What is behind fare evasion? The case of Transantiago

Pablo Guarda, Pontificia Universidad Católica de Chile, paguarda@uc.cl
Patricia Galilea, Pontificia Universidad Católica de Chile, pga@ing.puc.cl
Laurel Paget-Seekins, Pontificia Universidad Católica de Chile, lpaget@ing.puc.cl
Juan de Dios Ortúzar, Pontificia Universidad Católica de Chile, jos@ing.puc.cl

ABSTRACT
Fare evasion is a problem in many public transport systems around the world and policies to reduce it are generally aimed at improving control and increasing fines. We use an econometric approach to attempt explaining the high levels of evasion in Santiago, Chile, and guide public policy formulation to reduce this problem. In particular, a negative binomial count regression model allowed us to find that fare evasion rates on buses increase as: (i) more people board (or alight) at a given bus door, (ii) more passengers board by a rear door, (iii) buses have higher occupancy levels (and more doors) and (iv) passenger experience longer headways. By controlling these variables (*ceteris paribus*), results indicate that evasion is greater during the afternoon and evening, but it is not clear that it is higher during peak hours. Regarding socioeconomic variables, we found that fare evasion at bus stops located in higher income areas (municipalities) is significantly lower than in more deprived areas. Finally, based on our results we identified five main methods to address evasion as alternatives to more dedicated fine enforcement or increased inspection; (i) increasing the bus fleet, (ii) improving the bus headway regularity, iii) implementing off-board payment stations, iv) changing the payment system on board and v) changing the bus design (number of doors or capacity). Our model provides a powerful tool to predict the reduction of fare evasion due to the implementation of some of these five operational strategies, and can be applied to other bus public transport systems.

Keywords: *Fare Evasion, Transantiago, Cheating, Unethical behaviour, Count Regression Models*

1. INTRODUCTION

Fare evasion is a problem in many public transport systems around the world. A survey from the *International Association of Public Transport* (IAPT) found that fare evasion averaged 4.2% across their sample of (primarily) bus routes in 31 systems and 18 countries (Bonfanti and Wagenknecht, 2010). Policies to reduce fare evasion are generally aimed at improving control systems and increasing fines for offenders (Killias et al., 2009). Likewise, many studies have focussed on the design of punishment strategies to tackle fare evasion (Barabino et al., 2013; Lee, 2011; Thorlacius et al., 2010). However, viewing fare evaders as rational actors who maximize utility by weighting the costs of buying a ticket with the costs of being caught without one (Boyd et al., 1989; Kooreman, 1993), does not consider the different social and contextual aspects in which fare evasion takes place.

In February 2007 Chile´s capital city inaugurated an integrated bus-Metro public transport system called Transantiago, which became widely unpopular due to its poor initial implementation (Muñoz et al., 2008). The bus services changed from passengers paying cash to bus-drivers in a fairly informal system, to a cashless system using smartcards (BIP cards) tapped on electronic fare collection machines located at the entrance of buses, but not handled by the driver. The old system, where many bus drivers owned their buses and drivers' wages were based partly on the fares collected, changed to buses operated by large companies overseen by a government authority with drivers paid a set wage and having much better working conditions.

Evasion was unmeasured in the previous system, but soon became a concern for the new system. When chaos originally ensued at the system's implementation (i.e. radical changes to the routes were not properly communicated, etc.) the Transport Minister decreed a not well-thought moratorium on payments until the system "performed well"; this has given arguments to many evaders until today as the system is still perceived to work less than well by many (EMOL, 2007). The rate of evasion in the system represents the ratio between the number of unpaid bus fares (evaders) and the total number of boarding passengers (DTPM, 2013a). The estimated monthly evasion rate on Transantiago buses during 2007 oscillated between 12% and 16%, whereas during 2012 it ranged between 20% and 27% (Figure 1). Determining the reasons for this increase over time is outside the scope of this paper (a cross-sectional study), but there are a number of potential factors such as increases in the nominal bus fares (Figure 1) or a *contagion effect* of the unethical behaviour (Gino et al., 2009) where potential fare evaders learn by observing others (Bucciol et al., 2013; Reddy et al., 2011).

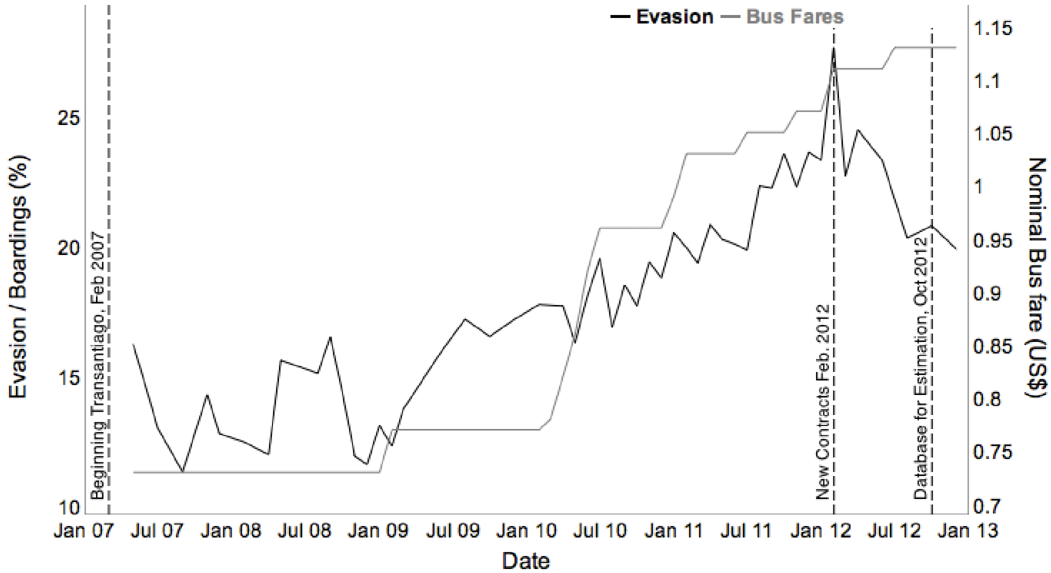

Figure 1. Evolution of the monthly rate of fare evasion and the nominal bus fares (DTPM, 2013a)

In response to this problem, the authorities have implemented various plans to reduce evasion, such as developing advertising campaigns, increasing the number of ticket inspectors, and adding more off-board payment stations. Since the highest rates of evasion and use of public transport are reported in low-income areas, public policies could be focused on increasing fine enforcement in these areas. However, limiting the solutions of a phenomenon as complex as evasion only to enforcement strategies might overestimate the benefits of some measures and create higher than necessary implementation costs. We believe that fare evasion levels are a combination of a number of factors including, the level of income, the perceptions of the service, cultural components, level of enforcement and the operation of the public transport system, among others (Torres-Montoya, 2014).

The main objective of this paper is to jointly analyse the impact of different factors behind the high levels of fare evasion in Santiago, with the purpose of guiding public policy aimed at reducing this problem. Using available cross sectional data from October 2012 (collected by the Chilean authorities) about the monthly mean rate of fare evasion on buses in Santiago, we explain the phenomenon using regression models, a methodology widely used to model statistical data both in transport research (Ortúzar and Willumsen, 2011) and in other areas of high relevance to public policy (Osgood, 2000). In particular, we estimated a multiple linear regression model and two types of count regression models, Poisson and negative binomial models. Subsequently, after trying different specifications and performing appropriate statistical tests, we found that the negative binomial regression model was the best approach.

In the existing literature some variables such as the period of the day (Lee, 2011), the operation of the system and the level of income of passengers have been discussed (Torres-Montoya, 2014), but not considered jointly to model evasion rates or to design public policy to tackle this problem. The main contribution of our paper is to analyse and quantify the joint influence of variables related to the level of income of the area where bus stops are located, period of the day, level of service (headways and overcrowding) and bus door operation (entrance door and number of doors) on fare evasion. Our econometric approach will hopefully provide decision-makers with a useful tool to predict the impact of some public policies to deal with evasion and also a better understanding of this complex issue.

The rest of the paper is organized as follows. Section 2 describes the information available, the procedures involved to collect the data, the variables used for modelling and the limitations of the data available. In addition, we describe the econometric method employed to explain evasion and compare the three types of regression models used. In Section 3 we discuss our results and the variables that may affect evasion, select the best model and analyse the impact of changes in the independent variables on fare evasion. In Section 4 we discuss public policies to tackle evasion, give some further recommendations and suggest avenues for further research.

2. METHODOLOGY

2.1. Data Description

In Transantiago, bus services are operated by a group of private firms overseen by the government authority. Each operating company manages several bus routes, which are identified by a number or number-letter combination, and serve a set of stops. In the Metro system component fare evasion is not a problem, but since the beginning of Transantiago the bus system component has faced high levels of fare evasion which are also increasing over the time (DTPM, 2014).

The Enforcement Commission of the Chilean Transport Ministry (MTT) collects data about evasion on a sample of bus routes of each private operating company every month using plain-clothes observers. Within the sample of bus routes a sample of runs (bus trips) are selected. Observers are stationed at each door of the sampled bus and register the number of people boarding and alighting at each bus stop, by each door of the bus, as well as other conditions such as bus occupancy. They also register the number of evasions, but do not interact with the passengers in an enforcement capacity.

In the database supplied by MTT, each observation was measured at a specific door and bus stop of each bus route sampled. With the variables supplied by the MTT we built evasion models using data from October 2012, which we chose as a representative month of the public transport system operation (without any major holidays in Chile). In addition, the database for model estimation includes only observations taken on Tuesdays, Wednesday, and Thursdays; we had no data available for the rest of the week but we have no reasons to suspect these might be unduly different (Section 2.5).

2.2. Modelling Procedure

Multiple Linear Regression (MLR) is one of the most used methods for modelling statistical data (Greene, 2007). However, when the conditional probability density distribution (PDF) of the dependent variable is closer to a counting process, the normality and homoscedasticity assumptions of MLR models is often violated, which invalidates the statistical inference and suggests the application of other kinds of regression models (Hilbe, 2011). In this study, the assumption of error normality was statistically tested using the Shapiro Wilk and the Lilliefors tests; the assumption of homoscedasticity in the error distribution was tested using the Breusch-Pagan test. These three tests are widely used in econometrics to validate these key assumptions of MLR models (Greene, 2007).

There is a wide range of regression models for counting data in the literature and each one requires a different set of assumptions. Count regression models can be estimated as generalized linear models (GLM) employing maximum-likelihood-estimation (MLE) and not using ordinary least squares (OLS) as MLR models (Hilbe, 2011). GLM assumes a PDF for the dependent variable coming from the set of exponential families (Normal, Binomial, Poisson, among others) and also defines a link function that allows setting a mathematical relationship between the mean of the distribution and a set of linear predictors (using the inverse link).

In particular, Poisson regression, based on the Poisson PDF (1), is the basic method used for modelling count response data (Hilbe, 2011). The canonical form of the Poisson Model is given by the following equations:

$$P(Y_i = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \tag{1}$$

$$\ln(\lambda_i) = \sum_{k=0}^{K} \beta_k x_{i,k} \tag{2}$$

Equation (1) indicates the Poisson PDF of the observed outcome of counts (y_i) for observation *i*. Equation (2), known as the Poisson *link function*, is a regression equation relating the natural logarithm of the mean (λ_i) or expected value of counts for observed value *i* in the Poisson Distribution, to the product between a set of linear predictors (x_k) and their estimated parameters (β_k).

Poisson models assume that one event has no effect on the likelihood of observing additional events in the same period. However, this assumption is violated when the data presents *overdispersion*, which occurs in Poisson models when the conditional variance of the dependent variable is greater than its conditional mean (Hilbe, 2011); in some cases, its source is linked to a dependence of the probability of occurrence between events. In Poisson regression model, this phenomenon is known as *positive contagion* (a type of *true contagion*) because the probability of an event increases with the occurrence of previous event (Cameron and Trivedi, 1998). In other cases, the increase in the conditional variance of the Poisson distribution can be explained merely by a high heterogeneity of the count response, which is known as *spurious contagion* or *apparent dependence* (Cameron and Trivedi, 1998). Whichever is the source of overdispersion, it may cause problems with the statistical inference in Poisson regression models due to deflated standard errors of the estimates, i.e. a variable may appear to be a significant predictor when it is in fact not significant (Hilbe, 2011).

The standard negative binomial regression model is one of the most widely used methods to deal with overdispersed data, especially when its source is unknown (Hilbe, 2011). Even though, there are many ways to solve this problem, different studies have shown that the negative binomial model for most purposes is satisfactory (Cameron and Trivedi, 1998; Hilbe, 2011). However, the canonical form of this model (NB-C) uses a logit link, which does not allow to model Poisson overdispersion (Hilbe, 2011). For this reason an alternative specification[1] must be estimated (NB2 model), which uses the PDF of the negative binomial family (3), but has a log link function (4) as the Poisson Regression model (2).

$$P\ (Y_i = y_i \mid \mu_i, \alpha) = \begin{pmatrix} y_i + \frac{1}{\alpha} - 1 \\ \frac{1}{\alpha} - 1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad (3)$$

$$\ln(\mu_i) = \sum_{k=0}^{K} \beta_k x_{ik} \quad (4)$$

Equation (3) represents the NB2 PDF, conditioned both on the mean (μ_i) of the observed outcome of counts (y_i) for observation i and on the *heterogeneity parameter* (α). Equation (4) shows the log-link function of the NB2 regression model, where $x_{i,k}$ is the value of the predictor for the observation i and β_k the estimated parameter of the k-th explanatory variable. The *heterogeneity parameter* (α) is a measure of Poisson overdispersion, or extra correlation, and allows to account for overdispersion using the NB2 model (Hilbe, 2011). When the heterogeneity parameter is zero, the NB2 model is equal to a Poisson regression model, and higher values of overdispersion are reflected in a greater value of α in NB2 model results (Hilbe, 2011).

To select the most appropriate count regression model between Poisson and NB2, three tests are suggested in the literature (Hilbe, 2011): i) the Wald Test to test the significance the heterogeneity parameter α; ii) the Likelihood Ratio (LR) test to compare the statistical adjustment of the two specifications; and iii) the Pearson Dispersion Statistic of the Poisson Model (i.e., if it is greater than one, it indicates overdispersion). The Wald test indicates that the sample is Poisson distributed when its null hypothesis is not rejected ($\alpha = 0$). However, as the GLM estimation of the NB2 model assumes that α is a parameter that only takes positive values, it is necessary to perform a one-tailed t-test to assess its statistical significance (Cameron and Trivedi, 1998). Furthermore, as the Poisson model is equal to NB2 when α is zero, the two models are nested and can be compared using the traditional boundary LR test. In this case, the LR test is calculated with one degree of freedom, since model NB2 only estimates an extra parameter (α). In contrast to the first two tests, the Pearson Dispersion Statistic is only an indicator of overdispersion, since there is no a statistical criteria for evaluating how far over one it must be to ensure the presence of overdispersion (Hilbe, 2011).

We also tested for the presence of multicollinearity between the explanatory variables in all three models. Although multicollinearity does not generate systematic bias in estimation it may lead to lack of precision in the estimates, (Greene, 2007). Since non-experimental data will never be orthogonal some degree of multicollinearity is always expected. The literature presents different indicators to detect this problem, among which one of the most reliable is the variance inflation factor (VIF). The VIF indicator is calculated for each explanatory variable included in the regression model and essentially gives the variance explained by each predictor as a function of the variance of the remaining ones; its lower upper bound is one and

[1] By "model specification" we refer to both the functional form defined for the dependent variable and the set of explanatory variables included in the model.

values in excess of ten are a clear indication of a multicollinearity problem. However, since there is no consensus about the critical value of the VIF indicator, we decided to use four, which the less optimistic value found in the literature (O'Brien, 2007). Although VIF indicators are typically used to detect multicollinearity in MLR models, the conclusions of this test can be also used in count regression models. Since only the predictors are involved in the calculation of these indicators (not the outcome variable), if the specification defined in the MLR model is equal to the specification of the link function used in the Poisson Regression or NB2 models, the conclusions obtained with the MLR model can be applied directly to the other two models.

2.3. Dependent variable

The data available includes the number of people boarding (distinguishing evasions and paid fares) and alighting through each door of the bus, at each stop used by a bus along its route for the sampled day. To use this information in the least aggregate way possible, in the Poisson and NB2 count regression models the dependent variable was defined as the amount of evasion at a given door of the bus, during a particular time of the day and at a given bus stop. On the other hand, the literature suggests including the logarithm of the rate of events for MLR models with count data as dependent variable; in addition, because the log of zero is not defined, a standard solution is to add an arbitrary constant term (k), and to model $\ln(y+k)$ by OLS, where y corresponds to the count dependent variable (Cameron and Trivedi, 1998). For this reason, in the MLR case, the dependent variable was defined as the logarithm of the sum of one and the rate of evasion (calculated as the ratio between number of evaders at a given door and the number of passengers boarding at that door).

2.4. Explanatory Variables

2.4.1. Bus door operation

The number of evaders naturally grows with the number of passengers boarding. As more people board at a given door, evasion has more opportunity to occur at that door. In the count regression models' literature this effect is known as *exposure effect*, and to control it, the natural logarithm of the *exposure variable* is included as an explanatory variable (Hilbe, 2011). It is easy to note that both in the Poisson and NB2 models, the exposure variable corresponds to the number of passengers boarding at a given door. To capture the *exposure effect*, we defined the variable ***Log(Boardings)*** as the logarithm of the number of evaders boarding at the door where the observations was collected. In contrast, in the MLR model, the number of passengers boarding was not included as an explanatory variable, since it was used to define the dependent variable (the logarithm of the sum of one and the rate of evasion).

The bus operating companies instruct their bus drivers to regulate the bus door operation. Even though passengers are only supposed to board through the front door and leave the bus by the back doors, the data collected shows that this does not always happen. To capture the effect of the number of people leaving the bus on the number of fare evaders at a given door, initially we added a variable defined as the number of passengers exiting at that door. However, the variable was only statistically significant at the 95% level when included as a dummy **(*Exiting*)** that took the value one when someone exited from that door and zero when nobody left the bus through that door.

There are multiple types and sizes of buses operating in Transantiago. They range from large articulated buses with four doors to small buses with only two doors. We defined the variable ***Doors*** as the number of doors of the bus. In addition, to capture the effect of the type of bus door on the number of fare evaders counted at that door (not explained by the other variables included in the model), we defined the variable ***Front Door***, which takes the value one if the observation was taken in the front door, and zero otherwise (second, third or fourth door).

2.4.2. Temporal variables

The data used to estimate the regression models was collected at different time periods of the day during October 2012. Transantiago defines 12 time periods for weekdays (DTPM, 2013b). The raw data contained only information collected between 6:30 and 22:30 corresponding to seven time periods, but we decided to not consider observations taken after 21:30 (only six observations). Finally, we regrouped the time periods into four blocks: Morning Peak (6:30 - 8:29), Morning Off-Peak (8:30 - 12:29), Afternoon[2] (12:30 - 20:29) and Evening (20:30 - 21:29), and defined a dummy for the last three periods (***Morning Off-Peak***, ***Afternoon*** and ***Evening***), taking the morning peak period as reference. Note that the time periods were regrouped based on the statistical significance of the estimated parameters and the models' goodness of fit.

To capture the potential effects of the day of the week, we created the dummy variables ***Wednesday*** and ***Thursday***, which take the value one if the observation was taken on Wednesday or Thursday, respectively, and zero otherwise (leaving Tuesday as the reference). Since the data only covered a month, we could not include variables to capture potential differences on fare evasion among the months and years of Transantiago's operation (not explained by the other variables included in the models).

2.4.3. Quality of service

The higher evasion rates observed on more crowded buses suggest that reducing overcrowding could reduce evasion. Our dataset includes an occupancy index with five categories (A, B, C, D and E), depending on the proportion of people sitting or standing in the bus. It should be noted that this classification is not perfect and it is subject to the perception of the plain-clothes observers. To model the effect of bus occupancy, we defined the dummy variable ***Low Occupancy***, which takes the value one if the occupancy index is C, D or E and zero otherwise (A, B). Buses are assumed to have low occupancy if less than a half of the total bus floor area has people standing regardless of how many seats are available. Instead, if more than a half the total bus floor has people standing, it is assumed that buses have high occupancy. The aggregation of the bus occupancy index used to create the dummy ***Low Occupancy*** was defined on the basis of the statistical significance of the estimated parameters and the goodness of fit of the regression models.

Buses in Santiago can be so crowded that passengers have to wait for the next bus in order to squeeze in. Average waiting times are closely related to the bus frequency of the operating companies, and if high they produce a bad perception of the service and could increase fare

---

[2] The dummy variable ***Afternoon*** added in our models includes the midday peak (12:30-13:59), afternoon peak (14:00-17:29) and afternoon off-peak (17:30-20:29) periods defined by Transantiago (DTPM, 2013b).

evasion (Torres-Montoya, 2014). However, information about the bus frequency at the time our data was collected was not available in the database. To capture the effect of the perceived bus frequency on fare evasion, we created a proxy variable using the operational program of Transantiago for the second semester of 2012 (DTPM, 2013b). This program defines, for each time period (using the classification of day periods mentioned in Section 2.4.2), the average frequency (buses/hour), capacity (passengers/hour) and average speed (among others) that bus operators must comply for their bus routes in the system. Depending on the mean and variability (e.g. regularity of the frequency) of these operational indicators, bus operators may receive financial penalties from Transantiago. Since the level of compliance of the frequency goal is over 90% for all companies (DTPM, 2013a), this information is a good proxy of the actual frequency experienced by users in the network. Now, as the frequency defined in the operational program of Transantiago varies with the direction of the bus route (i.e. outward or return) and the time period classification in Transantiago, we assigned a frequency value to each observation in the database, depending on the time when the observation was taken and the direction of the bus route.

Finally, to capture the effect of the bus frequency in the regression models, two alternative specifications were tested: (i) adding the frequency as linear variable, i.e. ***Frequency*** (buses/hour) and (ii) adding its reciprocal, i.e. ***Headway*** (min) as a linear variable. Note that the headway has a theoretical meaning in transport since it is directly related to the waiting times perceived by users (Ortúzar and Willumsen, 2011), and these may partly explain fare evasion. Interestingly, the statistical significance and goodness of fit in all models were far superior when using the variable ***Headway***.

2.4.4. Bus route operation

Transantiago is divided into 10 zones (A-J) each covering different municipalities. The system's network was designed as a trunk and feeder scheme, where Metro is a key part of the trunk network. The larger buses also operate on major trunk corridors and can serve multiple zones. Feeder buses, on the other hand, operate in neighbourhoods serving the trunk bus routes or Metro lines, and are restricted to a single zone (Muñoz et al., 2008). To evaluate the effect of the type of bus route on fare evasion, we created the dummy variable dummy ***Feeder***, which takes the value one if the observation was collected on a feeder bus route, and zero in a trunk route.

Since June 2012, bus routes in Transantiago have been managed by seven private operators; Alsacia, Subus, Vule, Express, MetBus, RedBus and STP. Initially, we created six dummy variables to control for possible average differences on fare evasion among the seven operators (not captured by the other explanatory variables included in the models). Subsequently, based on the statistical significance of the estimated parameters and the goodness of fit of the models under different groupings of bus operators (as discussed in Section 3), we defined four groups of operators. The first includes Alsacia and Vule, the second Express and MetBus, the third Subus and RedBus and the fourth only STP. Three dummy variables were then created: ***Operators I***, ***Operators II*** and ***Operators III***, which take the value one if the observation was taken in the first, second or third group of bus operators respectively, and zero otherwise (the operator STP was taken as reference).

2.4.5. Socioeconomic level (SEL)

The data collected by the MTT did not include socioeconomic information about the SEL of individuals boarding the bus at a given bus stop or the exact location where each observation was taken. For this reason, we developed an algorithm that inferred the location of the bus stop for each observation. Then, to capture potential income effects in our data, we created a socioeconomic proxy variable based on the average income of the location estimated by the algorithm for each observation in the database.

To infer the location the algorithm used information of a public databank corresponding to the second semester of 2012 (DTPM, 2013b) listing all bus routes of Transantiago and the characteristics of the set of bus stops on their routes (municipality, address, etc.). In the first step, the algorithm constructs a vector for each bus route with the proportion of bus stops that were located in each municipality (which are quite small in area in the Santiago metropolitan region). In the second step, each vector is ordered sequentially, based on the direction of the bus route (outbound or inbound) and which municipality is visited first. In the third step, the algorithm proportionally replicates the order in which the municipalities are visited by each bus route[3], among the list of observations collected by each plain-clothes observer at a given bus run (which was also ordered sequentially based on the time when each measurement was taken). For example, if the first 50% of the bus stops in a route are located in municipality A, it is assumed that the first half of the observations taken by each plain clothes observer at a given bus run corresponds to this municipality. Finally, the algorithm was applied to the whole set of observations of each bus route, yielding the municipality where each observation was taken.

To capture the effect of the SEL in the regression models, we divided the municipalities into four clusters using exogenous information about the average monthly household income in each municipality (BCN, 2013). To define the income thresholds in each cluster, we used income household data by quintile available in the National Survey of Economic Characterization (CASEN) conducted in 2009 (MDS, 2009). According to this data, the average monthly household incomes by quintile in Santiago are US$330, US$683, US$1065, US$1674 and US$5135[4], respectively. Since the average monthly household income of each municipality was always greater than the first two quintiles, we created four groups initially: lower than US$1065; between US$1065 and US$1674; between US$1674 and US$5135, and higher than US$5135. Then, we defined three dummies as proxies for income: ***Lower Middle Income***, ***Upper Middle Income*** and ***High*** **Income**, which take the value one, if the municipality where the observations were collected belonged to the second, third or fourth group respectively, and zero otherwise (the first group of municipalities was used as reference).

The difference between the average socioeconomic characteristics of passengers who travel in each municipality could explain the significant difference in the average level of fare evasion among clusters. Due to the relatively high cost of the bus fare in Transantiago compared to the average household income of low-income groups, we would expect a significant increase on fare evasion in these areas (Section 4.2.6); note that Santiago is characterised by strong spatial segregation by income (Sabatini et al., 2009) and there is high correlation between the SEL of the municipality and that of the neighbours living in that municipality. However,

[3] As stated in Section 2.1, the plain- clothes observers only register observations at those bus stop where the bus actually stops, thus, it is not possible to link sequentially the measurements with the list of bus stops served by a given bus route in the actual network.
[4] Using the average US dollar exchange rate in 2009, 1 US$ was equal to 559.7 CLP (BCC, 2015)

passengers boarding at a given bus stop do not necessarily live in that municipality, except if they board during the morning peak when it is more likely that bus riders live in the bus stop's influence area. Therefore, the proxy variables for income during the morning peak hour should capture more accurately the effect of the SEL of individuals on fare evasion. In Section 3.3, we will test statistically our hypothesis, by estimating a new regression model using only data collected during the morning peak and analysing the change in the SEL parameter.

2.5. Descriptive Statistics

The raw data contained 22,259 observations taken on four days of October 2012: Thursday October 18th, Tuesday October 23th, Wednesday October 24th, and Thursday October 25th. In total, 311 buses were monitored, equivalent to 4.9% of all buses operating in 2012 (DTPM, 2013a); these covered 40 different bus routes (10.7% of the overall services), the seven bus operating companies, and the ten zones of the system. During these four working days, 23,670 passengers were observed boarding Transantiago's buses and 18,726 were observed paying the bus fare; these figures may be compared with the 3,184,289 bus boardings recorded by Transantiago's fare system on an average working day in 2012 (DTPM, 2013a).

For the data analysis, we used a clean dataset removing all inconsistencies detected, yielding 21,244 observations; we eliminated duplicate observations (570), observations with missing data (176) and observations with other basic inconsistencies, such as the number of evaders being greater than the number of people boarding (20) and where the reported number of the door in which the observation was taken was greater than the number of doors in the bus (37). In addition, we did not include six observations taken after 9:30 PM (as discussed in Section 2.4.2) and 206 observations corresponding to off-board payment stations[5].

Table 1 presents summary statistics for the categorical variables available in the database and the variables created using our algorithms. The columns of the table show the proportion of total observations, the proportion of total boardings, the proportion of total evasions and the average sample rate of evasion at a given door and bus stop, for the different levels of each categorical variable (calculated as the ratio between the total number of unpaid bus fares and the total number of passengers boarding).

Table 2 shows the mean, standard deviation and coefficient of variation for the non-categorical variables, based on our unit of analysis[6].

[5] Plain-clothes observers collect observations sitting on the bus and are not able to verify if passengers pay the fare using the off-board payment machines. Even though they are instructed to register no fare evasion at these bus stops, fare evasion could still exist and this would distort the results obtained with the models.
[6] The unit of analysis corresponds to each measurement taken by a plain-clothes observer, who registers the number of people evading/boarding/exiting at a given door, bus stop and given time of the day.

Table 1: Summary statistics for the categorical variables, October 2012 (N = 21,244)

Variable	Proportion of sample (%)	Proportion of total boardings (%)	Proportion of total evasions (%)	Evasion rate (%)
Boarding door				
Front door	37.2	96.7	87.9	18.4
Back door 2	33.1	1.4	5.4	76.0
Back door 3	24.8	1.3	4.9	74.3
Back door 4	5.0	0.5	1.8	72.2
Bus doors				
2	21.3	23.1	19.4	17.0
3	59.3	55.3	56.4	20.6
4	19.5	21.7	24.2	22.6
Day of the week				
Tuesday	26.0	26.2	23.1	17.8
Wednesday	23.1	25.4	26.0	20.7
Thursday	51.0	48.4	51.0	21.3
Period of the day				
Morning peak (6:30-8:29)	13.5	18.8	20.9	22.5
Morning off-peak (8:30-12:29)	32.3	26.7	20.4	15.4
Noon peak (12:30-13:59)	6.5	4.6	4.5	20.0
Afternoon off-peak (14:00-17:29)	24.9	26.7	29.8	22.5
Afternoon peak (17:30-20:29)	20.9	22.0	22.9	21.1
Evening (20:30-21:29)	1.9	1.2	1.4	21.7
Bus Occupancy[a]				
A (low)	53.1	37.4	34.3	18.5
B (low)	19.5	20.7	21.0	20.0
C (low)	15.8	21.0	18.1	17.4
D (high)	8.8	15.4	17.8	23.5
E (high)	2.8	5.5	9.3	34.2
Type of bus route				
Trunk	53.6	54.3	56.8	21.1
Feeder	46.4	45.7	43.2	19.1
Bus operator				
Alsacia	15.7	17.0	21.5	25.5
Subus	21.1	19.9	16.1	16.3
Vule	13.5	13.4	14.7	22.2
Express	13.3	13.3	13.2	20.0
Metbus	12.9	13.3	11.2	17.0
Redbus	10.2	12.7	8.8	13.9
STP	13.3	10.4	14.5	28.6
Socioeconomic level (SEL)				
Low (< US$ 1065)	37.4	35.0	42.1	24.3
Lower middle (US$ 1065-1674)	36.0	36.0	34.5	19.4
Upper middle (US$ 1674-5175)	17.2	18.9	17.4	18.6
High (> US$ 5175)	9.4	10.1	6.0	12.1

[a] The indices A, B, C represent low occupancy, whereas D, E correspond to high occupancy (Section 2.4.3)

Table 2: Summary statistics for non-categorical variables, October 2012 (N=21,244)

Variable	Mean	Standard deviation	Coefficient of variation
Evasion (passengers)	0.207	0.901	4.352
Boarding (passengers)	1.025	3.088	3.013
Exiting (passengers)	1.050	2.448	2.331
Doors	2.982	0.638	4.674
Frequency (buses/hour)	9.263	4.377	0.473
Headway (minutes)	7.802	2.960	0.379

3. RESULTS AND DISCUSSION

In this section we will first present the modelling results for MLR, Poisson, and NB2 regression models. Second, we will test their assumptions and select the best specification. To do this we will try different sets of explanatory variables and functional forms of the dependent variable. After trying different specifications and validating the statistical model assumptions we selected the NB2 model. Third, we will test the impact of income effects on fare evasion by formulating an additional model specification based on the NB2 regression model. Finally, we will perform a sensitivity analysis with this model.

3.1. Model estimation

The final dataset used for estimating the three regression models (5,334 observations) does not include those observations corresponding to the case when nobody boarded through one door of the bus (15,910). As we explained in Section 2.1, each observation represents the count of how many passengers boarded/exited/evaded at one door of the bus at a certain bus stop. Therefore, some observations may record no boarding, but they still may record the number of people leaving the bus. However, in these cases the count of evasion is always zero, and thus, these observations do not add useful information for model estimation.

The sets of explanatory variables included in each type of regression model were defined on the basis of the following criteria: (i) the consistency of the signs of the estimated parameters; (ii) the parameters' statistical significance (at the 95% level) and (iii) the goodness of fit of the models (Ortúzar and Willumsen, 2011). Finally, the models consider the following common set of predictors to explain fare evasion: number of passengers boarding[7] and alighting at a door, number of doors, entrance door, bus occupancy, bus route headway, bus operator and time of day. Table 3 shows the estimation results of the MLR, Poisson and NB2 regression models using the set of explanatory variables and dependent variables described previously.

[7] Except in the RLM model since the number of boardings was used to define the dependent variable.

Table 3: MLR, Poisson and NB2 regression model estimation results (N = 5,334)

Variable (t-test)	MLR	Poisson	NB2	VIF
Bus door operation				
Log (boardings)	-	0.976 (59.3)	1.079 (42.7)	1.105
Exiting	0.044 (3.8)	0.292 (5.5)	0.221 (2.8)	1.384
Bus doors	0.032 (6.3)	0.178 (6.6)	0.210 (5.6)	1.385
Front door	-0.391 (-24.0)	-1.105 (-18.3)	-1.336 (-14.0)	1.520
Time period				
Morning off peak (08:30-12:30)	-0.022 (-2.3)	-0.231 (-4.6)	-0.166 (-2.4)	1.978
Afternoon (12:30-20:30)	0.026 (2.9)	0.156 (3.8)	0.231 (3.9)	1.958
Evening (20:30-22:00)	0.064 (2.7)	0.343 (2.6)	0.499 (2.8)	1.102
Quality of service				
Low occupancy	-0.043 (-4.8)	-0.137 (-3.6)	-0.150 (-2.6)	1.185
Headway	0.004 (3.0)	0.016 (2.6)	0.026 (2.9)	1.572
Bus route operation				
Operators I	-0.043 (-4.4)	-0.265 (-5.5)	-0.271 (-4.0)	2.363
Operators II	-0.075 (-7.4)	-0.528 (-10.2)	-0.567 (-7.8)	2.261
Operators III	-0.083 (-8.2)	-0.613 (-11.6)	-0.661 (-8.9)	2.395
Constant	0.485 (16.8)	-0.748 (-5.4)	-0.871 (-4.4)	-
Adjusted R^2	0.176	-	-	-
Log-likelihood	-	-5,831.7	-5,477.8	-
Akaike Information Criteria (AIC)	-	11,689.4	10,981.5	-

Note that in these three regression models, the dummy variables defined to capture the effect of the day of the week (***Wednesday***, ***Thursday***) and of the type of bus route (***Feeder***) were not significant at the 95% level. The low statistical significance of these variables could indicate that the remaining variables in the model are capturing the average differences in fare evasion observed both between Tuesday, Wednesday and Thursday, and between trunk and feeder bus routes (Table 1).

3.2. Testing model assumptions

To validate the theoretical assumptions underpinning MLR, Poisson Regression and NB2 models, the literature suggests different statistical tests (Section 2.5). All these tests were performed (at the 99% confidence level) using the model specifications shown in Table 3. In addition, we compared graphically the fit of the Poisson and NB2 regression models with the distribution of the observed data.

Regarding the MLR model, two statistical tests were performed (Shapiro Wilk and Lilliefors) to validate the assumption of error normality, and one (Breusch-Pagan) to test the homoscedasticity assumption. The results of these tests allow to reject both null hypotheses (Table 4). Subsequently, we tested the assumption of independence of random variables in the Poisson regression model. Since the presence of overdispersion violates the Poisson independence assumption, overdispersion was first tested; the results of the Pearson Dispersion Statistic, Wald Test and LR test indicate the presence of overdispersion in the data (Table 4), which violates the Poisson Independence assumption.

Table 4: Statistical diagnostic tests to check regression model assumptions (N = 5,334)

Test	Value	Table Value
Shapiro-Wilk[8]	0.00[a] (W=0.85)	0.01
Lilliefors	0.23[b]	0.01
Breusch-Pagan	197.18 (dof=11)	26.22
Pearson Dispersion Statistic	1.31	1.00
Wald Test ($\alpha = 0.660$)	14.64 (dof=5,332)	1.65
LR test for overdispersion	709.88 (dof=1)	3.84

[a] P-values lower than the chosen significant level chosen (1%) reject the null hypothesis of normality.
[b] Abdi et al. (2009) present in detail the methodology to calculate the Lilliefors statistic in large samples

Figure 2 shows the percentage of the total sample that represents each possible value of the amount of evasion at a door on the y-axis. The black solid line represents the observed data and the grey dashed and grey solid lines are obtained using the predictions of the Poisson and NB2 regression models, respectively (Table 3).

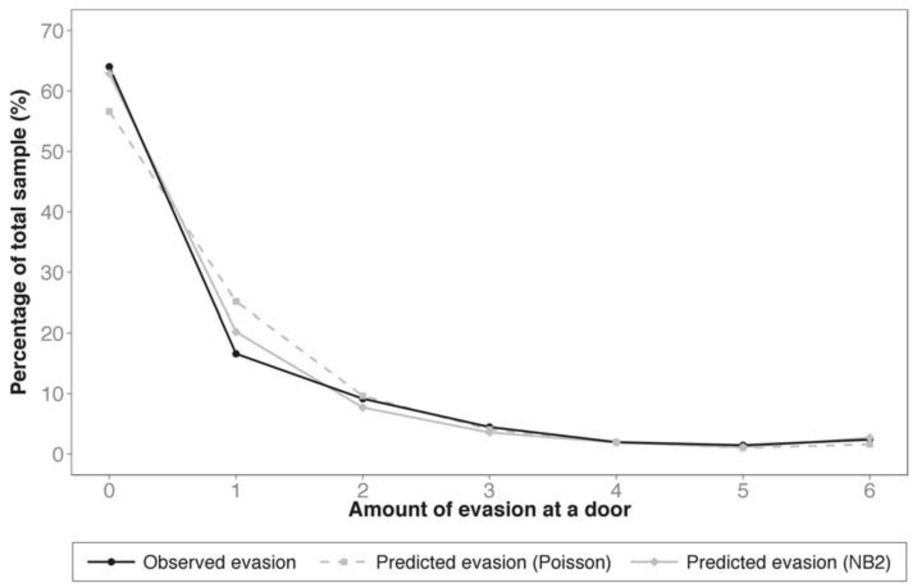

Figure 2. Observed vs. predicted amount of evasion at a bus door of NB2 and Poisson models (October 2012).

First, note that the distribution of the data has the typical form of a counting process distribution; also, the NB2 model shows a better fit overall than the Poisson model. Even though, the Poisson regression model predictions are marginally closer to the observed data when the amount of evasion is two and three units, the Poisson model presents a large bias in predicting for zero and one counts (corresponding to more than the 80% of the total observations).

Finally, we tested the presence of multicollinearity in the NB2 model. Based on the discussion in Section 2.2 the degree of multicollinearity of the NB2 model can be tested on a

[8] Royston (1995) re-formulated the W-test statistics including an algorithm for obtaining the P-value in large samples, which is implemented in R Studio. Since the R Package accepts a maximum of 5,000 observations, we selected the first 5,000 observations of our dataset to perform the Shapiro-Wilk test.

MLR model using the same specification defined in the link function of the NB2 model. Since the VIF indicators of each explanatory variable are lower than four (Table 3), these results indicate absence of multicollinearity between the set of variables included in the NB2 model. Note that the VIF indicators obtained for all explanatory variables included in the NB2 model are closer to the lower bound of the VIF indicator (one) than the critical value (four).

In conclusion, the violation of error normality in the MLR model, the presence of overdispersion in the Poisson regression model, and the absence of multicollinearity in the NB2 model (and its better fit), support the application of the NB2 regression model.

3.3. NB2 model adding income effects

The baseline Negative Binomial Regression model (NB2) does not include any proxy variable to capture income effects (Table 3). To explain the average differences in the level of evasion observed between the municipalities where each observation was taken (*ceteris paribus*), an additional NB2 model specification was estimated, which we call NB2-I. Since we expected that these differences would be mainly explained by the SEL corresponding to each location, model NB2-I included dummy variables based on the level of income of the municipality (Section 2.4.5).

Table 5 shows the NB2-I estimation results, and the semi-elasticities (ε) of the NB2-I model (will be analysed in Section 3.4). Results show that most parameters are statistically significant at the 95% confidence level, except for the variable ***Headway***, which is significant at the 94% confidence level. Once again, the variables designed to capture the effects of the type of bus route (***Feeder***) and of the day of the week (***Wednesday***, ***Thursday***) were not significant (Section 3.1). Regarding the variables to control for SEL, only the variables ***Upper-Middle Income*** and ***High Income*** were significant at the 95% confidence level, whereas the variable ***Lower-Middle Income*** was significant only at the 80% confidence level. Thus, the base dummy variable defined in the NB2-I model includes observations taken both in low income and lower-middle income municipalities.

The table also show that the goodness of fit of the NB2-I model is significantly greater than that of the NB2 regression model. Since the difference between the Akaike information criteria (AIC) of NB2 and NB2-I is greater than 10 ($\Delta\text{AIC} = 23.37$), regardless the number of observations used to estimate the models, NB2-I is preferred (Hilbe, 2011). Note that AIC is one of the most commonly used fit statistics for GLM models, and a lower value of the indicator, suggest a better-fitted model. Likewise, a log-likelihood Ratio (LR) test on NB2-I vs. NB2 with two degrees of freedom (two additional dummy variables for income) confirms clearly that NB2-I has the highest explanatory power (27.374> $\chi^2_{2,99\%}$) at the 99% level.

Table 5: Negative binomial regression model estimation results

Variable (t-test)	NB2-I	ε^a(%)
Bus door operation		
Log (boardings)	1.079 (42.8)	-
Exiting	0.215 (2.7)	23.985
Bus doors	0.186 (4.9)	20.431
Front door	-1.344 (-14.1)	-73.911
Time period		
Morning off peak (08:30-12:30)	-0.140 (-2.0)	-13.067
Afternoon (12:30-20:30)	0.253 (4.3)	28.783

Evening (20:30-22:00)	0.547 (3.1)	72.877
Quality of service		
Low occupancy	-0.174 (-3.0)	-16.008
Headway (minutes)	0.018 (1.9)	1.775
Bus route operation		
Operators I	-0.228 (-3.3)	-20.349
Operators II	-0.445 (-5.9)	-35.909
Operators III	-0.608 (-8.1)	-45.582
Socioeconomic level (SEL)		
Upper Middle Income (US$ 1674-5175)	-0.233 (-3.9)	-20.813
High Income (>US$ 5175)	-0.366 (-4.1)	-30.663
Constant	-0.722 (-3.6)	-
Observations	5,534	-
Log-likelihood	-5,463.1	-
Akaike Information Criteria (AIC)	10,958.2	-

[a] Percentage change in the expected number of evasions to a unit change in an independent variable (NB2-I model)

Finally, we conclude that the stratification using the monthly average household income by municipality is a good criteria to define a set of proxy variables to capture the effect of income on fare evasion. The inclusion of these variables in the NB2-I model increases significantly the explanatory power of the baseline model (NB2). In addition, the estimated parameters of the income variables (***Upper-Middle Income*** and ***High Income***) in the NB2-I model present correct signs, high statistical significance, and magnitudes consistent with our expectations (as discussed in Section 3.4). Likewise, the estimated parameters associated with the rest of the explanatory variables included in the NB2 model are still highly significant and present signs consistent with our expectations.

In the following section, we perform a sensitivity analysis to argue the relevance of the variables included in the selected model (NB2-I) and measure the relative impact of changes of these variables on fare evasion.

3.4. Sensitivity Analysis

To study the impact of the explanatory variables on the dependent variable in the NB2-I model, we calculated the Incidence Rate Ratio (IRR) for each variable; this is an indicator widely used in statistics to interpret the results obtained both in Poisson and NB2 models (Hilbe, 2011). The IRR factor represents how many times the dependent variable increases or decreases with respect to its expected value, when one independent variable increases in one unit (Hilbe, 2011). These changes in the dependent variable are calculated *ceteris paribus* (i.e. analysing the impact of one variable while all others are held constant). Then, considering that the dependent variable in the link function of the Poisson and NB2 models is a logarithm (2), as long as the independent variable has a linear form, its IRR is obtained by the exponential of its parameter in the regression.

To simplify the interpretation of results we calculated the semi-elasticity (ε) of each explanatory variable included in the model as $\varepsilon_k = 100 \cdot (\mathrm{IRR}_k - 1)$, for any variable k with a linear form; this represents the percentage change in the expected number of counts (i.e. number of evasions at a given door and bus stop) to a unit change in the variable k. Now, since the exposure variable has a log form, its estimated parameter gives directly the percentage elasticity of this variable[9] (i.e. percentage change in the dependent variable to a unit percentage change in the exposure variable). Consequently, if the estimated parameter is exactly one, regardless of the exposure variable's value, the rate of occurrence of the dependent variable remains constant. Instead, if the parameter is greater or lower than one, the rate of occurrence increases or decreases, respectively, as the value of the exposure variable increases.

The parameter of the variable ***Log(Boardings)*** in the NB2-I model is significantly greater than one (Table 5), according to our expectations; this suggests that the entry of larger groups promotes higher evasion rates (e.g. due to anonymity or overcrowding). The analysis of the percentage elasticity of this variable shows that an increase of 1% in the number of people boarding at a given door increases 1.08% the amount of evasion through that door. In addition, a two-tailed t-test allows us to reject the null hypothesis that this parameter is equal to one ($2.96 > t_{n-2,99\%}$). Note that in the Poisson model this parameter was lower than one, a result that is not consistent with the increase on fare evasion due to overcrowding observed at the boarding doors (see Table 3).

The number of people alighting through a given door is also a variable with high statistical significance in the model. The semi-elasticity of the variable ***Exiting*** indicates that if a passenger boards when someone is alighting[10], evasion increases 24% on average, regardless whether the observation was collected at the back or front doors. In the back doors an increase in the average level of fare evasion when somebody exits through that door is expected, because back doors lack smartcard validators. In contrast, the result for the front door is not obvious as evasion can occur at this door, regardless of someone being alighting and there are fare card validators. In this case, we think that the increase on fare evasion may be due to a higher perception of anonymity when a mass of passengers alights at the door.

Fare evasion is expected to grow with an increase in the number of doors. Results show that an increase of one unit in the number of doors increases evasion 20.4% on average. Since the model controls for the level of occupancy, this variable could be capturing the effect of the size of the bus and the bus driver's difficulty to monitor more doors.

Regarding the boarding door chosen by passengers, our results indicate that at the front door the amount of evasion is 73.9% less than at the back door (*ceteris paribus*). As Transantiago buses have validators for smartcard payment only at the front door, it is hard for passengers boarding through the back doors to validate their cards (especially those forced to use the back door, due to overcrowding, but who would pay otherwise). Since the model already controls for the bus occupancy level, the average increase of evasion associated with boarding through the back doors could be also explained partially by the group of fare evaders who, regardless the chances to board through the front door, board by the back doors in order to evade. Therefore, even if this group of passengers was forced to board through the front door they could still evade and thus we must be careful with the conclusions drawn from the effect

[9] For linear variables the semi-elasticities keep constant, regardless of the value taken by the explanatory variables. In contrast, for logarithmic variables only the percentage elasticity has this property.

[10] Note that to estimate the regression models we only used observations where someone was boarding (Section 3.1). Then, if an observation records a passenger alighting at a given door, someone must have also been boarding through that door.

of a change in this variable. Furthermore, part of the reduction of fare evasion at the front door could be explained by the greater proximity and eye contact between the passengers and the bus driver, and the availability of a fare card validator.

A higher level of bus occupancy increases evasion. The semi-elasticity of the variable ***Low Occupancy*** shows that if the bus has a low occupancy level (less than half of the total bus floor area has people standing), evasion at each door is expected to fall by 16% (compared to high occupancy buses). This result could be explained by the difficulty to pay and the greater anonymity associated with a crowded bus either at the front or back doors. In addition, overcrowding promotes a bad perception of public transport quality of service; so part of the increase in fare evasion could be attributed to this effect.

As mentioned previously, we found that the best functional form to capture the effect of the bus route frequency was using its reciprocal, which represents a proxy for the headway of the bus route. The semi-elasticity obtained for the variable ***Headway*** shows that an increase of one minute in the average headway of a given bus route increases 1.8% the amount of evasion at each door. Since the model includes a variable to control the bus occupancy level, the ***Headway*** captures mainly the effect of waiting times at bus stops. Like overcrowding, longer waiting times deteriorate the perception of the public transport system; thus, the high fare evasion rates reported in bus routes with longer headways could be explained mainly by a poor perception of service quality.

Regarding the changes in fare evasion observed throughout the day, our results show significant differences between some periods (*ceteris paribus*). The value of the estimated parameters of the variables used to capture the effect of the time periods (Table 5) indicate that: (i) fare evasion rates during the afternoon are expected to be greater than in the morning period; (ii) the highest fare evasion rates are during the evening[11]; (iii) fare evasion levels are not significantly different during the various afternoon periods. Regarding the first two findings, the semi-elasticities of the variables ***Afternoon*** and ***Evening*** show that during the afternoon and evening increases of 28.9% and 72.9% are expected an in fare evasion, respectively (compared to the morning peak period). The third finding is interesting, since it challenges the idea that fare evasion rates are necessarily higher during peak hours. Even though, during the morning off-peak fare evasion falls by 13.1% compared to the morning peak (*ceteris paribus*), during the afternoon time periods (peak noon, afternoon off-peak, afternoon peak) there are no significant differences in fare evasion. In general, the dummy variables used to control for the time of day may be capturing systematic changes on some factors during the day, which impact fare evasion significantly but are not included explicitly in the regression models such as socio-demographics (i.e. age, gender and occupation, among others), travel purpose and the level of ticket inspection (Bucciol et al., 2013).

Our results also indicate that fare evasion levels are significantly different between some of the bus operating companies, due to effects that are not captured by the rest of variables in the

[11] The San Francisco Municipal Transportation Agency (SFMTA) reported similar trends using a descriptive statistics analysis (Lee, 2011). Unlike our regression models, their approach does not allow to isolate the effect of the period of the day, since other variables tend to be highly correlated with the time period (level of occupancy and headway, among others) explaining jointly fare evasion.

model. The semi-elasticities of these variables indicate that the four groups of private operators exhibit different average evasion levels (*ceteris paribus*): 20.3%, 35.9% and 45.6% less evasion than the control group. The Transantiago private operators use different strategies to tackle evasion, such as turnstiles for passengers boarding at the front door and inspectors monitoring the entrance at bus stops. In addition, bus drivers may have a more important role in reducing fare evasion in some operating companies. Since these factors are not included in the model, the high statistical significance of the dummy variables designed to control for each group of private operators could be capturing, partially, the observed differences in fare evasion among them (Section 3.4).

Finally, the level of income of the municipality is also an important factor to explain the evasion behaviour of passengers. The semi-elasticities of the variables ***Upper-Middle Income*** and ***High Income*** in the NB2-I regression model indicate that a 20.8% decrease on the average evasion at each door is expected if a stop is located in an upper middle-income municipality, compared to that in low-income and lower middle-income municipalities (control group); in high-income municipalities the model suggests a 30.7% decrease with respect to the control group. However, the socioeconomic level (SEL) of the area of influence of the bus stop does not always match with the SEL of the individuals boarding. To capture the effect of individual's SEL more accurately, we re-estimated the NB2-I model using data collected only during the morning peak (Section 3.3). In this case, results confirm our expectations once more, as an increase in the average household income associated with the municipality where the observation was collected decreases evasion significantly (*ceteris paribus*).

4. POLICY IMPLICATIONS AND DISCUSSION

In this section we propose five strategies to tackle fare evasion derived from our analyses. After, we state other important considerations for the design of fare evasion policies, which are related with the role of bus drivers, psychosocial factors, public perceptions, institutional design, enforcement strategies and equity issues.

4.1. Strategies to tackle fare evasion

Based on the results of our models, to tackle fare evasion we recommend changes to the operation of the bus system using the following five strategies: i) increase the bus fleet; ii) improve the bus headway regularity; iii) implement off-board payment stations; iv) change the payment system on board of buses; v) modify the bus design by either changing the number of doors or the bus capacity (number of passengers/bus).

Note that although a cost-benefit exercise is out of the scope of this paper it should be performed to evaluate the feasibility of these strategies. If decision makers can estimate the costs associated with changes in the independent variables, the model can be used to predict the expected benefits as a function of the reduction in fare evasion.

4.1.1. Bus fleet size

A larger bus fleet operating on a given route and period will increase bus frequency. In terms of the variables included in the model, this policy should reduce headways, the level of bus occupancy, the number of passengers boarding simultaneously at a given door, the cases when someone boards using the same door where somebody else is alighting (captured by the variable ***Exiting***) and the proportion of people boarding by the back doors (as discussed in Section 3.4, some people are going to evade even if they have the chance to enter by the front door). Further, improving the bus route frequency in this way has an impact on all the bus stops served by a bus service, and thus, it can reduce significantly the level of fare evasion in the bus route overall.

4.1.2. Regularity

As improving bus frequency by increasing the bus fleet could be an expensive proposition, introducing control strategies to improve headway regularity (Delgado et al., 2009) could reduce overcrowding both on buses and at bus stops, without adding additional buses or hiring more bus drivers. In this scenario, our regression model shows that fare evasion would decrease by improving bus regularity, since it reduces the number of people boarding simultaneously at a given door, the number of cases of passengers boarding and exiting by the same door, and the proportion of buses with high occupancy (by balancing the number of passengers/bus). Note that there are other factors, not captured by our models, which could reduce fare evasion as well (i.e. achieving a more positive perception of the system).

In a field experiment designed to improve bus headway regularity on some bus routes in Transantiago, bus control strategies were able to reduce fare evasion and headway irregularity (Lizana et al., 2014). The outcome of this experiment is very interesting and supports our findings, since part of the reduction in fare evasion could be explained by the change in the value of the operational variables included in our models (i.e. the headway regularity reduced bus occupancy, and thus, evasion is expected to be lower).

4.1.3. Improved on-board fare payment system

Improvements in the on-board payment system through the implementation of rear smart card validators, for example, could decrease fare evasion. Back-door entry can speed up the boarding process helping to reduce loading times (Stewart and El-Geneidy, 2014), yielding increases in bus frequency (as cycle times decrease); this could reduce fare evasion significantly at all bus stops served by the bus route (Section 4.1.1); bus overcrowding is also linked to entry through the back doors (where there are no card validators). In some cases, people board through back doors when the front of the bus is too full, even if they are not attempting to evade. Thus, rear door validators could help to reduce fare evasion for this "group of fare evaders", but it would not ensure that all passengers boarding would validate their card. In particular, *opportunistic evaders* could take advantage of the anonymity that occurs in the back door and still evade (Torres-Montoya, 2014). However, the cost of fare card validators and the technical feasibility of its implementation must be considered (in Transantiago the current technology only allows for two validators per bus which are usually located at the front of the bus). Overall, this technology could be cost-effective depending on the balance between the revenues for operational improvements (Stewart and El-Geneidy, 2014), the net benefits due to the change (reductions or increases) in fare evasion at each door, and the cost of its implementation.

4.1.4. Off-board payment stations

Off-board payment stations are enclosed bus stops where payment must be made to an inspector before boarding the bus; they are usually located in areas with high passenger flows and high levels of evasion (DTPM, 2013a). However, off-board payment stations may not be the best method to directly reduce evasion, as they are associated with high infrastructure and operational costs, and require public space intervention. In contrast, by speeding up the loading process at a bus stop (i.e. passengers can board through all doors), off-board payment stations can improve the bus route frequency and, thus, can reduce fare evasion indirectly at all bus stops in the route (Section 4.1.3). In addition, by allowing all cards to be validated prior to boarding, off-board payment stations solve the problem of overcrowding at the bus stop (even more than allowing for back door boarding), which also contributes to reduce fare evasion at the bus stop.

Note that while off-board payment platforms have the benefit of reducing fare evasion in all bus stops served by the bus route, it is not entirely clear if fix inspectors are effective in reducing evasion overall, since passengers wishing to evade will avoid using fixed prepaid stops (a portion of evaders could move to a nearby bus stop). Thus, fixed inspection at off-board payment stations should be evaluated carefully, especially if there are a high proportion of dedicated evaders skipping inspections. Ex-post and ex-ante studies should be performed to evaluate changes in the flow of passengers among bus stops close to off-board payment stations at the period when it was implemented. This would help to determine, inter alia, the evasion benefits of fixed inspection at off-board payment stations (as opposed to their operational benefits) and could be compared with the effectiveness of using random inspectors.

4.1.5. Bus design

Our results shows that fare evasion is expected to be higher in buses with high occupancy and with more doors (Section 3.4), and thus, it could be tackled by either reducing the number of doors or increasing the bus capacity (number of passenger per bus). However, in both cases we suggest that operators perform a cost-benefit analysis weighting the net change on fare evasion and the cost of the implementation of any of these strategies. The investment required to increase the bus fleet capacity (i.e. buses with two floors) could be an expensive proposition. Furthermore, the reduction of the number of doors per bus could increase the average headway significantly (due to the increase in loading times), and thus, produce a net increase of fare evasion.

4.2. Recommendations

In this section we state some considerations derived from our findings. However, in contrast to the strategies proposed previously, the impact of these policies cannot be measured accurately, either because the model does not include explanatory variables to capture the effect of these policies (i.e. the relation between the rate of inspection and fare evasion) or because it is not possible to measure their impact on the independent variables (i.e. the change of the public perception).

4.2.1. Psychosocial factors

Fare evasion is a psychosocial problem that arises from the interaction between a person (or group of people) and a context that imposes restrictions on its welfare (Sluzki, 1996). There are psychological components involved in a fare evader's behaviour that are also observed in other types of illegal acts (or cheating) that violate widely held moral principles. There are also group influence processes that affect the individuals and seem to explain the behaviour of fare evaders (Bucciol et al., 2013). Generally, people are sensitive to social pressures and find comfort in anonymity (i.e. in bus overcrowding) or following group behaviour (Gino et al., 2009). In these situations individual standards can be replaced by group norms leading to *social contagion* (Hatfield et al., 1994), especially when members of our own group are involved, as this reduces its unethical perception (Gino et al., 2009). Recent studies have shown that under certain conditions fare evasion is expected to increases if people observe others evading (Bucciol et al., 2013). Evasion contagion has not been well studied and could help to explain the progressive increase of fare evasion in some public transport systems. From a public policy standpoint, it is possible to design dissuasive campaigns to change the perception of evasion and increase the social costs of evading, such as shame, fear and guilt (Torres-Montoya, 2014); however, its effectiveness will depend on the motivations to not pay the fare behind each fare evader.

4.2.2. Public perception

The close relationship between unlawful conduct and the public perception about institutions should also be considered (Karstedt and Farrall, 2006). In early 2013 consultants revealed that 61% of people were unhappy with Transantiago´s performance (EMOL, 2013). In this sense, an unknown proportion of evasion could be associated with dissatisfaction with the system (Torres-Montoya, 2014). Thus, evasion could be a form of protest, triggering an incentive for unethical behaviour, a phenomenon known as the *cheater's high* (Ruedy et al., 2013). In addition, evasion can also be seen as "… voluntary unethical behaviour without obvious harm or a salient victim" (Ruedy et al., 2013), by people that do not consider losses to the public transport system to be actually harmful. Consequently, increasing fines and levels of control could be effective measures to reduce evasion, but could worsen the public perception about the system even more. In contrast, improvements to the system's level of service, besides reducing fare evasion (Section 3.4), would increase the comfort of passengers, yielding a more positive perception of the system.

4.2.3. Role of bus drivers

Bus drivers can also play a potentially significant role in reducing evasion, as long as the passengers see them as an authority and to the extent that sufficient proximity exists, making eye contact possible (Blass, 1999). In this regard, improvements on the operation of the system could help with the enforcement role of bus drivers (i.e. as bus occupancy is lower, eye contact between passengers and driver increases). Bus drivers could also need economic incentives to enforce passenger payment, and this would require changes to the contractual relationships between drivers and firms. However, the modifications required should be looked carefully, since they could produce undesired effects on the system's operation as happened in the system prior to Transantiago (i.e. buses stalling or racing, with dangerous consequences). In addition, the relationship between passengers and drivers in Transantiago is complicated and some drivers feel that their safety would be at risk if they take on an enforcement role (Bonfanti and Wagenknecht, 2010).

4.2.4. Institutional design

Without social pressure enforcement has to come from some form of authority. One of the problems that Transantiago faces is designing institutional relationships that clearly identify who is responsible for addressing evasion (Torres-Montoya, 2014). The system is overseen by a government agency, but private firms operate the buses. While the firms can hire inspectors, the legal authority to fine passengers remains only with the national police. In February 2012, the contracts were changed shifting the financial incentive to reduce evasion to the firms; this produced a significant downturn in the evasion rates observed in the system afterwards (Figure 1). However, more data should be collected to see if contractual changes at different points in time (i.e. February 2012) have had a long-term impact on reducing evasion levels overall. Furthermore, our results indicate that some bus operating-companies exhibit lower levels of fare evasion due to effects that are not captured by the variables included in the model. In this sense, it would appear interesting to review if some elements of the contractual relationship between the government agency and these operators may have promoted better policies to tackle fare evasion.

4.2.5. Inspection scheduling

Our results showed no average differences (*ceteris paribus*) in the level of evasion during the whole afternoon (i.e. including the midday peak, afternoon off-peak and afternoon peak hours), and indicated that afternoon evasion is expected to be higher than in the morning (i.e. morning off-peak and morning peak hour). This suggests, inter alia, that fare inspection in some bus routes operating during the afternoon off-peak period could be cost-effective (as buses are less crowded, making it easier to inspect passengers without generating any delay in the network, e.g. increasing travel times and generating traffic congestion). Using another dataset supplied by the MTT for 2010, we observed that the fare inspection levels varied considerably between time periods: 35.7% of all inspections occurred during the afternoon, whereas the morning off-peak had 30.2%, the morning peak 7.4% and the evening only 2.7%. This data could help to explain the increase in fare evasion rates during the evening captured in the regression models (*ceteris paribus*), as the flow of passengers in this period represent more than 10% of the total trips (SECTRA, 2001) and inspection is less frequent than what it should be. Thus, increasing inspections in the evening could be cost effective too. It is interesting to note that other public transport agencies have concluded that ticket inspection schedules during the afternoon and evening can be more cost-effective than in other time periods (Lee, 2011).

4.2.6. Equity issues

Fare increases affect a large proportion of public transport users in Santiago, especially those belonging to low-income groups who tend to be captive. Up until 2006, the average household income of the first quintile of households in Santiago was less than US$ 312.2/month[12] and its average family composition was 4.3 people (Observatorio Social, 2006). In this setting, roughly 30% of public transport users belonged to families with an average monthly household income in this quintile (SECTRA, 2006). Further, a traditional family within the first quintile, where both parents commute twice per day[13], would have spent 25% of their monthly budget in public transport in 2006. The situation is even worse if we add the trips made during the weekend.

In the context of Transantiago, the high fare evasion rates measured in low-income areas could be legitimately linked to an inability to afford public transport, and could explain, partially at least, the average differences on fare evasion observed between municipalities in the regression models (Table 5). Thus, from a public policy standpoint there is a need to consider transport subsidies to reduce the cost of the fare, whether targeted to low-income riders or overall. In contrast, enforcement strategies in low-income areas could reduce the rate of fare evasion marginally, but at the cost of limiting the accessibility of people who cannot afford the ticket.

[12] In 2006, 1 US$ was equal to 530.26 CLP in average (BCC, 2015).

[13] SECTRA(2006) reported that families with an average household income lower than US$ 500 made 1.3 daily trips per person in motorized modes.

4.3. Final thoughts

Our view of fare evasion is more complicated than passengers simply acting as rational actors who attempt to maximize utility. We believe that in order to address evasion it is necessary to examine the social and urban system contexts where fare evasion takes place. Our econometric approach allowed capturing correctly the joint influence of different variables on the level of fare evasion. The models provide a useful tool for decision makers to predict the impact on fare evasion of different public policies that modify the level of some of the variables included in the models. Our methodology could be applied to other transport systems, with the caution that the values of the estimated parameters might change and, with them, their effect on public policy geared to tackle fare evasion. For example, in some transport systems overcrowding could be a major issue, a higher effect of operational measures to reduce bus occupancy should be expected.

In further research, a cost-benefit analysis should be performed to evaluate the feasibility of the various strategies proposed to tackle fare evasion. In addition, other model specifications could be tested using more accurate data about headways (i.e. bus GPS data instead of theoretical values), bus occupancy (i.e. the exact number of passengers per bus) and the socio-economic level of individuals (i.e. GIS data could allow estimates of the level of income of neighbours living within a radius of influence defined for each bus stop). Furthermore, a longitudinal analysis using all available data about evasion measurements between 2007 and 2012 could be very interesting, integrating variables related to public opinion, ticket price and rate of inspection, among others. Further, longitudinal models would allow testing for auto regressive effects (i.e. of the evasion rate in previous periods), i.e. that the perception of more evasion in the system produces a greater willingness to evade in subsequent periods.

The effect of some psychosocial variables should be also captured in new specifications of our models. People in collective contexts may experience *deindividuation*, leading them to adopt group norms, increasing their susceptibility to social group influence (Zimbardo and Leippe, 1991). Under these conditions, they tend to replace their individual standards with those of the group, increasing imitation and experiencing social contagion (Hatfield et al., 1994). Thus, the perception that others are evading may actually increase evasion and even legitimize it. To test these hypotheses it could be useful to analyse if the source of overdispersion detected in the Poisson regression model is linked to true contagion rather than *apparent dependence* between the random variables of the probability distribution. The presence of *true contagion* would support the hypothesis that fare evasion increases if people observe others evading, and this would be also consistent with the social contagion processes involved in unethical behaviour (Gino et al., 2009).

Finally, we suggest that future work on fare evasion integrates an interdisciplinary perspective, mixing both qualitative and quantitative research methods (Clifton and Handy, 2001). Relying exclusively on statistical analyses could generate a mis-understanding about a problem as complex as fare evasion. Even though, the methodologies used in this paper were mainly based on quantitative methods, literature from psychology related to unethical behaviour (Gino et al., 2009; Karstedt and Farrall, 2006; Ruedy et al., 2013) and group behaviour (Blass, 1999; Hatfield et al., 1994; Zimbardo and Leippe, 1991) was essential to interpret and support the results obtained with the regression models. We hope our findings will generate an impact on public policy discussions that are currently under debate and will provide guidelines for further research. Observed evasion is only the "tip of the iceberg" of a much deeper psychosocial problem.

ACKNOWLEDGMENTS

We would like to thank Jorge Manzi, Juan Carlos Muñoz, Angelo Guevara, Susan Handy and Mariana Torres-Montoya for helping us to improve this work. We are grateful to the Programa Nacional de Fiscalización, Ministerio de Transporte y Telecomunicaciones, Chile, for giving us access to their data. We also wish to thank the Millennium Institute in Complex Engineering Systems (ICM: P-05-004-F; CONICYT: FBO16), FONDEF D10I1049 "Una herramienta táctico-estratégica de gestión y planificación de sistemas de transporte público urbano", the Centre for Sustainable Urban Development, CEDEUS (Conicyt/Fondap/15110020) and the Bus Rapid Transit Centre of Excellence funded by VREF (www.brt.cl), for their support. Finally, we would like to thank two anonymous referees for their helpful comments, which certainly contributed to improve this paper.

REFERENCES

Abdi, H., Edelman, B., Valentin, D. and Dowling, W.J. (2009) **Experimental design and analysis for psychology**, Oxford University Press, Oxford.

Barabino, B., Salis, S. and Useli, B. (2013) A modified model to curb fare evasion and enforce compliance: empirical evidence and implications. **Transportation Research Part A: Policy and Practice**, 58, 29–39.

BCC (2015) Banco Central de Chile. **Base de Datos Estadísticos**. <si3.bcentral.cl/Siete/secure/cuadros/home.aspx> (retrieved 03.20.15).

BCN (2013) Biblioteca del Congreso Nacional de Chile. **Reportes Estadísticos Comunales 2013**. <http://reportescomunales.bcn.cl/2013/> (retrieved 01.10.14).

Blass, T. (1999) The Milgram paradigm after 35 years: some things we now know about obedience to authority. **Journal of Applied Social Psychology**, 29, 955–978.

Bonfanti, G. and Wagenknecht, T. (2010) Human factors reduce aggression and fare evasion. **Public Transport International**, 59, 28–32.

Boyd, C., Martini, C., Rickard, J. and Russell, A.R. (1989) Fare evasion and non-compliance. **Journal of Transport Economics and Policy**, 23, 189–197.

Bucciol, A., Landini, F. and Piovesan, M. (2013) Unethical behavior in the field: demographic characteristics and beliefs of the cheater. **Journal of Economic Behavior & Organization**, 93, 248–257.

Cameron, A.C. and Trivedi, P.K. (1998) **Regression Analysis of Count Data**. 1st Ed, Cambridge University Press, New York.

Clifton, K.J. and Handy, S.L. (2001) Qualitative methods in travel behaviour research. In: *International Conference on Transport Survey Quality*. Kruger National Park, South Africa.

Delgado, F., Muñoz, J.C., Giesen, R. and Cipriano, A. (2009) Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. **Transportation Research Record**, 2090, 59–67.

DTPM (2013a) Directorio de Transporte Público Metropolitano. **Informe de Gestión 2012**. <http://www.dtpm.cl/archivos/Informe_de_gestión_2012_4_de_Septiembre.pdf> (retrieved 01.10.14).

DTPM (2013b) Directorio de Transporte Metropolitano. **Programa de Operación Segundo Semestre 2013**. <http://www.dtpm.cl/index.php/plan-operacional-historico2> (retrieved 01.02.15).

DTPM (2014) Directorio de Transporte Público Metropolitano. **Informe de Gestión 2013**. <http://www.dtpm.cl/archivos/Memoria 2013-Web_Final (1).pdf> (retrieved 04.27.15).

EMOL (2007) **"Transantiago comenzó a operar con tres días de gratuidad en el pasaje."** <http://www.emol.com/noticias/nacional/2007/02/10/245411/transantiago-comenzo-a-operar-con-tres-dias-de-gratuidad-en-el-pasaje.html> (retrieved 01.10.14).

EMOL (2013) **Transantiago llega a su mayor desaprobación desde 2007, con 61 % de rechazo**. <http://www.elmercurio.com/blogs/2013/01/28/8825/Transantiago-llega-a-su-mayor-desaprobacion-desde-2007-con-61-de-rechazo.aspx> (retrieved 01.10.14).

Gino, F., Ayal, S. and Ariely, D. (2009) Contagion and differentiation in unethical behavior: the effect of one bad apple on the barrel. **Psychological Science**, 20, 393–398.

Greene, W.H. (2007) **Econometrics Analysis**. Seventh Ed, Pearson Education, Westford.

Hatfield, E., Cacioppo, J.T. and Rapson, R.L. (1994) **Emotional Contagion**, Cambridge University Press, Cambridge.

Hilbe, J. (2011) **Negative Binomial Regression**. Second Ed, Cambridge University Press, Cambridge.

Karstedt, S. and Farrall, S. (2006) The moral economy of everyday crime: markets, consumers and citizens. **British Journal of Criminology**, 46, 1011–1036.

Killias, M., Scheidegger, D. and Nordenson, P. (2009) The effects of increasing the certainty of punishment: a field experiment on public transportation. **European Journal of Criminology**, 6, 387–400.

Kooreman, P. (1993) Fare evasion as a result of expected utility maximisation: some empirical support. **Journal of Transport Economics and Policy**, 27, 69–74.

Lee, J. (2011) Uncovering San Francisco, California, Muni's proof-of-payment patterns to help reduce fare evasion. **Transportation Research Record**, 2216, 75–84.

Lizana, P., Muñoz, J.C., Giesen, R. and Delgado, F. (2014) Bus control strategy application: case study of Santiago transit system. **Procedia Computer Science**, 32, 397–404.

MDS (2009) Ministerio de Desarrollo Social. **CASEN 2009**. <http://www.ministeriodesarrollosocial.gob.cl/casen/Estadisticas/ingresos.html> (retrieved 01.10.14).

Muñoz, J.C., Ortúzar, J. de D. and Gschwender, A. (2008) Transantiago: the fall and rise of a radical public transport intervention. In: Saleh, W., and Sammer, G. (Eds.), *Travel Demand Management and Road User Pricing: Success, Failure and Feasibility*. Ashgate Publishing, pp. 151–172.

O'Brien, R.M. (2007) A caution regarding rules of thumb for variance inflation factors. **Quality and Quantity**, 41, 673–690.

Observatorio Social (2006) **Estadísticas Ingresos**. <http://observatorio.ministeriodesarrollosocial.gob.cl/casen/casen-documentos.php?c=85&a=2006> (retrieved 01.10.14).

Ortúzar, J. de D. and Willumsen, L.G. (2011) **Modelling Transport**. 4th Ed, John Wiley & Sons, Chichester.

Osgood, D. (2000) Poisson-based regression analysis of aggregate crime rates. **Journal of Quantitative Criminology**, 16, 21–43.

Reddy, A., Kuhls, J. and Lu, A. (2011) Measuring and controlling subway fare evasion: improving safety and security at New York city transit authority. **Transportation Research Record**, 2216, 85–99.

Royston, P. (1995) Remark AS R94: a remark on algorithm AS 181: The W-test for normality. **Applied Statistics**, 45, 547–551.

Ruedy, N.E., Moore, C., Gino, F. and Schweitzer, M.E. (2013) The cheater's high: the unexpected affective benefits of unethical behavior. **Journal of Personality and Social Psychology**, 105, 531–548.

Sabatini, F., Wormald, G., Sierralta, C. and Peters, P.A. (2009) Residential segregation in Santiago: scale-related effects and trends, 1992-2002. In: *Urban Segregation and Governance in the Americas*. Palgrave Macmillan, pp. 121–144.

SECTRA (2006) Secretaría de Planificación de Transporte. **Encuesta Origen Destino 2006**. <http://www.sectra.gob.cl/datos_e_informacion_espacial/gran_santiago/encuestas_movilidad.html> (retrieved 01.10.14).

Shields, T.G. and Huang, C. (1995) Presidential vetoes: an event count model. **Political Research Quarterly**, 48, 559–572.

Sluzki, C. (1996) **La red social: frontera de la práctica sistémica**, Gedisa, Barcelona.

Stewart, C. and El-Geneidy, A. (2014) All aboard at all doors: route selection and running time savings estimation for multi-scenario all-door bus boarding. In: *93rd Annual Meeting of the Transportation Research Board*. Washington, D.C.

Thorlacius, P., Clausen, J. and Brygge, K. (2010) Scheduling of inspectors for ticket spot checking in urban rail transportation. **Trafikdage på Aalborg Universitet 2008**,.

Torres-Montoya, M. (2014) Tackling fare evasion in Transantiago: an integrated approach. In: *93rd Annual Meeting of the Transportation Research Board*. Washington, D.C.

Zimbardo, P.G. and Leippe, M.R. (1991) **The Psychology of Attitude Change and Social Influence.**, McGraw Hill, New York.