. Types of fare evasion.

will be associated with either first-stage or total trip bus fare evasion.

Models have been developed to identify factors that explain the high levels of fare evasion at the trip stage level in Transantiago. Guarda et al. (2016) found that boarding conditions (crowding, boarding through the back door), high bus occupancy rates, and long headways are variables that contribute to fare evasion; they also found that fare evasion is greater during the afternoon and the evening, as well as in lower income areas. Troncoso and De Grange (2017) undertook a more aggregate analysis; using time series data over eight years, they found a positive effect of fare raises, a negative effect of the number of inspections (dissuasive pressure), and a negative effect of unemployment rates.

The Santiago metropolitan area is administratively divided into 34 municipalities. It is considered a spatially-segregated city with an uneven income distribution (Amaya et al. 2018). Table 1 shows average income per municipality. The average income in the richest municipality (Vitacura) is five times the average income in the poorest municipality (Recoleta).

At the time that this research was completed, Transantiago was a bus-Metro integrated system with a trip-based fare structure and smartcard payment system. One fare payment covers trips of up to three trip stages within a two-hour time window. Users are requested to validate their smartcard each time they initiate a trip stage, and the system deducts the corresponding amount depending on the time of day and the mode used to complete the trip stage (bus or Metro). Passengers take 4.5 million daily trips in this public transport system. The distribution of passengers between modes is 22% Metro, 52% bus, and 25% bus-Metro combinations (Muñoz et al., 2016).

Relevant AFC information, used in combination with GPS data obtained from devices installed on buses, is used for OD matrix estimation. AFC data usually contains the card ID, validation timestamp, and validator ID (which, in the case of bus rides, is associated with the vehicle). GPS data contains a sequence of location-time points for the vehicle. Both databases can be linked using the timestamp and vehicle ID to estimate the position of the boarding transaction. Methods have been proposed for estimating trips and building OD matrices using this information. They are based on a destination inference algorithm (required in the case of tap-in validation only), and a method to link trip stages and identify OD movements. Simpler cases are based on single mode systems, such as Farzin (2008) for buses in Sao Paulo and Rahbee (2008) for rail in Chicago. Other cases consider multi-stage trips in integrated systems. Examples are Zhao et al. (2007) in Chicago, Reddy et al (2009) in New York, Nassir et al (2011) in Minneapolis, and Gordon (2012) in London, as well as Munizaga and Palma (2012) for the bus-Metro system in Santiago. These matrices contain paid trips and trip stages only.

Due to fare evasion, some trips or trip stages are not registered in the fare collection system database. The correction and expansion of the OD matrix is not trivial when non-observed trips are unevenly distributed in time and/or space. For example, in Santiago, fare evasion in Metro, which has turnstiles and guards in all stations, is uncommon. Fare evasion rates in buses are higher in certain areas and during certain time periods. The methods developed to obtain public transport OD matrices from smartcard payment data do not take this into consideration, and usually assume that all trips, and all trip stages, are registered. If some are not registered (for any reason) the matrices will be incomplete and may be biased. For example, a non-paid bus trip will not be observed, and a bus-Metro trip where the Metro trip stage is paid, but the bus trip stage is not paid, will only be partially observed.

If a proportion of the trips were missing, and that proportion were homogeneously distributed across all services, then the construction of correction factors would be easy. One would determine the number of trips that were missing and apply a homogeneous expansion factor to the estimated OD matrix. However, as this is a non-homogeneous phenomenon that affects some areas more than others, as well as some trip stages more than others, developing correction factors is a much more complex problem. Particularly complex is the case of multi-stage trips where some stages are not registered, as this has an impact on the estimation of the origin or destination of the trip, and the correction needs to change the structure of those trips. This paper addresses this challenge.

The objective of this work is to develop and apply a method to incorporate fare evasion correction factors to public transport OD matrices obtained from AFC and GPS data. To the best of our knowledge, this is the first work that proposes correction factors for fare evasion. The method is applied to public transport OD matrices obtained from passive data in Santiago through cooperation between academia and the public transport authority (Gschwender et al, 2016). This work is a contribution towards establishing new methods to feasibly obtain OD matrices through the adequate merging of automatically-collected data with complementary traditional measurements and survey instruments.

The initial passive data OD matrices have been built using the methodology proposed by Munizaga and Palma (2012), Devillaine et al. (2012), and Munizaga et al. (2014). The effect of fare evasion on the estimated OD matrices depends on whether the evasion is total (during all trip stages) or partial (during some of the trip stages). Fig. 1 illustrates these two cases, considering a user who makes a trip from a to c, with a transfer at b. The left side of Fig. 1 corresponds to a case of total evasion, where none of the trip stages are validated; therefore, the whole trip is absent from the matrix, and a correction factor that adds this trip can be used to correct for it. The right side of Fig. 1 corresponds to a case of partial evasion, where the first trip stage is not paid, but the second trip stage is paid. This

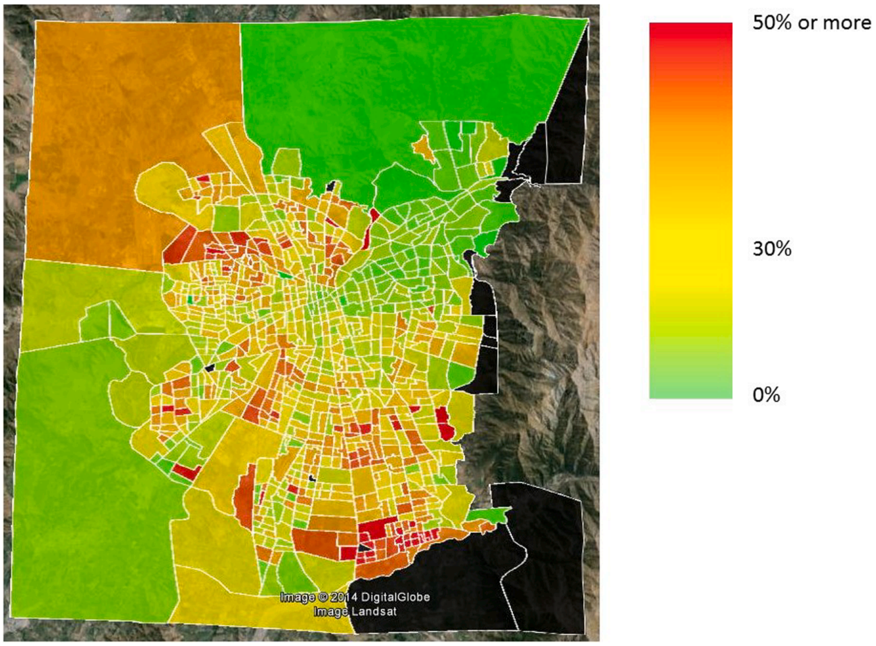


Fig. 2. Spatial distribution of measured fare evasion rate.

case is completely different from the OD matrix perspective, because the trip is accounted for in the matrix, but its origin is incorrectly estimated. Therefore, to correct for this case, we need a procedure or factor that modifies the trip structure, keeping the total number of trips constant.

Transantiago has a flat fare structure (independent of the distance travelled) and, as mentioned before, one fare payment covers up to three trip stages within a two-hour time window. Given this fare structure, the left-side case represents a financial benefit for the evader, because he or she is able to travel without paying. On the contrary, in the right-side case, there is no savings or financial benefit for the traveller. This case occurs when the passenger is not able to pay at the first stage; for example, when the smartcard does not have enough credit to pay for the trip and there is no fare loading point available at the trip origin, or when the bus is so crowded that the passenger enters the bus through one of the back doors (validation is only possible at the front door in Santiago), or the passenger forgot to bring their card. If the next trip stage is made on Metro, the user must solve the problem, either by charging their card or buying a new one, in order to access the Metro network, where fare evasion is almost inexistent. Fare evasion due to crowding conditions while boarding can also occur in the bus stage of a bus-Metro trip, but it is less common (Delbosc and Currie, 2019). As in the previous case, it does not represent a financial benefit for the passenger (the trip has already been paid).

2. Data

Fare evasion is regularly measured in Santiago on a sample of services, from which global estimations of fare evasion are obtained. These measurements are made by incognito observers located at each door of the sampled bus, who observe and count at each bus stop: (i) passengers who board and pay, (ii) passengers who board and do not pay (evaders), and (iii) passengers who alight. This information is used to estimate the load profile of the observed buses and the fare evasion rate at the observed bus stops. We used the 2012 sample (DICTUC, 2012) which contains 509,974 bus stop observations made during 10,155 bus expeditions. The spatial distribution of fare evasion is shown in Fig. 2, which shows the percentage of evaded trip stages over total trip stages for bus-stop observations in each zone, using an 800-zone zonification of Santiago. This spatial distribution is clearly non-homogenous.

We also use the paid trip and trip stage matrices obtained from smartcard data for a week in April 2013, and the results of the Metro Origin-Destination survey (Metro ODS) conducted in 2013. The matrix obtained from smartcard data is known as the "bip! Matrix", as bip! is the onomatopoeic name of the smartcard. The bip! Matrix from April 2013 is built with observations of 20 million trips made during the five workdays of a chosen week. These trips correspond to 29 million trip stages. The information that can be extracted from the data includes the service (or sequence of services) taken, boarding and alighting stops and timestamps for each trip stage, travel time and distance, and fare paid. GPS data is used to estimate the boarding and alighting point: using the location-at-time GPS information of every bus, which is measured and stored every 30 s, we estimate the time-at-location of bus stops for the route the bus is serving. The spatiotemporal information is very disaggregated, as we know the exact location of the bus stops and Metro stations, and from the smartcard transactions, we know the exact time of boarding. The boarding and alighting stops are estimated using a procedure described in Munizaga and Palma (2012); in the cases reported in that work, the proposed method is able to estimate the boarding and alighting stops in approximately 99% and over 80% of the cases, respectively. The validation of these estimations

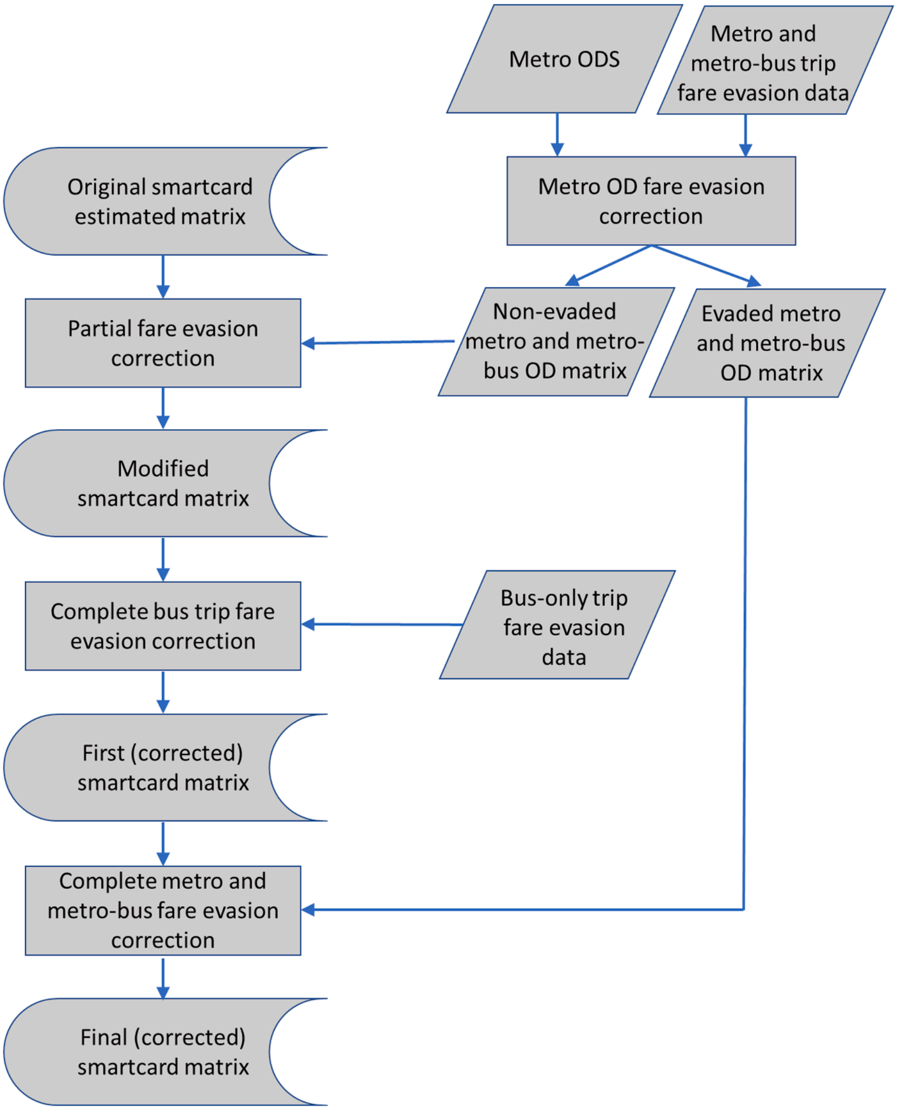


Fig. 3. Stepwise methodology for smartcard-estimated OD matrix.

reported in Munizaga et al. (2014), shows that the boarding location is correctly estimated in 99%, and the alighting stop in 84% of the cases.

AFC data, which is the basis for the OD matrix, is subject to possible errors in reliability, but the impact of these errors is minimized due to several factors. First, the information provided by the on-board smartcard validation machines ("validators") is used to define the amount paid to the network operators; therefore, it is in their best interest to maintain them. Second, there are at least two validators available at any boarding point, including buses. Third, there is no alternative payment method (no cash, passes, etc.). Fourth, because of the fare structure that allows up to three trip stages with one fare payment, errors with the validators could result in additional fare deductions, and users will complain.

GPS data is subject to more sources of error, such as GPS failure (there is only one GPS device per bus), signal interference, and inherent equipment errors. However, methods have been developed to detect and rectify errors. As described in Cortés et al. (2011), GPS pulses are projected to the route path, and observations that fall outside of a given distance buffer along the route path are omitted. These corrections are applied prior to the OD matrix estimation.

The Santiago Metro regularly carries out an ODS to evaluate system demand. The survey is undertaken at Metro stations, where passengers are asked about their mode of access to that station (walking or a specific bus route) and their intended destination and mode of egress (walking or a specific bus route), along with sociodemographic characteristics. The sample size for the 2013 Metro ODS is 156,350 observations (IPSOS, 2013). These observations are expanded, taking the sampling strategy into consideration to avoid potential bias.

3. Method

To correct the estimated smartcard OD matrices for fare evasion bias, we propose a method that sequentially addresses the two cases of fare evasion described above: partial fare evasion and complete trip fare evasion (whether intended or unintended). The general method is illustrated in Fig. 3. In the first stage, partial fare evasion is corrected using information about non-evaded Metro trips, obtained from the Metro Origin-Destination survey (Metro ODS). To extract these non-evaded trips from the Metro ODS, external information regarding fare evasion in Metro-only, bus-Metro, and Metro-bus trips is needed. With this information, cases of partial fare evasion in bus trip stages before and after a corresponding Metro trip stage are corrected. The second stage corrects the OD matrix to include bus-only trips that are completely evaded and, as they do not include a Metro trip stage, do not appear in the Metro ODS. To correct for complete bus trip fare evasion, external data about bus-only trip fare evasion is needed. Finally, a third stage corrects the OD matrix to include Metro, bus-Metro, and Metro-bus trips which are completely evaded. This correction considers information regarding evaded trips obtained from the Metro ODS and external Metro-involved trip fare evasion.

In Sections 3.1 and 3.2, we apply this general methodology to the case of Santiago. This case study has some particularities that simplify the procedure. First, fare evasion in the Metro system is negligible, making the third correction stage, regarding Metro-only trip fare evasion, unnecessary; furthermore, this means that we can directly obtain the Metro OD matrix from the Metro ODS without removing evaded trips. Second, partial fare evasion in post-Metro bus trip stages is much lower than partial fare evasion in bus trip stages prior to a Metro trip stage, as explained in the Introduction. Therefore, only the first-stage partial fare evasion case, where an initial bus stage is not paid but a subsequent Metro stage is paid, along with cases of complete trip fare evasion, need to be corrected. First-stage fare evasion bias is corrected using information from the Metro ODS. We assume that at Metro stations where the proportion of trips with bus access to Metro is greater in the Metro ODS than in the bip! OD matrix, fare evasion is causing a bias in the bip! OD matrix. This bias implies an overestimation of trips where the first trip stage is a Metro trip. Therefore, a correction factor is applied at the Metro station level to those stations where the assumed overestimation is observed. These factors reduce the number of trips with an initial Metro trip stage (Metro-only, Metro-bus) and increase the number of trips with an initial bus trip stage (bus-Metro, bus-Metro-bus) so that the trip structure observed in the Metro ODS is reproduced. Afterwards, the next module incorporates complete trip fare evasion, usually associated with intended evasion. This process uses fare evasion data at the bus stop level, applying factors for bus-only trips, so that the remaining fare evasion (that which is not explained by first-stage evasion) is explained. Given that the fare evasion measurements are taken at the trip stage level only, an iterative method is required to find those factors. The details and assumptions of the procedure to calculate both types of correction factors are provided in the next two sections.

In Section 3.3 we explain how the other modules of the methodology would operate if needed in other cases.

3.1. Partial fare evasion

In this section, we propose correction factors to adjust the smartcard-estimated OD matrices to the trip structure observed in the Metro ODS. The smartcard matrix is used as an a priori matrix, and correction factors are applied for each Metro station separately. We assume that the survey respondents declared all of their trip stages, regardless of whether they were paid or not, and that partial fare evasion is present only in trips that have at least one bus stage followed by a Metro stage.

As each trip reported in the Metro ODS provides information about the mode of access used to reach the Metro station, the surveyed trips can be classified into two types:

- Direct access to Metro: trips where the user walked to the Metro or accessed the Metro station using a non-integrated mode (taxi, bicycle, or car); i.e., a trip stage that is not registered in the smartcard database. In this case, the relevant geographic information for our method is the metro station accessed.
- Bus access to Metro: trips began with an initial bus trip stage prior to the Metro trip stage, according to the survey. In this case, we must identify the zone of origin, which is geocoded in the database, and the Metro station accessed.

For each Metro station where the smartcard matrix presents an overestimation of the number of trips initiated there (that is to say, it is considered feasible to correct for partial fare evasion), a proportional iteration factor procedure is applied. For those cases where no overestimation is observed, partial fare evasion correction is not necessary (or feasible). The sub-matrix of trips that originate or transfer in Metro stations according to the estimation from fare transactions is used to build the a priori matrix by extracting the pre-Metro and Metro trip stages only. Any trip stage that happens after the Metro trip stage is ignored. The a priori matrix is then iteratively adjusted to reproduce the trip distribution structure obtained from the Metro ODS. The model can be written as a doubly constrained growth factor model, as shown in Eq. (1), where $\left\{\widehat{t_{ij}^M}\right\}$ is the corrected element of the matrix, α_i is an expansion factor based on the trip origin, β_j is an expansion factor based on the trip destination, and $\left\{t_{ij}^0\right\}$ is the a priori matrix (estimated from fare transactions) considering pre-Metro and Metro boardings only.[2]

$$\widehat{t_{ij}^M} = \alpha_i \cdot \beta_j \cdot t_{ij}^0 \tag{1}$$

[2] A complete list of variables is presented in Appendix A.

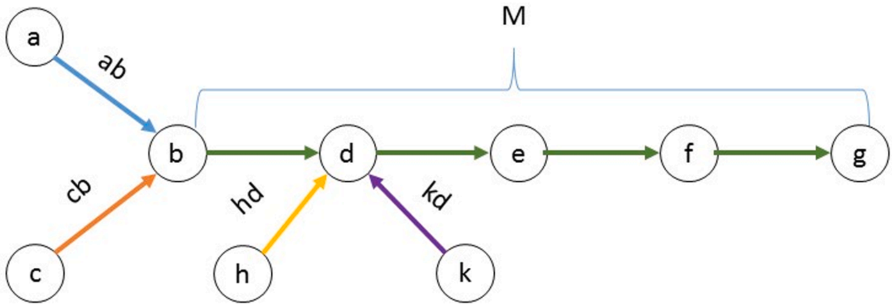


Fig. 4. Example of the sub-network used to test the method.

The constraints of the problem are that the summation of each column in the adjusted matrix must be equal to the total number of trips entering the subway system at Metro station M with destination j, D_j^M, and the summation of each row must be equal to the total number of trips produced by each zone i entering the subway system at station M $\left(O_i^M\right)$ according to the expanded Metro ODS.

$$\sum_i \widehat{t_{ij}^M} = D_j^M \tag{2}$$

$$\sum_j \widehat{t_{ij}^M} = O_i^M \tag{3}$$

To illustrate the method, we present a simple example in Fig. 4. To correct the smartcard matrices for the fact that some users may not validate onboard some feeder services (ab, cb, hd, kd), we build the a priori matrix as the sub-matrix of trips that go from the feeder areas to any station in the Metro network (b, d, e, f, g), along with those trips that initiate in a Metro station. In both cases, this is regardless of the final trip destination. This sub-matrix can be compared against a matrix obtained from a Metro OD survey, which only considers trips that include a Metro trip stage. If the Metro survey sample is sufficiently large, one would expect that it accurately reflects the real access structure, and the proposed method allows for its reproduction.

Additionally, we impose the constraint that corrected trip stages (S_i) as a proportion of total trip stages (S_{iODS}) must be less than or equal to the proportion of fare evasion ev_i measured for zone i. ev_i is obtained as the number of evaded trip stages divided by the total number of observed trip stages (DICTUC, 2012), i.e.:

$$\frac{S_i - S_{iODS}}{S_{iODS}} \leq ev_i \tag{4}$$

Afterwards, the iterative proportional fitting method described in de Ortúzar and Willumsen (2011) is applied to calculate the expansion factors α_i and β_j. As each trip in the original matrix is subject to a unique set of correction factors, the updated matrix can be built by applying the factors, estimated using the sub-matrices, directly to the general matrix.

Using the test network described above, we simulated several scenarios of demand and fare evasion, and evaluated the performance of the proposed method. The method worked well in all scenarios, with a final χ^2 below 0.03.[3] To further illustrate the method, a numerical example is provided in Appendix B.

3.2. Complete trip fare evasion

The method to correct for complete trip fare evasion uses trip stage evasion measurements (DICTUC, 2012) and the a priori matrix obtained from passive data sources (Munizaga and Palma, 2012) after correcting for partial fare evasion as described in Section 3.1. The trip stage evasion measurements are used to account for the total number of trip stages evaded, and the a priori matrix is used to represent the trip structure. We assume that it is not possible to observe the actual trip structure of fare evaders, as this would require following people during their complete trip. As this information is not available in Santiago, it is not possible to assess the difference in travel patterns between fare evaders and payers.

Therefore, to build correction factors for complete trip fare evasion, we make the conservative assumption that in the aggregate, fare evaders and paying bus users who board within the same zone have the same trip structure. This assumption holds at the trip stage level for bus trips only. We assume that complete trip fare evaders do not use Metro.

Considering that, at the zone level, the number of non-paid trips is a proportion of the total trips ev_i, as measured by DICTUC (2012), the number of observed trip stages associated with complete trip fare evasion S_{CTE_i} is calculated for zone i after discounting the partial evasion estimation S_{PE_i} from the total number of evaded trip stages, as shown in Eq. (5), where the first term on the left side if the

[3] $\chi^2 = \sum_{ij} \frac{(\widehat{t_{ij}} - t_{ij})^2}{t_{ij}}$, where $\widehat{t_{ij}}$ is the estimated matrix and t_{ij} is the observed matrix. This index, originally proposed for this kind of use by Gunn and Bates (1982), is used to evaluate prediction error.

equation corresponds to the number of evaded trip stages in zone i. This method is applied to an aggregated zonification that divides the Santiago metropolitan area into 803 zones.

$$S_{CTE\_i} = S_{Paid\_i} \cdot \frac{ev_i}{1 - ev_i} - S_{PE\_i} \tag{5}$$

To build the complete trip fare evasion correction factors, we use the trip distribution structure of trips made from zone i to any destination using bus only. We group them in the a priori matrix depending on their sequence, defined as the sequence of trip stages h as defined in Eq. (6).

$$h = \left(b_1, b_2, b_3, b_4\right) \tag{6}$$

We consider that trips may be comprised of up to four trip stages, which we denote b_1, b_2, b_3 and b_4. Most trips consist of up to three trip stages; four-stage trips are very uncommon. The total number of paid trips associated with sequence h is the summation of all trips whose sequence is h. The number of evaded trips that follow sequence h ($\widehat{t_h}$) can be expressed as a proportion γ_h of the paid trips that follow the same sequence h, as shown in Eq. (7).

$$\widehat{t_h} = \gamma_h \cdot t_h \tag{7}$$

In each iteration, the number of estimated trip stages associated with complete trip fare evasion in zone i is calculated as:

$$\widehat{S}^n_{CTE\_i} = \sum_h \gamma_h^n \cdot t_h \cdot \delta_i^h \tag{8}$$

where δ_i^h is equal to 1 if stop i belongs to sequence h; otherwise, it is equal to 0. The γ_h^n parameters are calculated with the proportional fitting method, and iteratively updated until the number of corrected trip stages is equal to the number of trip stages associated with complete trip fare evasion.

$$\gamma_h^0 = 1 \forall h \tag{9}$$

$$R_i^n = \frac{S_{CTE_i}}{\sum_h \gamma_h^n \cdot t_h \cdot \delta_i^h} \tag{10}$$

$$\gamma_h^{n+1} = \gamma_h^n \cdot \frac{\sum_h R_i^n \cdot \delta_i^h}{\sum_h 1 \cdot \delta_i^h} \tag{11}$$

Notice that R_i is the ratio between corrected and evaded stages associated with complete trip fare evasion at stop i. If it is greater than 1, the total expanded stages are overcorrected, and vice-versa. The correction factors are updated using these ratios, as expressed in Eq. (11), until the distance norm constraint (12) is satisfied. The objective of this step is to balance between zones where the different trip stages initiate.

$$|\varepsilon^n| = \sum_i \left| S_{CTE\_i} - \sum_h \gamma_h^n \cdot t_h \cdot \delta_h^i \right| \tag{12}$$

3.3. Other modules of the general methodology

In other cases, some (or all) of the other modules of the general methodology may be needed. If partial fare evasion in buses after a Metro trip stage is a concern, the partial fare evasion correction consists of two stages. In the first stage, the partial bus fare evasion occurring prior to a Metro trip stage is corrected according to the procedure detailed in Section 3.1. In the second stage, an analogous procedure is performed considering those trips where the bus partial evasion occurred immediately after a Metro trip stage. As both stages are independent and do not interact with each other, there is no problem in developing them in sequence. Special care would be needed while developing the Metro ODS to collect reliable information regarding bus trip stages made after the Metro trip stage. This is important because, as the survey is generally carried out at Metro stations, questions about post-Metro bus trip stages would be answered before the fact. This may cause a loss of quality of the answers and, subsequently, the data. This is not the case for questions regarding bus trip stages prior to the Metro trip stage, which are already complete when passenger is surveyed at the Metro station.

If fare evasion in the Metro system is relevant, the researcher must complete two additional steps. First, at the beginning of the general methodology, two matrices with Metro-involved trips are constructed; namely, a non-evaded Metro, Metro-bus, and bus-Metro OD matrix, and an evaded Metro, Metro-bus, and bus-Metro OD matrix. To do this, additional information regarding Metro-involved trips is needed. Only the non-evaded Metro OD matrix is used as an input for the partial fare evasion correction procedure. Second, the evaded trips that form the Metro OD matrix must be added at the end of the general methodology to complete the correction process and build the final corrected matrix.

4. Analysis of the results

To implement the proposed method in the case of Santiago, only the modules detailed in Sections 3.1 and 3.2 are needed. We used data from one week (April 15–21, 2013) to build the smartcard matrix using the Munizaga and Palma (2012) methodology; the 2013 Metro ODS, which contains 156,350 observations made between July 30 and November 26; and fare evasion measurements made during the second semester of 2012, which contain observations of 10,767 bus expeditions (DICTUC, 2012). We restrict our analysis to trips made on workdays between 6:30AM and 8:30PM. Running on an Intel(R) Core(TM) i5-3450 CPU @ 3.1 GHz, 4.00 GB of RAM and 64-bit Windows 7 OS, the total processing time was 23 min.

4.1. Partial fare evasion

Following the proposed methodology, the first step is to identify the Metro stations where the proportion of users who declared using the bus to access the Metro (according to the Metro ODS) was higher than the proportion observed in the smartcard matrix. This is true for 79 out of 100 stations. Therefore, a partial fare evasion correction factor is applied to these 79 stations.

Higher partial fare evasion factors were found at end-of-the-line Metro stations and at those connecting with main bus corridors. This is explained by the higher proportion of trips that reach the Metro station using the bus instead of walking.

This methodology calculates evasion rates at the trip level (instead of the trip stage level), where some trips are completely evaded, and others are only partially evaded. According to this methodology, 536,109 trips were made without paying the initial bus fare (i.e., partial evasion), out of a total of 10,357,930 trips observed in the smartcard sub-matrix of trips that originated or transferred at a Metro station. This corresponds to an overall figure of 5% for partial fare evasion; that is to say, on average, 5 of every 100 users that reach a Metro station do not pay for the prior bus trip stage.

The next steps are to discount this partial fare evasion from the matrix and then correct for complete trip fare evasion.

After obtaining the number of trip stages evaded between each zone and each Metro station corresponding to partial evasion, these trip stages are subtracted from the total trip stage evasion data obtained from DICTUC (2012), at the zonal level. This number of evaded trip stages is used in the next stage to correct for complete trip evasion.

Table 2 shows the total number of trip stages evaded and summarizes the evasion rates at the trip stage level. Partial evasion accounts for 12% of total trip stage evasion (3% over 26%). The other 88% corresponds to complete evasion.

Fig. 5 shows the spatial distribution of partial fare evasion. The green symbols show the 79 stations where a correction factor was applied, whereas the red symbols represent the remaining 21 stations where the correction factor was not feasible. The zone colour shows the balance between partial evasion and complete evasion in that area. This is calculated as the number of trip stages associated

Table 2
Bus evasion at trip stage level.

	Total	Paid	Evaded
Bus trip stages	18,511,304	13,698,635 74%	4,812,939 26%
Evaded bus stages of bus-only trips			4,276,830 23%
Evaded bus stages to reach Metro			536,109 3%

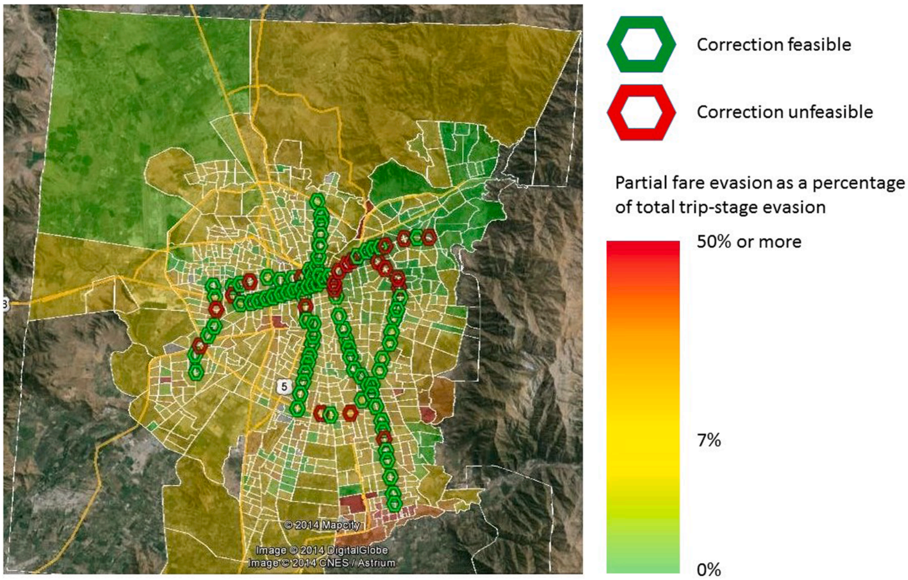


Fig. 5. Spatial distribution of partial fare evasion.

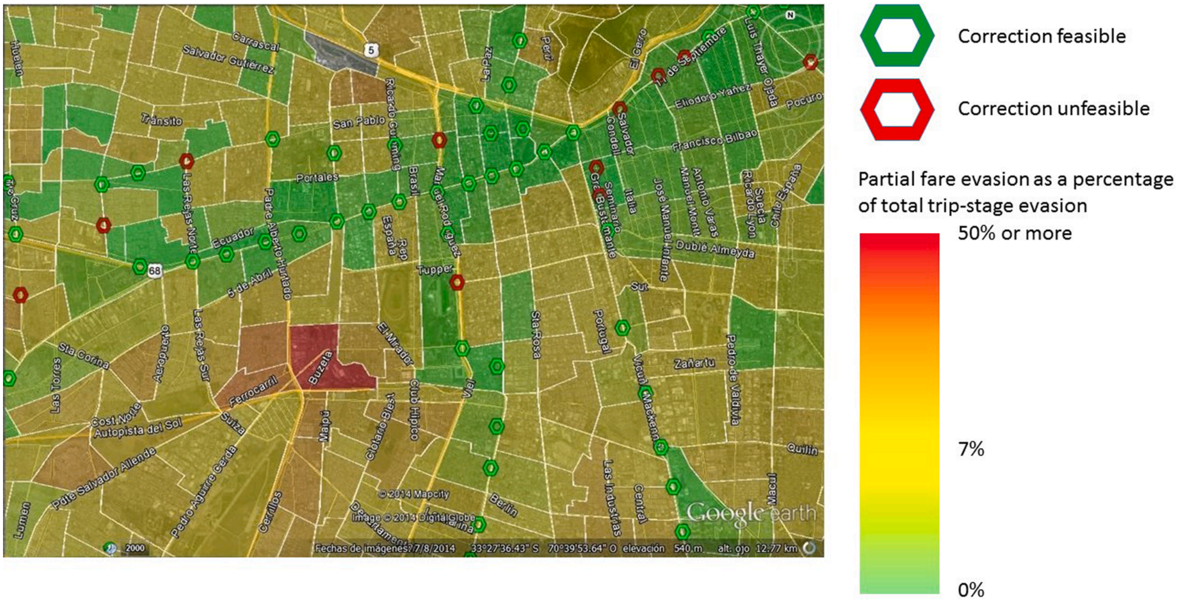


Fig. 6. Spatial distribution of partial fare evasion in the city centre.

with partial fare evasion, divided by the total number of trip stages evaded, expressed as a percentage. Green is associated with mainly complete trip fare evasion, whereas red indicates a more even split between complete trip and partial fare evasion.

Fig. 6 shows the city centre in more detail. It can be observed that a lower proportion of partial fare evasion is observed at zones near Metro stations. This can be explained by the fact that walking access to Metro is also an option in those zones.

4.2. Complete trip evasion

After discounting partially-evaded trip stages, the complete trip correction method is applied. The result of this process consists of a set of expansion factors based on the boarding zone of each trip stage, which allows for the expansion of the origin–destination matrix at the zonal level. To facilitate the interpretation of results, we aggregate the results to the municipality level (the Santiago metropolitan area is organized into 34 municipalities, as shown in Table 1).

The OD pairs with the highest and lowest complete trip evasion rates, and with more than 4000 trips, are shown in Tables 3 and 4 respectively. The seven highest pairs reach levels between 50% and 60%, while the lowest pairs remain below 18%. High-income municipalities such as Providencia, Las Condes, Vitacura, and Lo Barnechea (see income figures in Table 1) are predominant OD

Table 3
OD pairs with highest complete trip evasion rates.

Origin-Destination	Income	Metro connection	Paid	Evaded	Total	Evasion %
Puente Alto – Puente Alto	Low - Low	No	125,822	172,052	297,874	58
Puente Alto – La Pintana	Low - Low	No	6,994	9,643	16,637	58
Puente Alto – La Granja	Low - Low	No	6,057	7,866	13,923	56
Puente Alto – Santiago	Low - Medium	Yes	19,195	20,872	40,068	52
Renca –Pudahuel	Low - Low	No	4126	4228	8354	51
Renca – Cerro Navia	Low - Low	No	2306	2440	4747	51
La Granja – La Pintana	Low - Low	No	2121	2248	4369	51

Table 4
OD pairs with lowest complete trip evasion rates.

Origin- Destination	Income	Metro connection	Paid	Evaded	Total	Evasion %
Providencia – Las Condes	High - High	Yes	49,992	8724	58,716	15
Providencia – Vitacura	High - High	No	9824	1896	11,720	16
Santiago – Providencia	Medium - High	Yes	66,141	13,965	80,106	17
Las Condes – Providencia	High - High	Yes	51,889	10,544	62,432	17
Santiago – Independencia	Medium - Low	No	35,233	7,183	42,416	17
Las Condes – Santiago	High - Medium	Yes	30,740	6460	37,200	17
Santiago – Las Condes	Medium - High	Yes	25,731	5515	31,246	18

pairs with low rates complete trip evasion. This type of income effect was also observed by Guarda et al. (2016). Furthermore, we observed that Metro accessibility seems to have an effect as well. Most of the OD pairs with a high percentage of complete trip evasion do not have an easily-accessible Metro connection.

Our estimation yields an evasion rate for bus-only trips of 37%; in other words, 37 of every 100 trips that do not use Metro are not paid for (see Table 5). This figure is greater than the 26% evasion rate measured at the trip stage level. This is because, even though the numerator is smaller (as some evaded trip stages correspond to partial evasion), the denominator is much smaller (as it only accounts for trips without Metro usage, as Metro usage is associated with practically no evasion). Table 5 also shows the total trip evasion rate (20%). Note that the total number of trips evaded does not include trips with partial evasion, as those trips are indeed considered paid (not in the initial bus trip stage, but in the subsequent Metro trip stage). Note also that the 0% of evaded trips with a Metro trip stage is not a result, but rather an assumption, of the model. The Metro trip stage evasion rate is very small, according to manual measurements.

5. Conclusions

We propose a method to correct public transport OD matrices obtained from smartcard data, incorporating information regarding trips not registered in the smartcard database. This method incorporates two types of unobserved trips: partial fare evasion, in the case of bus-Metro combinations where the initial bus trip stage is not paid; and complete trip fare evasion. These two types of evasion have different effects on the estimated OD matrices. Partial evasion generates a bias towards Metro-only trips. Complete trip fare evasion progressively underestimates the trips in OD pairs with higher evasion rates. To be able to apply these methods and make reliable corrections, exogenous observations of passenger flows are required. These observations must be independent of payment technology systems and provide information about Metro access (proportion of bus/walking access) and evasion rates at the stop/station or zone level. The proposed methods consider independent factors, which are sequentially applied to the smartcard OD matrix.

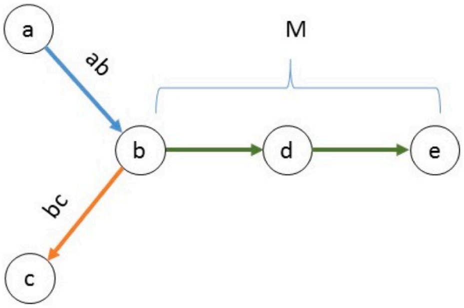
We apply this methodology to the case of fare evasion in Santiago, where the smartcard is the only available payment option and fare evasion within the Metro system is negligible. The partial evasion correction is applied only for bus-Metro combinations where the initial bus trip stage is not paid, as this is the predominant partial fare evasion case in Santiago. The negligible Metro fare evasion rate implies that the complete trip evasion correction procedure needs only be applied to bus-only trips. The implementation of the method shows that 5% of multiple-stage Metro trips include partial fare evasion. These figures are higher at some stations, depending on the public transport network structure and other variables. Stations that connect with important bus corridors or are at the end of a line require higher correction factors. In other stations, the effect is less dramatic. Partial evasion rates are lower for trips originating in zones close to Metro stations, where users can walk to the station instead of relying on bus access. At the trip stage level, partial evasion accounts for only 12% of the observed evaders, whereas the other 88% corresponds to trip stages evaded by complete trip evaders. Additionally, we estimate that 37% of bus-only trips and 20% of total trips (bus-only, Metro-only, and combined) correspond to cases of complete trip fare evasion. Complete trip fare evasion is found mainly in short intramunicipal trips and/or trips originating in low-income areas, especially those without Metro access.

This methodology allows for the determination of evasion rates not only at the trip stage level, but also at the complete trip level. This is important, as complete trip evasion is harder and more expensive to measure than trip stage evasion (evaders would need to be followed during their complete trip, instead of simply counting how many users pay and how many do not). Complete trip evasion affects system revenue, as opposed to partial evasion, where the fare is eventually collected at the Metro station. These calculations are made assuming that, for a given origin and time period, the trip structure of fare evaders is similar to that of the passengers who pay for their trips. This assumption is difficult to verify, as there is no readily-available information about the actual trip structures of fare evaders. They are not registered in the system databases and there are no surveys available in Santiago that address fare evasion behaviour in terms of trip structure. An interesting topic for future research would be to enrich the model with additional sources of information, such as mobile phone traces, automatic passenger counting (APC) systems, or surveys about the trip behaviour of fare evaders. This enriched data would allow us to explore the differences in the trip structure of fare evaders and payers.

In terms of other future research avenues, the methodology could be adapted to more complex fare structures such as zone fare systems, where partial evasion not only distorts the OD matrix, but also implies reduced revenue collection, because the registered trip is shorter than the actual trip. If fares change during the day (reduced off-peak fares), partial evasion could also lead to a different fare being charged, because the registered start time of the trip may differ from the actual start time. The methodology can also be considered to address other problems arising from partial smartcard database information. For example, if other payment methods (e.

Table 5
Evasion at trip level.

	Total	Paid	Evaded
Bus-only trips	12,104,766	7,623,735 63%	4,481,031 37%
Trips with Metro stage (Metro-only and bus + Metro)	10,357,930	10,357,930 100%	0 0%
Total trips	22,462,696	17,981,665 80%	4,481,031 20%



	b	c	d	e	Σ
a	500	400	350	130	1,380
b		200	300	100	600
d				150	150
Σ	500	600	650	380	**2,130**

Fig. B1. Public transport network.

Table B1
Partial fare evasion correction iterations.

Iteration	α_a	α_b	β_d	β_e	Difference
1	1.04167	0.95000	1.00000	1.00000	–
2	1.04167	0.95000	1.00064	0.99819	0.191871%
3	1.04169	0.94997	1.00064	0.99819	0.003586%
4	1.04169	0.94997	1.00064	0.99819	0.000106%
5	1.04169	0.94997	1.00064	0.99819	0.000002%
6	1.04169	0.94997	1.00064	0.99819	0.000000%

Table B2
Evasion survey expansion.

Boarding bus stop	Paid stages S_{paid_i}	Evasion rate ev_i	Unpaid $S_{Paid\_i}\cdot\frac{ev_i}{1-ev_i}$	First stage evasion S_{PE_i}	Bus-only trips evaded stages S_{CTE_i}
a	1380	9.21%	140	20	120
b	600	6.25%	40	0	40

Table B3
Complete trip fare evasion correction factor iterations.

Iter	γ^n_{ab}	γ^n_{bc}	γ^n_{ac}	$\widehat{t_{ab}}$	$\widehat{t_{bc}}$	$\widehat{t_{ac}}$	$\sum_h \gamma^n_h \cdot t_h \cdot \delta^h_{ab}$	$\sum_h \gamma^n_h \cdot t_h \cdot \delta^h_{bc}$	R_{ab}	R_{bc}	$\lvert\varepsilon\rvert$
1	1.000	1.000	1.000	500	200	400	900	600	0.13	0.07	1,340
2	0.133	0.067	0.100	66.67	13.33	40.00	106.7	53.3	1.13	0.75	26.7
3	0.150	0.050	0.094	75.00	10.00	37.50	112.5	47.5	1.07	0.84	15.0
4	0.160	0.042	0.089	80.00	8.42	35.79	115.8	44.2	1.04	0.90	8.4
5	0.166	0.038	0.087	82.91	7.62	34.74	117.6	42.4	1.02	0.94	4.7
6	0.169	0.036	0.085	84.57	7.20	34.12	118.7	41.3	1.01	0.97	2.6
7	0.171	0.035	0.084	85.50	6.97	33.76	119.3	40.7	1.01	0.98	1.5
8	0.172	0.034	0.084	86.03	6.84	33.56	119.6	40.4	1.00	0.99	0.8
9	0.173	0.034	0.084	86.32	6.77	33.45	119.8	40.2	1.00	0.99	0.5
10	0.173	0.034	0.083	86.48	6.73	33.39	119.9	40.1	1.00	1.00	0.3
11	0.173	0.034	0.083	86.57	6.71	33.36	119.9	40.1	1.00	1.00	0.1
12	0.173	0.034	0.083	86.62	6.70	33.34	120.0	40.0	1.00	1.00	0.1
13	0.173	0.033	0.083	86.65	6.70	33.33	120.0	40.0	1.00	1.00	0.0

g., single paper tickets or travel cards) are available, this information may not be included in the automatically collected smartcard data and could greatly distort the estimated OD matrices. Adapting the methodology presented in this paper, OD matrices could be constructed from the smartcard databases and corrected to include the unregistered information from paper tickets.

The correction procedures presented assume that the information provided by traditional surveys and measurements is good enough to be used as a basis for correction. The veracity of this assumption obviously depends on the quality of the surveys and measurements themselves. Besides methodological considerations, the quality of the results of surveys and measurements strongly depends on the amount of funding that is available for this type of manual data collection. As measurements, and especially surveys, can be quite expensive, automatically-collected data can free up funds to be used for other surveys that cannot be replaced using passive data collection. In fact, we could expect that traditional surveys and measurements will be replaced as information gleaned from automatically-collected data increased the need for simpler (and less expensive) surveys and manual measurements, for which more funds could be available, thus improving their quality. For example, by freeing up funds used for traditional OD surveys, larger samples of trip stage evasion measurements could be gathered, allowing for evasion estimations for different periods of the day. It is improbable that automatically-collected databases will provide all information needed. Therefore, a key issue to resolve in the near future will be the consolidation of joint estimations of OD matrices, considering both information source types: smartcard data and complementary surveys. The incorporation of other sources of automatically-collected data, such as smartphone or toll road data, is another interesting challenge. This work presented a real example of how a successful merger of both types of information can be achieved to not only obtain an experimental OD matrix exclusively based on smartcard data, but also to adequately incorporate information about non-registered trips using information from traditional surveys, therefore obtaining a more complete public transport OD matrix. These methodologies, which merge automatically-collected data with traditional surveys and manual measurements to construct richer, more relevant databases, can help build OD matrices that are more frequently-updated, lower-cost, and reliable when compared to traditional survey methods.

CRediT authorship contribution statement

Marcela A. Munizaga: Conceptualization, Methodology, Validation, Writing - original draft, Writing - review & editing. **Antonio Gschwender:** Conceptualization, Methodology, Validation, Writing - review & editing. **Nestor Gallegos:** Methodology, Software, Data curation, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the Complex Engineering Systems Institute, Chile (ANID PIA/BASAL GRANT: AFB180003) and Fondecyt (1161589). We would like to thank the contributions made by anonymous reviewers, who helped us to improve our submission with their valuable comments and suggestions.

Appendix A. List of variables

$\left\{t_{ij}^{0}\right\}$ a priori matrix estimated from fare transactions considering pre-Metro and Metro boardings only
$\left\{\widehat{t_{ij}^{M}}\right\}$ matrix corrected for partial fare evasion
α_i trip origin expansion factor
β_j trip destination expansion factor
D_j^M the summation of each column of the adjusted matrix
O_j^M the summation of each row of the adjusted matrix
ev_i fare evasion proportion measured at stop i
χ^2 distance indicator calculated as $\sum\left\{(\widehat{t_{ij}} - t_{ij})^2/t_{ij}\right\}$
S_i number of trip stages (corrected) with origin in i
S_{iODS} total trip stages in the origin–destination survey matrix with origin in i
S_{CTE_i} number of stages associated with complete trip fare evasion in zone i
S_{PE_i} number of stages associated with partial fare evasion in zone i
S_{paid_i} number of paid trip stages in zone i
h observed sequence of trip stages
$\widehat{t_h}$ number of evaded trips that follow sequence h
$t_h$number of paid trips that follow sequence h
γ_h proportion of the paid trips that follow the same sequence h
δ_i^h dummy variable that takes the value 1 if stop i belongs to sequence h, and zero otherwise
R_i ratio between corrected and evaded stages associated with complete trip evasion in zone i
ε_i difference between the observed boardings and those predicted by the model at stop i

Appendix B. Numerical example

To illustrate the method, we propose a simple public transport network, described in Fig. B.1, where *ab*and *bc* are two different bus routes, and *M* is a Metro system with two segments. The matrix shown in Fig. B.1 is the OD matrix estimated from AFC and GPS data.

In this network, there is only one route to go from one stop to another. The types of trips that can occur are bus-only $\{t_{ab}, t_{ac}, t_{bc}\}$, bus-Metro $\{t_{ad}, t_{ae}\}$ and Metro-only trips $\{t_{bd}, t_{be}\}$. Let us also suppose that the proportion of users who declare bus access to Metro in the Metro ODS is 0.57 (25 from 44 observations), and that the expanded Metro trips are 500 with origin in *a* and 380 with origin in *b*, which are the target values related to the ODS matrix. The observed evasion rates are 9% at bus stop *a* and 6% at bus stop *b*. There is no fare evasion in the Metro.

Partial fare evasion correction (bus-Metro trips):

According to the methodology, the first step is to compare the proportion of users who declare using the bus to access the Metro in the Metro OD survey with the proportion observed in the smartcard matrix.

To apply the proposed method, we build the a priori matrix t_{ij}^{0} as the sub-matrix that contains the trips that originate or transfer at a certain Metro station. In our case, Metro station b.

$$t_{ij}^{0} = \begin{array}{c|cc|} & d & e \\ a & 350 & 130 \\ b & 300 & 100 \end{array} \tag{B.1}$$

With these values, we calculate the proportion of trips with bus access to Metro $\frac{O_a^0}{O_a^0+O_b^0}$

$$O_a^0 = t_{ae}^0 + t_{ad}^0 = 350 + 130 = 480 \tag{B.2}$$

$$O_b^0 = t_{be}^0 + t_{bd}^0 = 300 + 100 = 400 \tag{B.3}$$

Resulting in a partition of 54.5%. As this figure is lower than the proportion obtained from the Metro ODS (57%), we assume that partial fare evasion exists, and the partial fare evasion correction is applied. The Metro survey is expanded to adjust the origins to the new total (880 total trips), resulting in$\widehat{O_a} = 500$, $\widehat{O_b}$=380.

We then apply a bi proportional method to obtain the corrected matrix, which explains each trip with an a priori matrix (t_{ij}^0), and two factors, $\alpha_i$explaining the expansion for origins, and β_j, adjusting for destinations.

$$\widehat{t_{ij}} = \alpha_i * \beta_j * t_{ij}^0 \tag{B.4}$$

Since we assume that destinations are correct in the original matrix, we will not make any correction to these totals[4,5]. The constrains are laid out in Eqs. (B.5) to (B.8).

$$\widehat{t_{ad}} + \widehat{t_{ae}} = \widehat{O_a} = 500 \tag{B.5}$$

$$\widehat{t_{bd}} + \widehat{t_{be}} = \widehat{O_b} = 380 \tag{B.6}$$

$$\widehat{t_{ad}} + \widehat{t_{bd}} = D_d = 650 \tag{B.7}$$

$$\widehat{t_{ae}} + \widehat{t_{be}} = D_e = 230 \tag{B.8}$$

Following our methodology, we iterate to adjust equations α_i and β_j as shown in Eqs. (B.9) and (B.10) respectively

$$\frac{O_i}{\sum_j \beta_j * t_{ijM}^0} = \alpha_i \tag{B.9}$$

$$\frac{D_j}{\sum_i \alpha_i * t_{ijM}^0} = \beta_j \tag{B.10}$$

Initializing $\beta_j = 1\forall j$, we run iterations on each group of parameters, α_i and β_j. The iterations are shown in Table B.1.

$$\widehat{t_{ij}} = \begin{array}{c|cc} & d & e \\ \hline a & 364.8 & 135.2 \\ b & 285.2 & 94.8 \end{array} \tag{B.11}$$

The corrected sub-matrix B.11 increases by 20 (364.8–350 + 135.2–130) the number of trips with origin in *a* (bus access to Metro) and decreases trips beginning in the Metro station by the same amount. The new matrix, corrected for partial fare evasion, is presented in B.12.

	b	*c*	*d*	*e*	*Σ*
a	500	400	365	135	1,400
b		200	285	95	580
d				150	150
Σ	500	600	650	380	**2,130**

(B.12)

Complete trip correction (bus-only trips):

[4] $\widehat{t_{ad}} + \widehat{t_{bd}} = t_{ad}^0 + t_{bd}^0$

[5] $\widehat{t_{ae}} + \widehat{t_{be}} = t_{ae}^0 + t_{be}^0$

Using the trip stage evasion measurements at bus stops a and b, we estimate the total number of trip stage-level evaded trips to be corrected by the expansion factors (see Table B2).

Since we already have explained 20 trip stages evaded between bus stop a and the Metro station, corresponding to cases of partial evasion, these are subtracted from the total trip stage evasion data at the zonal level, following Eq. (5). This modified number of evaded trip stages is used in the next step to correct for complete trip evasion.

Note that S_{paid_i} includes both bus and bus-Metro trips. Bus-only trips can be isolated from the OD matrix. Here we have three cases of bus only trips. Following Eq. (6), we define: h = 1 going from a to b boarding service ab ($\delta^h_{ab} = 1$, δ^h_{bc}=0), h = 2 going from b to c boarding service bc ($\delta^h_{ab} = 0$, δ^h_{bc}=1) and h = 3 going from a to c boarding services ab and bc ($\delta^h_{ab} = 1$, δ^h_{bc}=1). A correction or expansion in sequence 3 trips will explain trip stage evasion for both services.

	b	c	Σ
a	500	400	900
b		200	200
Σ	500	600	**1.100**

(B.13)

Following Eq. (9) we initialize every γ_h as 1, and apply Eqs. (9) and (10). The first iteration results in

$$R_a = \frac{S_{CTE_a}}{\sum_h \gamma^n_h \cdot t_h \cdot \delta^h_i} = \frac{120}{500\cdot 1 + 400\cdot 1} = \frac{120}{900} = 0.1333 \tag{B.14}$$

$$R_b = \frac{S_{CTE_b}}{\sum_h \gamma^n_h \cdot t_h \cdot \delta^h_i} = \frac{40}{200\cdot 1 + 400\cdot 1} = \frac{40}{600} = 0.0666 \tag{B.15}$$

$$\gamma^1_{ab} = 1\cdot\frac{\sum_h R_i \cdot \delta^h_{ab}}{\sum_h 1\cdot \delta^h_{ab}} = \frac{0.1333}{1} = 0.1333 \tag{B.16}$$

$$\gamma^1_{bc} = 1\cdot\frac{\sum_h R_i \cdot \delta^h_{bc}}{\sum_h 1\cdot \delta^h_{bc}} = \frac{0.0666}{1} = 0.0666 \tag{B.17}$$

$$\gamma^1_{ac} = 1\cdot\frac{\sum_h R_i \cdot \delta^h_{ac}}{\sum_h 1\cdot \delta^h_{ac}} = \frac{(0.1333 + 0.0666)}{1+1} = \frac{0.2}{2} = 0.1 \tag{B.18}$$

According to Eq. (12), the difference $|\varepsilon|$in the first iteration is:

$$|\varepsilon| = |120 - 900| + |40 - 600| = 1340 \tag{B.19}$$

Applying these equations iteratively, we obtain a result that converges when $|\varepsilon|$approaches zero. These steps are illustrated in Table B.3.

The ratios R_{ab} and R_{bc} represent the ratio between corrected and evaded stages; as they approach 1, the error $|\varepsilon|$ in total trips approaches 0.

The final corrected matrix, where the evaded trips are added to the matrix already corrected for partial fare evasion, is presented in B.20. Besides the difference in the OD matrix structure, the corrected matrix has 6% more trips than the original matrix.

	b	c	d	e	Σ
a	587	433	365	135	1,520
b		207	285	95	587
d				150	150
Σ	587	640	650	380	**2,257**

(B.20)

In order to analyze the effect that the assumption that fare evaders have the same trip structure as those who pay might have on the results, we carried out sensitivity analysis scenarios considering two alternative assumptions:

Alternative Assumption 1: If the proportion of two-stage trips made by fare evaders is only 50% of those made by paid passengers, we obtain the matrix described in B.21. Besides having more *ab* and *bc* trips, the total number of trips is also greater in this case (6,7% greater than that of the original matrix).

	b	c	d	e	Σ
a	604	416	365	135	1,520
b		224	285	95	604
d				150	150
Σ	604	640	650	380	**2,273**

(B.21)

Alternative Assumption 2: Following the same tendency, if all fare evaders make only one-stage trips, the total number of trips would be 7,5% greater than that of the original matrix.

	b	c	d	e	Σ
a	620	400	365	135	1,520
b		240	285	95	620
d				150	150
Σ	620	640	650	380	**2,290**

(B.22)

References

Amaya, M., Cruzat, R., Munizaga, M.A., 2018. Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis. J. Transp. Geogr. 66, 330–339.

Barabino, B., Salis, S., Useli, B., 2014. Fare evasion in proof-of-payment transit systems: Deriving the optimum inspection level. Transport. Res. Part B: Methodol. 70, 1–17.

Barabino, B., Salis, S., Useli, B., 2015. What are the determinants in making people free riders in proof-of-payment transit systems? Evidence from Italy. Transport. Res. Part A: Policy Pract. 80, 184–196.

Boyd, C., Martini, C., Rickard, J., Russell, A., 1989. Fare evasion and non-compliance: A simple model. J. Transp. Econ. Policy 189–197.

CASEN, 2013. Encuesta de Caracterización Socioeconómica Nacional. Santiago, Gobierno de Chile.

Correa, J., Harks, T., Kreuzen, V.J.C., Matuschke, J., 2017. Fare Evasion in Transit Networks. Oper. Res. 65 (1), 165–183.

Cortés, C., Gibson, J., Gschwender, A., Munizaga, M.A., Zúñiga, M., 2011. Commercial bus speed diagnosis based on GPS-monitored data. Transport. Res. Part C: Emerg. Technol. 19, 695–707.

Dai, Z., Galeotti, F., Villeval, M.C., 2017. Cheating in the lab predicts fraud in the field: An experiment in public transportation. Manage. Sci. 64, 1081–1100.

Dauby, L., Kovacs, Z., 2007. Fare evasion in light rail systems. Transport. Res. E-Circular, (E-C112).

Delbosc, A., Currie, G., 2016. Cluster analysis of fare evasion behaviours in Melbourne, Australia. Transp. Policy 50, 29–36.

Delbosc, A., Currie, G., 2019. Why do people fare evade? A global shift in fare evasion research. Transp. Rev. 39 (3), 376–391.

Devillaine, F., Munizaga, M., Trépanier, M., 2012. Detection of activities of public transport users by analyzing smart card data. Transport. Res. Rec.: J. Transport. Res. Board 2276, 48–55.

DICTUC, 2012. Mediciones de perfil de carga con evasión en la ciudad de Santiago. Estudio realizado por DICTUC para la Coordinación Transantiago.

DTPM (2016) Informes de Gestión, Tablas de Datos DTPM Período 2015, Directorio de Transporte Público, Ministry for Transportation and Telecommunications Chile. http://www.dtpm.gob.cl/index.php/2013-04-29-20-33-57/informes-de-gestion retrieved 13-02-2017.

Farzin, J., 2008. Constructing an automated bus origin-destination matrix using farecard and global positioning system data in Sao Paulo, Brazil. Transport. Res. Rec.: J. Transport. Res. Board 2072, 30–37.

Factor Estratégico, 2010. Informe final estudio cualitativo de opinión: El fenómeno de la evasión. Estudio realizado por Factor Estratégico para el Ministerio de Transportes y Telecomunicaciones de Chile.

Gordon, J.B., 2012. Intermodal passenger flows on London's public transport network: automated inference of full passenger journeys using fare-transaction and vehicle-location data. Doctoral dissertation. Massachusetts Institute of Technology.

Gschwender, A., Munizaga, M., Simonetti, C., 2016. Using smart card and GPS data for policy and planning: The case of Transantiago. Res. Transport. Econ. 59, 242–249.

Guarda, P., Galilea, P., Paget-Seekins, L., de Ortuzar, J.D., 2016. What is behind fare evasion in urban bus systems? An econometric approach. Transport. Res. Part A: Policy Pract. 84, 55–71.

Gunn, H.F., Bates, J.J., 1982. Statistical aspects of travel demand modelling. Transport. Res. Part A: Policy Pract. 16, 371–382.

Ipsos, 2013. Encuesta Origen Destino de viajes 2013. Realizado por Ipsos para Metro S.A.

Lee, J., 2011. Uncovering San Francisco, California, Muni's proof-of-payment patterns to help reduce fare evasion. Transport. Res. Rec.: J. Transport. Res. Board 2216, 75–84.

Munizaga, M., Devillaine, F., Navarrete, C., Silva, D., 2014. Validating travel behavior estimated from smartcard data. Transport. Res. Part C: Emerg. Technol. 44, 70–79.

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. Transport. Res. Part C: Emerg. Technol. 24, 9–18.

Muñoz, V., Thomas, A., Navarrete, C., Contreras, R., 2016. Encuesta origen-destino de Santiago 2012: Resultados y validaciones. Ingeniería de Transporte 19, 21–36.

Nassir, N., Khani, A., Lee, S., Noh, H., Hickman, M., 2011. Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. Transport. Res. Rec.: J. Transport. Res. Board 2263, 140–150.

Ortúzar, J.de.D., Willumsen, L.G., 2011. Modelling Transport. Fourth Edition, Wiley.

Rahbee, A., 2008. Farecard passenger flow model at Chicago transit authority, Illinois. Transport. Res. Rec.: J. Transport. Res. Board 2072, 3–9.

Reddy, A., Kuhls, J., Lu, A., 2011. Measuring and controlling subway fare evasion: improving safety and security at New York City transit authority. Transport. Res. Rec.: J. Transport. Res. Board 2216, 85–99.

Reddy, A., Lu, A., Kumar, S., Bashmakov, V., Rudenko, S., 2009. Entry-only automated fare-collection system data used to infer ridership, rider destinations, unlinked trips, and passenger miles. Transport. Res. Rec.: J. Transport. Res. Board 2110, 128–136.

Smith, M.J., Clarke, R.V., 2000. Crime and public transport. Crime Justice 27, 169–233.

Troncoso, R., de Grange, L., 2017. Fare evasion in public transport: A time series approach. Transport. Res. Part A: Policy Pract. 100, 311–318.

Yin, Z., Jiang, A.X., Johnson, M.P., Kiekintveld, C., Leyton-Brown, K., Sandholm, T., Sullivan, J.P., 2012. TRUSTS: Scheduling Randomized Patrols for Fare Inspection in Transit Systems. AI Magazine 33, 59–71.

Zhao, J., Rahbee, A., Wilson, N.H., 2007. Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. Comput.-Aided Civ. Infrastruct. Eng. 22, 376–387.