FREQUENCY AND CAPACITY COMPETITION IN DEREGULATED URBAN BUS SYSTEMS

Marco Batarce, Universidad San Sebastián, Santiago, Chile – marco.batarce@uss.cl

Palabras clave: public transport, competition, regulation

ABSTRACT

This paper helps to explain why the frequency is high and bus capacity low when the urban bus transit market is not regulated. To do so, we assume bus services compete on frequency and capacity in a two-stage game, given a regulated flat fare. We also compare the effect of different operation schemes for bus services on the provided frequency. This comparison allows us to identify tools for the regulation of urban bus transit services.

1. INTRODUCTION

A feature of the liberalization of bus transit services has been the increase of frequency in comparison to that in a regulated industry (Evans, 1990, Paredes and Baytelman, 1996, Estache and Gomez-Lobo, 2005). On the one hand, this phenomenon has produced significant externalities such as congestion, pollution, and accident risk in developing countries (Estache and Gomez-Lobo, 2005). However, on the other hand, it has improved some aspects of the service quality: waiting time has been reduced significantly, and provided transport capacity has increased, reducing passenger overcrowding.

If the externality problems could be overcome without limiting competition in the field, a valid question is whether the competitive equilibrium is better than the monopolistic one, either regulated or not, for users. This paper aims to analyze the frequency competition in urban bus transit. We compare the effect of different industrial organizations for bus services on the supplied frequency and the consumer's surplus. Notably, we consider competition on frequency, two unregulated monopolies with different operation schemes, and a regulated monopoly with welfare-maximizing frequencies.

Another relevant feature of deregulated of urban bus markets is the reduction in bus capacity. Small buses or minibuses wait for passengers in long lines in CBD streets of big cities in developing countries. But the capacity reduction also happens in developed countries. For instance, after deregulation in Great Britain, minibuses spread rapidly, thereby providing more frequent service

(Gómez-Ibañez and Meyer, 1997).

To study the effects of competition on bus capacity, we assume that the operators play a game in frequency and bus capacity in two stages. In the first stage, the operators choose the capacity, and then they choose the frequency in the second stage, given the capacity. We solve the game assuming open-loop strategies for the sake of tractability. This assumption implies that the bus operators do not take into account the effect of their decision about bus capacity on the frequency chosen by their rivals in the second stage. The assumption is also consistent with pre-commitment, if the firms choose the bus size before starting operation and cannot change it in the short term. When firms realize the effect of their bus capacity decision on the rivals' frequency, they might play closed-loop strategies.

2. THE MODEL

We distinguish three components of the model: users, firms, and network. The users' behavior is summarized in the demand model. The firms operate the bus lines and maximize profits for the specified cost function. The network is the representation of the bus lines running in the city and stops where demand concentrate. We assume that the total demand at every bus stop in the network is inelastic to the frequency. We also consider the network and the number of firms as given. Hence, the optimal network configuration and the number of firms are not studied.

2.1. Demand

The demand modeling approach is based on what is known as the bus stop problem in the literature on transit network assignment (Bouzaiene-Ayari et al. 1998; Nokel and Wekeck, 2007). These models aim to determine the passenger distribution across all available (or attractive) bus lines in a stop and the passengers' waiting time at the stop.

Research in the context of transit assignment models indicates that bus stop models distributing passengers between lines based only on the frequency give unrealistic results (Gendreau, 1984; Nokel and Wekeck, 2007). For instance, the frequency share (FS) model used in some air transportation studies (Douglas and Miller, 1974; Vander Weide and Zalkind, 1981; Kawasaki and Li, 2013; Brueckner and Flores-Fillol, 2007). To overcome this lack of realism, different authors propose to take into account the effect of bus capacity and passengers aboard the bus from upstream stops. Gendreau (1984) uses the share of the residual capacity (RC) of each line. In their model, the residual capacity corresponds to the nominal capacity minus the total passengers onboard the bus right after the stop. Since the RC model delivers better results than the FS model only when the network is very congested, Bouzaiene-Ayari (1988) proposes the adjusted residual capacity model (ARC). In this case, passengers are distributed in proportion to the frequency multiplied by the ratio between residual capacity and nominal capacity. Thus, the ARC model approximates the FS model when congestion is low because the residual capacity is similar to the nominal capacity, and approximates the RC model when congestion is high. More recently, Bouzaiene-Ayari et al.

(2001) propose a general bus stop model. Their model is based only on the assumption that there is an attraction factor for each line, which is a function of frequency, nominal capacity, and total passengers on the bus after the stop. Given these factors, the demand for each line is distributed in proportion to its attraction factor. These models are founded on the queue theory. For instance, Cominetti and Correa (2001) assume that bus arrivals follow a Poisson process and use bulk queue theory to derive expressions for attraction factors in a transit network. They show that the attraction factors are the inverse of the expected waiting time for any attractive line at the bus stop. This paper adopts an approach based on the general stop model by Bouzaiene-Ayari et al. (2001) along with attraction factors as the proposed ones by Cominetti and Correa (2001).

Assumption 1. *The share of an attraction factor, ξ, gives the bus line's demand share at any stop. The line l's attraction factor is a function of its frequency, f_l, vehicle capacity, k_l, and total passengers on board from the preceding to the next stops in its route, v_l. Then, if L firms serve the stop, we define by λ_l the demand share attracted by the firm l as*

$$\lambda_l \equiv \lambda(f_l, k_l, v_l, f_{-l}, k_{-l}, v_{-l}) = \frac{\xi(f_l, k_l, v_l)}{\sum_{j=1}^{L} \xi(f_j, k_j, v_j)} = \frac{\xi(f_l, k_l, v_l)}{X(f, k, v)}, \tag{1}$$

where $f = (f_1, \ldots, f_L)$, $k = (k_1, \ldots, k_L)$, and $v = (v_1, \ldots, v_L)$.

By doing the attraction factor a function of the frequency, vehicle capacity, and passengers on board, we take into account the congestion on the bus network, which plays a crucial role in the firms' market share in the case of urban public transportation. The determinants of the attraction factors and the demand share are a result of the application of queue theory to the phenomenon of boarding a bus. In general, it is assumed that passengers wait for being served at the stop (the server), and the service process corresponds to the bus arrivals. If the bus and passenger arrivals at the bus stop are random, the attraction factor for a bus line is the inverse of its expected waiting time. This assumption implies that the longer the wait for a bus line is, the smaller the probability of taking this line will be, which is realistic. Besides, if several bus lines serve a stop, the random arrivals imply that the expected waiting time for any bus is the inverse of the sum of all attraction factors.

Assumption 2. *The attraction factor $\xi_l \equiv \xi\left(f_l, k_l, v_l\right)$ satisfies the following conditions:*

$$\xi(0, k, v) = 0 \quad \forall (k, v), \tag{2}$$

$$\xi(f, k, v) \geq 0 \quad \forall (f, k, v), \tag{3}$$

$$\frac{\partial \xi}{\partial f}(f, k, v) > 0, \quad \frac{\partial \xi}{\partial k}(f, k, v) > 0, \quad \frac{\partial \xi}{\partial v}(f, k, v) < 0 \quad \forall (f, k, v), \tag{4}$$

$$\frac{\partial^2 \xi}{\partial f^2}(f, k, v) \leq 0, \quad \frac{\partial^2 \xi}{\partial k^2}(f, k, v) \leq 0 \quad \forall (f, k, v). \tag{5}$$

$$\frac{\partial^2 \xi}{\partial f \partial k}(f, k, v) \geq 0 \quad \forall (f, k, v). \tag{6}$$

These assumptions imply that λ_l is an increasing and concave function of (f_l, k_l).

A particular case is the adjusted residual capacity (ARC) model (Bouzaiene-Ayari, 1988). In this case, the attraction factor is

$$\xi_l = \frac{f_l k_l - v_l}{f_l k_l} f_l = f_l - \frac{v_l}{k_l}. \tag{7}$$

The last member of the equation (7) is the *effective frequenc*y, which corresponds to the nominal frequency corrected by the total number of passengers on board going from the preceding to the next stops. If there are few passengers aboard, the effective frequency approximates the nominal one.

In general, if the total demand at any stop is Q passengers per hour, the line l's demand is given by

$$q_l = Q\lambda_l = Q\frac{\xi(f_l, k_l, v_l)}{\sum_{j=1}^{L} \xi(f_j, k_j, v_j)}. \tag{8}$$

2.2. Network

Bus routes are defined by a sequence of stops that concentrate the demand. In a bus network, several firms operate on the same roads and compete with each other. The degree of competition depends on the number of shared stops in the route. The higher the number of shared stops, the higher the rivalry.

We analyze the case of L lines coming from different stops, O_l, $l = 1, \ldots, L$, to a common route segment SD (see Figure 1). Each line l serves three stops and competes in the common segment of the route. Figure 1 shows the bus network in this case.

To understand this network representation, consider a circular city with a CBD surrounded by a residential belt where people live and from where they travel to the CBD to work and do other activities. In the residential zone, the road network is dense such that bus services locate separately one each other to avoid competition. This idea is similar to the Salop model for horizontal differentiation (Salop, 1979). However, in the CBD, the road network is concentrated, and only a few major streets can accommodate the bus services; therefore, the bus lines need to share the available road capacity to access the CBD and compete for the demand. For instance, in Santiago, only twelve main streets concentrated 80 per cent of the bus routes entering the CBD in 2001 (Malbran et al., 2001). Figure 1 represents several lines from the residential zone, which concentrate in one street, represented by route segment SD, when they enter the CBD.

Assumption 3. *The bus network is that represented in Figure 1, and the total demand for each served pair is perfectly inelastic to price and frequency (i.e., fixed demand) such that:*

- Q_l *is the total demand at the origin* O_l*,*
- Q_{lD} *is the demand between the origin* O_l *and the destination* D*, and*
- Q *is the total demand at the stop* S*.*

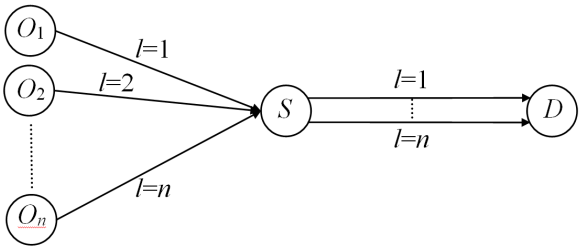

Figure 1: Network.

Thus, firms compete for the demand Q. When the supplied capacity, $k_l f_l$, is less than the total demand Q_l, the firm's demand equals the available capacity. Then, line l's demand at O_l, q_l^O, is the minimum between the total demand at the stop and the line's capacity $q_l^O = \min\{Q_l, k_l f_l\}$.

We assume the passengers boarding a bus distribute between the destinations S and D in the same proportion as the total demand at the stop. In other words, a passenger at O_l going to D boards an arriving bus with probability $\alpha_l = Q_{lD}/Q_l$, which is the fraction of the demand at O_l traveling to D. Hence, the number of passengers on board the line l at S and going from O_l to D is $v_l = \alpha_l q_l^O$. The assumption implies that all passengers waiting for the bus have the same probability of boarding when it arrives. This idea is consistent with the impossibility of distinguishing the passengers' destination when boarding the bus. Finally, the firm l's demand at S, q_l^S, is determined by the right-hand side of Equation (8) after replacing v_l by $\alpha_l q_l^O$, and the firm l's total demand is $q_l = q_l^O + q_l^S$.[1]

2.3. Firms

We consider a firm operating a bus line under a regulated environment. The regulator fixes the price ticket and the route of every line in the city. We assume that, under this regulation scheme, the firm can choose frequency and bus capacity to maximize profits. However, in the short term, only the frequency is a decision variable. In the medium or long term, firms can choose the bus capacity when renewing the fleet. We remark that the firms cannot change the route, and they must operate all of the buses from the initial stops to the final one.

Route regulation does not imply the transport authority specifies every street where the buses pass. But it means that authority can fix some routes segments to reduce some externalities of bus operations. For instance, in Santiago, the regulator set the routes only inside a perimeter around the densest zone of the city. Route regulation also means the firms cannot change their routes without approval from the authority.

We assume firms maximize profits by playing a two-stage game. In the first stage, firms play a game in (average) bus capacity. In the second stage, they play a game where the strategy is the frequency for a given capacity. Thus, the firm's strategy set is a subset of $\mathbb{R}^+ \times \mathbb{R}^+$, and the payoff

[1] We implicitly assume the total residual capacity provided by the firms is greater than the total demand at S.

function is

$$\pi_l(f_l) = p_l q_l - C_l(q_l, f_l, k_l), \tag{9}$$

where p_l is the fare, and $C_l(\cdot)$ is the cost function that depends on the firm l's frequency, bus capacity, and carried passengers. In what follows, we assume a cost function linear in passengers and frequency for a given capacity. However, we consider that the operation costs depend on the bus size (k_l). Specifically, we assume the cost per kilometer depends on the bus capacity. The larger the bus, the higher the cost per kilometer, and therefore the operation cost. Moreover, the operation cost, as a function of frequency, is different among firms because of the route length it is. Hence, we define a linear relationship between bus capacity and operation cost as Jansson (1980). By assuming the capacity is a continuous variable, we mean the firms can mix buses of different size to get the average bus capacity they want.

We assume that firms are heterogeneous because both fixed and operating costs depend on the firm's operation variables, such as route length, vehicle capacity, fleet average age. However, we assume the marginal cost per passenger is the same across firms. The firms' behavior is summarized in the following assumption.

Assumption 4. *(a) The firms are risk neutral and maximize their utility by choosing bus frequency and capacity.*

(b) For a given bus capacity, their cost functions are linear such that

$$C(q_l, f_l) = c_{0l} + cq_l + c_l(k_l) f_l, \tag{10}$$

where is the fixed cost, is the operation cost, and is the marginal cost per passenger.

(c) The operation cost depends on the bus capacity such that

$$c_l(k_l) = \rho_l \left(\gamma_0 + \gamma_1 k_l\right). \tag{11}$$

where ρ_l is the route length, and $\gamma_0 > 0$ and $\gamma_1 > 0$ are constants corresponding to fixed and variable operation cost per kilometer, respectively. In addition, bus capacity is a continuous variable in a compact interval in $\mathbb{R}^+$.

The essential requirement for a firm to operate is obtaining positive profits for some operation scheme. We consider a condition independent of the competition level and assume that the firm decides to enter the market by taking into account only the demand where it is a monopoly, which is a riskless decision. Moreover, since the demand is inelastic, the firm produces at level Q_l and without excess supply, which means that $Q_l = k_l f_l$. Thus, given the profit (9) and the cost function, it is straightforward to show the entry condition boils down to

$$p_l - c \geq \frac{c_l}{k_l} + \frac{c_{0l}}{k_l f_l} = \frac{\rho_l \gamma_0}{k_l} + \rho_l \gamma_1 + \frac{c_{0l}}{k_l f_l} \tag{12}$$

The intuition behind this condition is that the firm operates only if the marginal net revenue per passenger is higher than the average fixed cost of providing a seat (or place for a passenger). It is worth to notice that the last condition should hold even if the fixed cost is zero, because the operation cost is independent of the demand. We assume that the entry condition hold.

Assumption 5. *The entry condition of Equation (12) holds for every firm in the market.*

3. EQUILIBRIUM FREQUENCIES

Now, we study the equilibrium on frequency on the network of Figure 1. We assume the bus capacity is given by the firms' choice in the first stage of the game. The goal is to determine the equilibrium frequencies and their relationship with the parameters characterizing firms and demand. All proofs are omitted because the page number is limited.

3.1. Oligopolistic equilibrium

We assume firms behave non-cooperatively, and they consider the rivals' frequency as fixed when choosing their own. We distinguish two cases to analyze the equilibrium: the firm operates with an excess of demand in the first route segment, or it operates with an excess of supply. This distinction is necessary because the profit function has a discontinuous derivative in the frequency level where supplied capacity equals demand at the origin O_l. That implies different expressions for the equilibrium frequency in each case. However, the next lemma states a condition that simplifies the analysis. Formally, we denote the strategy set as $F_{Ll} = \{f_l \in \mathbb{R}^+ : k_l f_l < Q_l\}$ in the case of excess of demand, and $F_{Ul} = \{f_l \in \mathbb{R}^+ : k_l f_l \geq Q_l\}$ in the case of excess of supply.

Lemma 1. *Under Assumptions 1 to 5, for every line l, the set F_{Ll} is dominated by the set of strategies F_{Ul}.*

Lemma 1 implies that firms always choose a frequency such that it provides enough capacity to satisfy the demand in the first route segment, where it is a monopoly. Thus, we focus on the case where $f_l \in F_{Ul}$ hereafter. Lemma 1 also implies that $q_l^O = Q_l$, and $v_l = \alpha_l Q_l$.

Proposition 1. *Under Assumptions 1 to 5, a unique Nash equilibrium on frequency exists and the equilibrium frequency, f_l^*, is the solution to*

$$\xi(f_l^*, k_l, \alpha_l Q_l) = Q(p-c)\left[\frac{\tilde{c} - \tilde{c}_l}{\tilde{c}^2}\right] \tag{13}$$

where $\tilde{c}_l = (c_l - \mu_l)(\partial \xi_l / \partial f_l)^{-1}$, $\tilde{c} = \sum_{j=1}^{L} \tilde{c}_j/(L-1)$, and $\mu_l \geq 0$, $l = 1, \ldots, L$, are the multiplier associated with the constraints $k_l f_l^* - Q_l \geq 0$, and satisfy the complementarity conditions $\mu_l(k_l f_l^* - Q_l) = 0$.

Furthermore, frequencies are strategic substitutes because the reaction functions are decreasing and every firm's equilibrium frequency is a decreasing function of its bus capacity.

Finally, if firm l's is not dominant, such that its market share is less than 0.5, its equilibrium frequency decreases with the rivals' bus capacity. Otherwise, if firm l's is dominant, such that its market share is greater than 0.5, its equilibrium frequency increases with the rivals' capacity.

The parameter $\tilde{c}_l$ summarizes two effects: the constraints $k_l f_l^* - Q_l \geq 0$, through the multipliers, and the functional form of the attraction factors. The multipliers μ_l reduce the effect of operation

costs that enter Equation (14) in such a way that firms with high operation cost provide just the bus frequency needed to satisfy only the demand at O_l. Also, the multipliers ensure that the difference $(\tilde{c} - \tilde{c}_l)$ is always positive. In turn, the inverse of the derivative of ξ_l transforms the operation cost into the marginal cost of increasing the attraction factor at S. The effect of these derivatives offsets the impact of high operation costs such that, in equilibrium, firms with a higher marginal effect of the frequency on the attraction factor tend to deliver higher frequencies.

Firm heterogeneity also has an effect on competition: the more alike the firms are, the more incentives to compete for the demand at S and provide excess capacity. In fact, the firms with the highest operation cost only provide the residual capacity due to passengers alighting at S. In turn, firms with low per-kilometer cost have incentives to extend their routes to increase their potential monopoly demand. In this way, they can increase their frequency lower bound, which increases their equilibrium frequency and demand share at S. In the symmetric case, with identical firms, all of them provide the minimum frequency when the total residual capacity is greater than the total demand at S. Otherwise, all firms provide frequencies that equal their attraction factors.

As the equilibrium frequencies are the solution to a fixed point equation, the effects of the model parameters on the equilibrium are not evident. Therefore, the following corollary allows us to analyses the effects of the model parameters on the equilibrium frequency when the ARC model gives the attraction factors.

Corollary 1. *Under the same conditions as Proposition 1 and $\xi_l = f_l - v_l/k_l$, then if f_l^{ARC} is an interior solution*

$$f_l^{ARC} = \alpha_l \frac{Q_l}{k_l} + Q(p-c)\left[\frac{\bar{c} - c_l}{\bar{c}^2}\right], \tag{14}$$

where $\bar{c} = \frac{1}{L-1}\sum_{j=1}^{L} c_j$.

In the ARC model, the optimal frequency is composed of two terms. The first term on the right-hand side of Equation (14) corresponds to the frequency level needed to satisfy the demand from O_l to D. Since the demand is inelastic, the optimal frequency does not explicitly depend on the demand for trips going from O_l to S because the supplied capacity in that segment is higher than the demand. Indeed, the second term on the right-hand side of Equation (14) determines the additional frequency required to carry passengers boarding at S. As the frequency is an interior solution, it must satisfy $Q(p-c)(\bar{c} - c_l)/\bar{c}^2 > (1-\alpha_l)Q_l/k_l$. Therefore, only the demand at O_l going to D is relevant for the optimal strategy because such demand reduces the effective capacity in the route segment where the firms compete.

Both Corollary 1 and Proposition 1 show that the equilibrium frequency increases as a function of net revenue per passenger. This result helps to explain the excess of supply, which is a well-documented fact in the literature of public transportation deregulation (Evans, 1990; Estache and Gomez-Lobo, 2005; Gomez-Lobo, 2007; Fernandez and Muñoz, 2007; Paredes and Baytelman, 1996). Indeed, fare higher than marginal cost is a necessary condition for the operation of the service because of the existence of fixed costs and operation cost independent of the demand level. This feature holds in environments both competitive and price-regulated. However, in deregulated-price markets, equilibrium fare is higher than the necessary one to cover total operation cost or

long-run marginal cost, as Gomez-Lobo (2007) and Fernández and Muñoz (2007) show. Therefore, in competitive environments, there is a strong incentive to increase the frequency and operate with an excess of supply. Besides, the equilibrium frequency is increasing in the rival's operation cost, which implies that competition improves efficiency. Indeed, the less efficient firms offer a lower frequency and obtain a smaller demand share.

As the equilibrium frequencies are linear in demand at O and S, from some demand level, the competitive frequency will be higher than the socially optimal one. Indeed, the frequency that maximizes total welfare, including the users' waiting time, is proportional to the square root of total demand (Mohring, 1972; Jansson, 1980; Jara-Díaz and Gschwender, 2003).

3.2. Monopolistic equilibrium

We analyze the two cases. The first case corresponds to one firm operating all the lines in the network of Figure 1. We call this case direct-lines monopoly. In the second case, we assume the regulator adopts a different network configuration and reduces the route lengths. Thus, L lines operate from O_l ($l = 1, \ldots, L$) to S and one line operates from S to D. This structure is similar to a hub-and-spoke configuration, where the line in the segment SD is a trunk line. We call this case a hub-and-spoke monopoly. We assume that Assumption 5 holds in both cases.

The direct-line monopoly can be thought of as a system without integrated fares. In such a case, the lines serve all destinations without transfers from the origin. In the network of Figure 1, this operation scheme is reasonable, if the demand from O_l to D is higher than the demand from O_l to S. The hub-and-spoke monopoly may be a system with integrated fares and unbalance demand between route segments. The configuration depends on the demand and cost structure (see Fielbaum et al., 2016).

In the case of direct lines, all of them operate with a frequency equal to Q_l/k_l, because it maximizes revenues. However, if the residual capacity is less than the total demand at the stop S, the firm increases the frequency of the line with the lowest operation cost, because this leads to the highest marginal profit.

Lemma 2. *In the direct-lines monopoly, provided Assumptions 1 to 5 hold, the equilibrium frequencies are:*

(a) If the passengers alighting at S are more than the passenger boarding, such that $Q \leq \sum_{l=1}^{L}(1-\alpha_l)Q_l$, then the frequencies are $f_l^{DL} = Q_l/k_l$, for all $l = 1, \ldots, L$.

(b) If the passengers boarding at S are more than the passenger alighting, such that $Q > \sum_{l=1}^{L}(1-\alpha_l)Q_l$, then the frequencies are $f_l^{DL} = Q_l/k_l$ for all $l \neq m$, and $f_m^{DL} = (Q_l + Q - \sum_{j=1}^{L}(1-\alpha_j)Q_j)/k_m$, where the firm m has the lowest operation cost.

In the hub-and-spoke monopoly case, since the entry condition holds, the firms operate at the capacity level. Indeed, this level maximizes firms' profits because the demand is perfectly inelastic, and the marginal revenue is higher than the marginal cost because of the entry condition. We summarize this result in the following lemma.

Lemma 3. *Under the hub-and-spoke monopoly, provided Assumptions 1 to 5 hold, the equilibrium frequencies are $f_l^{HS} = Q_l/k_l$ for all $l = 1, \ldots, L$, and $f_S^{HS} = (Q + \sum_{j=1}^{L} \alpha_j Q_j)/k_S$ for the trunk bus line operating the segment SD with vehicle capacity k_S.*

3.3. First-best frequencies

We assume an individual's travel utility depends on the waiting time. In turn, the waiting time depends on the bus line's frequency. We define a linear utility $U(t_w, p) = u_0 - u_w t_w - p$, where u_0 and $u_w > 0$ are parameters, t_w is the waiting time at the bus stop, and p is the ticket price. The parameter u_0 represents the utility of travel. It includes the utility the individual obtains at the destination, which is related to the activity motivating the trip and the (dis)utility produced by the spent travel time. The total welfare depends on the value of u_0. Consequently, the demand level and the optimal frequencies also depend on its value. The minimum condition for an individual to travel is $u_0 \geq p + u_w t_w$.

In general, at uncongested stops served by one bus line with frequency ϕ, the expected waiting time is $t_w = \theta/\phi$. The parameter θ depends on the distribution of bus arrivals.[2] We assume exponential arrivals such that $\theta = 1$. At congested stops, the expected waiting time is the inverse of the attraction factor (Bouzaiene-Ayari et al., 2001; Cominetti and Correa, 2001). In the case of the stop S, where L lines serve it, the expected waiting time is the inverse of the sum of their attraction factors. Therefore, we assume the users' utility increase with the bus frequency since travel utility decreases with the waiting time, which is inversely proportional to the frequency. That is, the higher the frequency, the higher the user's utility. We summarize these assumptions as follows.

Assumption 6. *(a) The individual travel utility is $U(t_w, p) = u_0 - u_w t_w - p$, and for every passenger $u_0 \geq p + u_w t_w$.*

(b) The expected waiting times at every stop O_l ($l = 1, \ldots, L$) is

$$t_{wl}(f_l) = \frac{1}{f_l} \tag{15}$$

and the expected waiting time at S is

$$t_{wS}(f) = \frac{1}{\sum_{l=1}^{L} \xi_l(f_l)} = \frac{1}{X(f)}. \tag{16}$$

The social planner (or regulator) maximizes the total welfare, which corresponds to the consumers' surplus plus the sum of firms' profit. The consumers' surplus is

$$CS = \sum_{l=1}^{L} q_l^O \left(u_0 - \frac{u_w}{f_l} - p \right) + \sum_{l=1}^{L} q_l^S \left(u_0 - \frac{u_w}{X(f)} - p \right) \tag{17}$$

[2]For instance, in the case of exponential arrivals, equals to one (Spiess and Florian, 1989). The case $\theta = 1/2$ is an approximation of constant headway $1/f$. This measure of waiting time is widely used in practice (e.g., Mohring, 1972, and Jansson, 1980), even though it is based on a rough approximation (Spiess and Florian, 1989).

and the firms' profit is

$$P = \sum_{l=1}^{L} \left(q_l^O + q_l^S\right)(p - c) - c_l^s f_l - c_{0l} \tag{18}$$

where c_l^s is the social operation cost, which includes the effect of externalities from bus operation as congestion, pollution, and noise (see Rizzi and De la Mazza, 2017).

Following the literature on optimal bus frequency (Mohring, 1972; Jansson, 1980; Jara-Díaz and Gschewnder, 2003), we could assume the optimal level is high enough to satisfy all the demand in the network. Hence the demand is a parameter in the objective function. However, the condition that supplied capacity is higher than demand can be derived from the previous assumptions. The following lemma states that Assumption 5 implies that the social planner chooses frequencies at levels higher or equal to the demand at every origin O_l.

Lemma 4. *If the social welfare is $W = CS + P$, then the first-best frequencies satisfy all passenger demands at origins such that $f_l^{FB} \geq Q_l/k_l$, for all $l = 1, \ldots, L$.*

Lemma 4 follows from the monotonicity of the social welfare function when $f_l < Q_l/k_l$. In turn, this property steams from the linearity of the consumers' surplus at O_l and firms' costs and the positive marginal benefit from waiting time reductions at S. Thus, it is optimal for the regulator to choose frequencies greater or equal to Q_l/k_l. The case for the provided capacity at S depends on the cost parameters and demand levels nonlinearly. Thus we cannot show that the residual capacity at S is always higher than the demand. Therefore, we need to make an assumption.

Assumption 7. *The provided transport capacity at the first-best frequencies are higher than the demand at S. Thus the social planner's objective function is*

$$W = \sum_{l=1}^{L} Q_l \left(u_0 - \frac{u_w}{f_l} - c\right) + Q\left(u_0 - \frac{u_w}{X(f)} - c\right) - \sum_{l=1}^{L} (c_{0l} + c_l^s f_l). \tag{19}$$

Proposition 2. *Under Assumptions 1 to 7, the frequencies that maximize the total welfare are the solution to*

$$f_l^{FB} = \sqrt{\frac{Q_l u_w}{c_l^s - b_l(f^{FB})}} \tag{20}$$

with $b_l(f^{FB}) = \frac{Qu_w}{X^2(f^{FB})} \frac{\partial \xi_l}{\partial f_l}(f_l^{FB})$ and decreasing in f_l for any $l = 1, \ldots, L$.

Moreover, the first-best frequencies are decreasing functions of bus capacity.

Note that the right-hand side of Equation (21) is a continuous and monotone decreasing function of f, then the solution for f^{FB} exists and is unique.

Equation (21) matches the classical result for optimal frequency: the 'square root formula' (Mohring, 1972). However, it is corrected to take into account the competition and congestion effects at the stop S. Indeed, if the demand at S were zero, the optimal frequency becomes $\sqrt{Q_l u_w / c_l}$, which

corresponds to the optimal frequency in a bus corridor (Mohring, 1972; Jansson, 1980). The deviation from the square root formula depends on the difference between the line l's operation cost and $b_l(f^{FB})$. This function $b_l(\cdot)$ is the marginal benefit of waiting time savings at S due to a line l's frequency increase. The higher the marginal time saving at S, the higher the first-best frequency (every else equal). In turn, the marginal time saving increase with the demand and the expected waiting time at S. Besides, the difference $c_l - b(f^{FB})$ may be interpreted as an operation cost corrected by marginal waiting time benefits at , which leads to a corrected standard square root formula too.

Note that the operation cost in the first-best frequencies must include all externalities produced by buses running in the city. That is, the cost in Equation (21) might not be the same cost faced by the firms when fixing their frequencies. If the operation of buses does not produce externalities, then the firm's operation cost is the right parameter.

Corollary 2. *Under Assumptions 1 to 7, if the ARC model determines the attraction factors, the frequencies that maximize the total welfare are the solution to*

$$f_l^{FB} = \sqrt{u_w Q_l}\left(c_l - u_w Q\left[\sum_{j=1}^{L} f_j^{FB} - \frac{\alpha_j Q_j}{k_j}\right]^{-2}\right)^{-1/2}. \tag{21}$$

3.4. Comparison of operation schemes

First, we compare monopolistic frequencies with competitive ones. In this case, we do not need to make any assumption on the parameter values, such that costs or utility weights. We study the two monopoly cases: direct lines, and hub-and-spoke.

Proposition 3. *The frequencies under competition are higher than those under direct-lines monopoly at all stops O_l with $l \neq m$, where m is such that $c_m = \min\{c_1, \ldots, c_n\}$. Moreover, if the total demand at S is not too large, such that $Q \leq \sum_{l=1}^{L}(1-\alpha_l)Q_l$, then the competitive frequencies are higher than the monopolistic ones at all stops.*

If the demand condition of Proposition 3 does not hold, we cannot compare the frequencies in stops O_m and S without assuming values for the model parameters. For instance, we may assume the total demand at the stop S is lower than the total demand coming from O_l ($l = 1, \ldots, L$) to S, such that $\sum_{l=1}^{L} k_l f_l \geq Q + \sum_{l=1}^{L} \alpha_l Q_l$. This assumption means the total capacity provided by the competing firms is higher than the total demand traveling between S and D. In this case, the frequency at S delivered under competition is higher. However, we can say nothing about the frequency at O_m without additional assumptions. In the next section, we use some parameters estimated with data from Santiago for the comparison of operational schemes. Now, we compare the oligopolistic competition with the hub-and-spoke monopoly.

Proposition 4. *The frequencies under competition are higher than those under hub-and-spoke monopoly in all stops O_l, for all $l = 1, \ldots, L$.*

Again, we can say nothing about the frequency at S without additional assumptions on the parameters. Finally, we compare the first-best frequencies with the monopolistic ones. In the case of direct-lines monopoly, the result is the same as in Proposition 3. In the case of the hub-and-spoke monopoly, the result is analogous to that in Proposition 4.

Proposition 5. *The first-best frequencies are higher than those under direct-lines monopoly at all stops O_l with $l \neq m$, where m is such that $c_m = \min\{c_1, \ldots, c_n\}$. Moreover, if the total demand at S is such that $Q \leq \sum_{l=1}^{L} (1 - \alpha_l) Q_l$, then the first-best frequencies are higher than the monopolistic ones at all stops.*

Proposition 6. *The first-best frequencies are higher than those under hub-and-spoke monopoly in all stops O_l, for all $l = 1, \ldots, L$.*

3.5. Simulation-based comparison of operation schemes

Since the comparison between operation schemes depend on the parameters such as c_l, Q_l, and Q, we adopt some values from Santiago. We simulate the network of Figure 1 and assume the ARC model determines the attraction factors. The simulation allows us to compare the operation schemes in terms of frequency, consumer surplus, firms' profit, and total welfare. Because of the limited space, we do not present the detailed results from simulation but its general conclusions.

We estimate firms' costs and capacity using data from Santiago in 2001 (SECTRA, 2003). At that moment, the bus operation was regulated in price and routes, but the frequency was unregulated. The external costs due to bus operation are adapted from Rizzi and De la Maza (2017). They estimated external cost per kilometer for Santiago including congestion, road damage, road crashes, air pollution, and noise. Table 1 summarizes the information we use to compute frequencies. The waiting time parameter is obtained from SECTRA (2005) and corresponds to three times the value of travel time. The adopted value is higher than the value of travel time because waiting produces more significant discomfort than in-vehicle travel. It is also consistent with Jara-Díaz and Gschwender (2003) and Mohring (1972), who adopt a value of waiting time three times the value of in-vehicle travel time. For the travel utility, we assume that u_0 equals the ticket price plus the value of waiting for a bus with low frequency (2 bus/hour). In the case of HS monopoly, the consumer surplus includes the welfare loss that users undergo because of bus transfer at the hub. In Santiago, recent research shows that transfers metro network reduce passenger's welfare in an amount equivalent to ten minutes of travel time (Raveau et al., 2014). Therefore, we value the bus transfer as ten minutes at the value of time.

We assume that five firms face the same demand at the origin and the same proportion of demand going to D (i.e., α_l). We also assume two demand levels according to the frequency that matches the capacity with the demand. If a high-frequency line operates with a headway of 5 minutes, for the bus capacity equal to 80 passengers, a high level of demand is 960 passengers/hour. In turn, if a low-frequency line operates with headway equal to 10 minutes, the demand is 480 passengers/hour. These levels of frequencies are similar to the observed levels in Santiago in the peak hour morning. The firms are asymmetric in their operation cost. These differences arise from different route lengths. We adopt a route length of 15 kilometers for the segment SD, and routes lengths of 20,

25, 30, 35, and 40 for the segments $O_l S$. We choose lengths to vary the operation cost 33% above and below the mean. The most relevant results of the simulations are the following.[3]

- Frequencies under competition are always the highest at the stops O_l, except when the demand is low, and the operation cost is high. At stop S, the competitive nominal frequency is the highest in all cases, but the effective frequency is the highest only when the demand at S is high.
- If the total demand at S decreases, the competition softens because the strategies tend to be dominated by $f_l = Q_l/k$. Therefore, the firms with high cost operate at the lowest levels $f_l = Q_l/k$, and only the low-cost firms provide capacity above the residual one at the stop S.
- The first-best frequency is greater or equal to the monopolistic frequencies always. In turn, when the demand at O_l is small, and the cost high, the first-best frequency tends to be above the competitive one because the waiting time savings at O_l compensate the costs.

4. EQUILIBRIUM IN BUS CAPACITY

Now, we solve the game in bus capacity. We focus the open-loop equilibrium, which is consistent with pre-commitment in bus capacity. The equilibrium in capacity is characterized by an equation. We do not solve analytically the closed-loop equilibrium because there is no additional insights by doing it. Moreover, we believe the pre-commitment assumption is consistent with bounded rationality that would characterize bus firms operating in cities without public transport regulation. However, we show that the firms have incentives for reducing the bus capacity when they are not dominant and for increasing capacity when they are. Next proposition describe the equilibrium in the capacity competition game.

Proposition 7. *Under Assumptions 1 to 5, if f_l^* is an interior solution for the equilibrium frequency given by Proposition 1, the open-loop equilibrium on bus capacity, k_l^*, is the solution to*

$$(\gamma_0 + \gamma_1 k_l)\,\frac{\partial \xi_l}{\partial k_l}(f_l^*, k_l^*, \alpha_l Q_l) - \gamma_1 f_l^* \frac{\partial \xi_l}{\partial f_l}(f_l^*, k_l^*, \alpha_l Q_l) = 0. \tag{22}$$

Moreover, firms have strategic incentives to increase the bus capacity beyond what they choose in the open-loop equilibrium defined by (23) if their market share is $\lambda_l < 0.5$. If a firm is dominant and its market share is $\lambda_l < 0.5$, then it has incentives to decrease bus capacity.

To study the effect of the model parameters on the equilibrium capacity, we adopt the ARC model for the attraction factors.

[3]To avoid problems with the demand functions at the kink points (where $k_l f_l = Q_l$), we use an approximation with a smooth derivative. Notably, we use $q_l^O(f_l) \approx (-1/2)\,\sqrt{(Q_l - k_l f_l)^2 + 4\delta^2} - (Q_l + k_l f_l)$. This approximation function tends to $\min\{Q_l, k_l f_l\}$ as δ tends to zero (Kanzow, 1996; Facchinei et al., 1999). This way, we do not need to use constrained optimization to obtain competitive and first-best frequencies. We neither need Assumption 6 on the demand level at S to compute the first-best frequencies because we approximate the total demand with the same function at the point $Q = \sum_{l=1}^{L} (1 - \alpha_l) k_l f_l$.

Corollary 3. *Under the same conditions as Proposition 7, if the attraction factors are given by the ARC model and f_l^{ARC} is an interior solution, then the equilibrium capacity is the solution to*

$$k_l^{ARC} = \sqrt{\frac{(\gamma_0 + \gamma_1 k_l^{ARC})\alpha_l Q_l}{\gamma_1 f_l^{ARC}}}. \tag{23}$$

From Corollary 2, on the one hand, the capacity under competition decreases with the demand in the common route segment. This helps to explain the observed reduction in the size of the vehicles in most of liberalized bus service markets (Evans, 1990; Gómez-Ibañez and Meyer, 1997; Estache y Gomez-Lobo, 2005). On the other hand, the equilibrium capacity is decreasing in the demand of the first segment, where lines operate as a monopoly. Corollary 2 also implies that the lower the net revenue per passenger, the smaller the vehicle. Note that if the attraction factor is given by the FS model, the equilibrium capacity is any positive value, which highlight the relevance of including limited bus capacity to explain the trend to reduce bus size in deregulated urban bus markets.

The role of operation cost in the equilibrium capacity is consistent with the maximizing-profit behavior. Indeed, if the fixed operation cost (independent of the capacity), γ_0, is high, firms tend to use big buses, because the effect of changing the bus size is small in the total cost. The opposite effect occurs if the variable operation cost, γ_1, is high. Firms choose small vehicles because that reduces the total cost. It is worth to notice that both effects are independent of the route length.

To compare the competitive capacity with the optimal ones, the next proposition and corollary show the first-best capacity.

Proposition 8. *Under Assumptions 1 to 7, the first-best bus capacity, k_l^{FB}, is the solution to*

$$\left(\gamma_0 + \gamma_1 k_l - \frac{Q_l u_w}{\rho_l (f_l^{FB})^2}\right) \frac{\partial \xi_l}{\partial k_l}(f_l^{FB}, k_l^{FB}, \alpha_l Q_l) - \gamma_1 f_l^{FB} \frac{\partial \xi_l}{\partial f_l}(f_l^{FB}, k_l^{FB}, \alpha_l Q_l) = 0. \tag{24}$$

Corollary 4. *Under the same conditions as Proposition 8, the optimal frequency for the ARC model is the solution to*

$$k_l^{FB} = \sqrt{\frac{\left(\gamma_0 + \gamma_1 k_l^{FB} - \frac{Q u_w}{\rho_l (f_l^{FB})^2}\right) \alpha_l Q_l}{\gamma_1 f_l^{FB}}}. \tag{25}$$

The last corollary shows that the capacity chosen by the firms when competing is lower than the socially desirable depending on the external operation cost and the marginal waiting time saving ($Q_l u_w / (f_l^{FB})^2$). Simulations with data from Santiago show that the competitive capacity is around 40% lower than the first-best one. This result is consistent with the observed trend to reduce bus size in Santiago and other cities in developing countries with liberalized urban bus transport markets.

5. CONCLUSION

Frequency and capacity competition seems to be a plausible explanation for the excess of supply observed in the liberalized or partially regulated market of public transport. Our general result

implies the frequency increases with the demand in the common route segments and with the net revenue per passenger.

In comparison with the first-best frequencies, the competitive frequencies are higher, except when demand is low in the monopolistic route segment. In terms of waiting time, the consumer is better off under competition, even if the total welfare is not the highest. Competition produces more externalities and operation costs, but the bus users directly perceive neither.

In turn, the first best frequencies are higher than the monopolistic ones and similar for low demand levels in the common route segment. Nevertheless, this result depends on the value adopted for the value of waiting time. The higher the value of waiting time, the higher the optimal frequency.

These implications are relevant for the design of contracts for public transport provision. For instance, if the provision of public transportation is placed in the hands of a monopoly firm that operates all the lines in the city, the frequencies will be low. These low frequencies reduce demand in the long term. Consequently, if the contract is a renegotiable or a short-term contract, the operating firm has no incentive to improve quality, such as setting high frequencies. The firm fixes the frequency maximizing profits, and the result is the monopolistic frequencies, as we show in Section 3. The regulator could set the frequency of the lines, but this requires monitoring the operation, which may be costly and technically difficult for big networks.

For instance, Muñoz et al. (2014) report the case Transantiago. The project defined exclusive areas where only two firms operate in a hub-and-spoke scheme with frequencies fixed by the authority with a social criterion, i.e., maximum social welfare. The scheme is such that the firms do not compete: one firm operates the feeder services on the network; the other firm operates the trunk service and few feeder services. In the initial period of the project, there was no frequency monitoring by the authority; therefore, the firms fixed the frequency as a monopoly. The service quality deteriorated dramatically: the waiting time increased, and the buses crowded. After a few months, the authority implemented some monitoring measures which better off the service (Beltran et al, 2011). However, the incentives to operate with low frequencies remain. In a current contract renegotiation, the transport authority is trying to reverse that situation.

The model gives us insights into how we can provide incentives for the firm to provide the desired frequency in routes with a segment where there is competition. Indeed, we show that the equilibrium frequency increases with the net revenue per passenger. Thus, the authority can design a contract with a payment per passenger (or proportional to patronage) and a lump-sum transfer in order to cover total costs. With such a contract, it would be possible to control the frequency by changing the payment per passenger.

REFERENCES

Beltran, P., Gschwender, A., and Palma, C. (2011). The impact of compliance measures on the operation of a bus system: the case of Transantiago. 12th International Conference on Competition and Ownership in Land Passenger Transport (Thredbo12), Durban, South Africa.

Bouzaiene-Ayari, B. (1988) Modélisation des arrêts multiples d'autobus pour les réseaux de transport en commun, M.S. thesis, Département de Génie Industriel, École Polytechnique de Montréal.

Bouzaiene-Ayari, B., Gendreau, M. and Nguyen, A. (1998) Passenger assignment in congested transit networks: a historical perspective, in P. Marcotte, S. Nguyen (Eds.), Advanced Transportation Modelling, Kluwer Academic Publishers, Massachusetts, pp. 47–71.

Bouzaiene-Ayari, B., Gendreau, M. and Nguyen, A. (2001) Modeling Bus Stops in Transit Networks: A Survey and New Formulations, Transportation Science, Vol. 35(3), pp. 304–321.

Brueckner, J. K., and Flores-Fillol, R. (2007). Airline schedule competition. Review of Industrial Organization, 30(3), 161-177.

Bulow, J. I., Geanakoplos, J. D., and Klemperer, P. D. (1985). Multimarket oligopoly: Strategic substitutes and complements. Journal of Political economy, 93(3), 488-511.

Cominetti, R. and Correa, J. (2001) Common-Lines and Passenger Assignment in Congested Transit Networks, Transportation Science, Vol. 35(3), pp. 250–267.

Douglas, G.W. and Miller, J.C. (1974) Quality Competition, Industry Equilibrium, and Efficiency in the Price-Constrained Airline Market, American Economic Review, Vol. 64 (4), pp. 657-669

Estache, A. and A. Gómez-Lobo (2005) The limits to Competition in Urban Bus Services in Developing Countries, Transport Reviews, 25(2), 139–58.

Evans, A. (1987) A theoretical comparison of competition with other economic regimes for bus services, Journal of Transport Economics and Policy, 21(1), pp. 7–36.

Evans, A. (1990) Competition and the Structure of Local Bus Markets, Journal of Transport Economics and Policy, Vol. 24(3), pp. 255-281.

Facchinei, F., Jiang, H., and Qi, L. (1999). A smoothing method for mathematical programs with equilibrium constraints. Mathematical programming, 85(1), 107.

Fernández, J.E. and Muñoz, J.C. (2007) Privatisation and Deregulation of Urban Bus Services: An Analysis of Fare Evolution Mechanisms, Journal of Transport Economics and Policy, Vol. 41(1), pp. 25-49.

Fielbaum, A., Jara-Diaz, S., and Gschwender, A. (2016). Optimal public transport networks for a general urban structure. Transportation Research Part B: Methodological, 94, 298-313.

Fudenberg, D., and Tirole, J. (1984). The fat-cat effect, the puppy-dog ploy, and the lean and hungry look. The American Economic Review, 74(2), 361-366.

Fudenberg, D., and Tirole, J. (1991). Game theory. MIT press.

Gagnepain, P., Ivaldi, M., and Muller-Vibes, C. (2011). The industrial organization of competition in local bus services. In A handbook of transport economics. Edward Elgar Publishing.