UNDERSTANDING FARE EVASION RATES IN PUBLIC TRANSPORT

Luis-Angel Cantillo · Juan Carlos Muñoz · Sebastián Raveau · Paula Iglesias

Abstract Fare evasion is one of the main problems of many transport systems around the world. This unethical behavior can reduce the income of a transportation system by millions of dollars, seriously affecting its viability. Common strategies to reduce fare evasion are to increase inspection and fines; however, analyzing the problem in this way is very generic and simplistic. It is necessary to understand fare evasion to effectively address it. This study presents an econometric model for estimating evasion rates, incorporating socioeconomic, accessibility and operational variables to explain the levels of evasion in Transantiago, Santiago’s (Chile) public transport system, and its spatial distribution across the city.

Keywords Fare Evasion · Unethical Behavior · logistic regression · Transantiago

Luis-Angel Cantillo
Pontificia Universidad Católica de Chile
Santiago, Chile
E-mail: lacantillo@uc.cl

Juan Carlos Muñoz
Pontificia Universidad Católica de Chile
Santiago, Chile
E-mail: jcm@ing.puc.cl

Sebastián Raveau
Pontificia Universidad Católica de Chile
Santiago, Chile
E-mail: sraaveau@ing.puc.cl

Paula Iglesias
DICTUC
Santiago, Chile
E-mail: piglesias@dictuc.cl
Fare collection is a fundamental component for the sustainability of any public transport system; therefore, fare evasion generates a negative impact on the finances and even affects the perception of the system (Reddy et al, 2011). Therefore, it is important for a system to know the level of fare evasion, at the highest possible spatial and temporal detail and its potential causes; to be able to implement effective measures.

However, measuring evasion in a detailed manner in a large city is extremely costly, and, therefore having specific indicators, for different moments of time and with regular updating, is not usually feasible. What is required, then, is to have adequate estimates of the phenomenon. For this, this study seeks to explain observed evasion, initially from socioeconomic operational variables, which have not been traditionally considered in other models found in the literature.

Levels of fare evasion in public transport have been explained by several factors; such as the level-of-service of the system, inspection rates and fines, the period of the day and the payment system, among others (Torres-Montoya, 2014). In addition, Lee (2011) found significant differences in fare evasion levels throughout the day, and a strong correlation between inspection levels and fare evasion was also detected (Clarke et al, 2010). Recently, Guarda et al (2016) estimated counting regression models for Santiago (Chile), using a database of the National Program for Transportation Supervision, and explain the number of evaders that board through a bus door at a certain location. The results indicate that evasion increases when: i) more people board or alight through the door, ii) passenger boarding is through the back doors, iii) the buses are busier and the interval between arrivals at the bus stop is greater, iv) the buses have more doors, and v) the income level of the commune where the bus stops are located decreases.

Regarding the type of model, Barabino et al (2015) tested the use of logistic regression to estimate the probability of evading. In their study, they sought to explain a binary variable of propensity to evade, based on categorical socio-demographic variables. Another alternative is the use of counting models (Guarda et al, 2016), which consider regression models. To implement this alternative the dependent variable would become "number of evaders", instead of considering the "evasion rate". With this alternative, a direct relationship was found between the total number of passengers boarding the bus and the number of people evaded. This variable represents most of the explanatory capacity of the model, reducing the effect of other relevant variables that affect the disposition to evade. If, on the contrary, the variable "people boarding" is excluded from the models, significant predictive capacity is lost. Therefore, it was preferred to test other functional forms, to avoid this problem. This same study suggests testing logistic regression models for estimating the evasion rate, since it allows limiting the dependent variable to the range of values between 0 and 1.
2 Problem Description and Model Formulation

2.1 Problem description

Fare evasion rate is the number of people who do not validate, over the total number of people who board, whether at bus, service, whereabouts or zone. The methodology applied for this stage of the study was as follows: create a database, establish the calibration sample and with it calibrate evasion estimation models and validate the models. As indicated in the literature review, it was considered appropriate to use a logistic regression model to estimate the overall rate of fare evasion.

Logistic regression corresponds to a statistical technique that has been widely used in practice Hosmer Jr et al (2013). This type of statistical models estimates the continuous probability of occurrence of an event based on the use of explanatory variables. The model computes a logistic function (1) that estimates the probability of occurrence of a phenomenon \( P_i \), in relation to the dependence that said phenomenon does not occur \( 1 - P_i \); for this, the vectors of dependent variables \( x_i \) are used as input data, with which the coefficients \( \beta \) (2) have to be calibrated.

\[
P_i = \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} = \text{Fare Evasion (1)}
\]

The calibration process of the parameters \( \beta \) is carried out by means of the maximum likelihood method (Eliason, 1993), which maximizes the estimated probability of obtaining the observed results. In summary, equation (2) is maximized, where \( NV_i \) is the number of individuals that evaded in observation \( i \), and \( V_i \) the number of individuals that validated in that observation.

\[
\max_{\beta} L = \prod_i (P_i(\beta))^{NV_i}(1 - P_i(\beta))^{V_i} \quad (2)
\]

2.2 Model formulation

Three different models were estimated, one for each of the periods of the day to be analyzed. This is based on the hypothesis that behavior associated to mandatory trips (such as those for "work" or "study", which are made in peak periods) may be different from the one observed in trips with other purposes (which are usually carried out in off-peak periods).

Certain variables evaluated in the models could have a different effect (different sign) in the evasion, depending on the period of the day; as indeed it happened with some of the included variables.
2.2.1 Dependent variable

For the estimated logistic regression models, the dependent variable was defined as the fare evasion rate of the bus stop, calculated according to the following equation, where $I = \text{busstops}$, $J = \text{buslines}$ and $L = \text{periods}$:

$$\text{FareEvasion}_{ijkl} = \frac{N_{\text{evaders}}_{ijkl}}{N_{\text{evaders}}_{ijkl} + N_{\text{payers}}_{ijkl}} \quad \forall i \in I \quad \forall j \in J \quad \forall l \in L \quad (3)$$

2.2.2 Independent variables

The explanatory variables used in the formulation of the model can be classified into: socioeconomic, accessibility and operational. The first is related to the socioeconomic conditions of the area where a bus stop is located. The next category includes variables that measure the proximity of the bus stop to other public transport services. Finally, the third category includes variables related to level-of-service offered by the transportation system, as well as binary variables to distinguish between operators (Transantiago is comprised by seven different operators).

Other variables associated with service typologies were also tested. On the one hand, services were classified into trunks, feeders and trunk-feeders; on the other hand, they were categorized according to the length of the route or according to the direction of the routes. All these variables turned out to be not statistically significant.

3 Experiments and results

3.1 Case study

The methodology proposed in this research was applied to the case study of Transantiago, the transportation system of the city of Santiago (Chile). We worked with the database developed for the Metropolitan Public Transport Directory (DICTUC, 2012), which measured fare evasion aboard the buses of Transantiago at a very detailed level: bus - bus stop - bus line - direction - period level, for the seven Transantiago operators.

The database is made up of 573,249 records, distributed in 143,312 records in “Morning peak hours” (end of travel time between 8:00-9:00), 166,252 records in “Evening peak hours” (start time of travel between 18:00 and 19:00) and the remaining 263,685 in “Off-peak hours” (mid travel time between 10:00-12:00).

The set of explanatory variables present in the best of the estimated models, for each of the analysis periods, is presented in Table 1. These variables were selected for inclusion on the model, following the criteria of consistency of signs and statistical significance of the estimated parameters (Ortuzar and
The table presents the value of coefficient and t-test for each of the variables, as well as the value of the log-likelihood of the model for each period.

### Table 1: Estimation of logistic regression models

<table>
<thead>
<tr>
<th>Variable</th>
<th>Morning peak hour</th>
<th>Off-peak hour</th>
<th>Evening peak hour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Test-T</td>
<td>Coef.</td>
</tr>
<tr>
<td><strong>Socioeconomic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Priority Index (SPI)</td>
<td>0.0101</td>
<td>6.17</td>
<td>0.0116</td>
</tr>
<tr>
<td>Average Income Zone Person (M$ CLP)</td>
<td>-0.001</td>
<td>-5.12</td>
<td>-0.0015</td>
</tr>
<tr>
<td><strong>Location and accessibility</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smart-card recharge point at 200 meters</td>
<td>-0.0782</td>
<td>-5.1</td>
<td>-0.0718</td>
</tr>
<tr>
<td>Subway station at 200 meters</td>
<td>-0.2768</td>
<td>-4.93</td>
<td>-0.3929</td>
</tr>
<tr>
<td>Frequency of bus-stop (veh/h)</td>
<td>-0.003</td>
<td>-4.67</td>
<td>-0.0031</td>
</tr>
<tr>
<td><strong>Operational</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupancy rate (%)</td>
<td>1.3834</td>
<td>17.31</td>
<td>0.8309</td>
</tr>
<tr>
<td>Corrected Frequency (Veh/h)</td>
<td>-0.0077</td>
<td>-2.48</td>
<td>0.0204</td>
</tr>
<tr>
<td>N of average doors</td>
<td>0.2444</td>
<td>6.84</td>
<td>0.0282</td>
</tr>
<tr>
<td>Percentage of transfer trips (%)</td>
<td>0.3511</td>
<td>4.96</td>
<td>0.249</td>
</tr>
<tr>
<td><strong>Operators</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UN1</td>
<td>0.4504</td>
<td>10.36</td>
<td>0.4875</td>
</tr>
<tr>
<td>UN3</td>
<td>0.4131</td>
<td>8.57</td>
<td>0.5567</td>
</tr>
<tr>
<td>UN4</td>
<td>0.2879</td>
<td>5.6</td>
<td>0.5548</td>
</tr>
<tr>
<td>UN5</td>
<td>0.2749</td>
<td>6.1</td>
<td>0.3789</td>
</tr>
<tr>
<td>UN6</td>
<td>0.3979</td>
<td>5.61</td>
<td>0.5571</td>
</tr>
<tr>
<td>UN7</td>
<td>0.6585</td>
<td>12.27</td>
<td>0.5949</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>-2.8576</td>
<td>-15.76</td>
<td>-2.1408</td>
</tr>
<tr>
<td>log-likelihood</td>
<td>-8891.94</td>
<td></td>
<td>-72116.87</td>
</tr>
</tbody>
</table>

### 3.2 Results

The explanatory capacity of the model variables is checked in the table (1), using the t test as a measure (for all cases, its value is higher than 1.645, reference value for a 90% confidence). It is important to clarify that, although there are variables that are not significant for a particular period, it was decided to keep them since they are very significant in the other periods, and maintaining the same variables in the three periods allows a better comparison between them, and allows to review how the effect produced by each of the variables changes according to the period of the day analyzed. Regarding the signs, it is found that:
- The communes with better economic conditions, education and health, have a lower IPS, for which the positive sign of the coefficient of this variable is considered correct, since it is expected that people with better economic conditions and better quality of life tend to evade less.

- The initial assumption is fulfilled that in areas with higher economic income the evasion rate is lower. The above is consistent with the negative sign of said variable.

- Having a smartcard top-up point nearby allows to reduce the evasion due to the lack of balances in the smartcard, which agrees with the negative sign that is observed for this coefficient. Additionally, it is found to have greater weight and significance in periods "morning peak" than in "evening peak", which is understandable since the situation of lack of balance is more limiting in the morning trip than in the afternoon, when having gone through recharging points throughout the day.

- Bus stops with a nearby subway station are more likely to come from the metro (transfer), where at least that part of users would have been forced to pay, which means that there is less evasion in said bus stop.

- Bus stops with lower "total frequency" tend to be less connected and usually in peripheral areas of the city, and it seems that this is related to greater evasion.

- A higher occupancy rate on buses makes it difficult for people to get close to the smartcard reader, which sometimes they evade even unintentionally. Additionally, it can be thought that there is less social pressure when evading, when this action is done in a group. For this reason, the positive sign obtained for this variable and its great significance is considered adequate.

- It was expected a priori that services with a higher frequency and a better Frequency Compliance Index (ICF), offer a better service, which increases the system compliance with the users and discourage them from evading the rate; however, it is striking that for the "off-peak" period the sign changes with respect to the other two periods. It may be due in part to the fact that during this period the metro system is not at capacity, so bus users who always pay the fare are changed, while users seeking to evade use the bus services that compete with them. metro, which tend to be those more frequently. This is only one possible reason for the observed behavior, which can not be confirmed from these results.

- On a bus with more doors, an evader has more possible access points, particularly in case when the rear doors open to allow passengers to alight at the bus stop. This is reflected in the positive sign obtained for this variable, and agrees with what is found in the literature.

- The coefficients of the variables of operators are difficult to analyze, since it is not possible to relate it to the socioeconomic characteristics of the areas where they operate, because the structure of the system, each operator runs many communes in simultaneous with other operators; in addition, the effect of the socioeconomic conditions of the commune is included in the variable of Social Priority Index (IPS). On the other hand, it is not possible to assume a relationship with the quality of the service offered
by the operator, since this effect is partly picked up by the variable Corrected Frequency, which includes the frequency compliance indicator (ICF) as a quality service offered. Finally, the analysis that can be given to these variables of operators is to reflect a bit the policies implemented by some or others to control evasion, such as the case of installation of turnstiles, inspection rates, and others; Unfortunately, there is not complete information, but it is one of the topics to continue investigating.

References


Lee J (2011) Uncovering san francisco, california, muni’s proof-of-payment patterns to help reduce fare evasion. Transportation Research Record: Journal of the Transportation Research Board (2216):75–84

