The impact of a financial constraint on the spatial structure of public transport services.

Sergio R. Jara-Díaz*, Antonio Gschwender and Meisy Ortega
Universidad de Chile
*Corresponding author. jaradiaz@ing.uchile.cl

Abstract

Using a single line model, it has been shown recently that the presence of a stringent financial constraint induces a less than optimal bus frequency and larger than optimal bus size. This occurs because the constraint induces a reduction of the importance of users' costs (their time); in the extreme, users' costs disappear from the design problem. In this paper we show that such a constraint also has an impact on the spatial structure of transit lines. This is done departing from the single line model using an illustrative urban network that could be served either with direct services (no transfers) or with corridors (transfers are needed). First, the optimal structure of lines is investigated along with frequencies and vehicle sizes when the full costs for users and operators are minimized (unconstrained case); the optimal lines structure is shown to depend upon the demand level, the values of time and the cost of providing bus capacity. Then the same problem is solved for the extreme case of a stringent financial constraint, in which case users' costs have relatively little or no effect in determining the solution; in this case the preferred outcome would be direct services under all circumstances, with lower frequencies and larger bus sizes. The impact of the financial constraint on the spatial structure of transit lines is shown to be caused by the reduction in cycle time under direct services; the introduction of users' costs in the objective function makes waiting times reverse this result under some circumstances.

Keywords: public transport, lines structure, design, financial constraint.

1. INTRODUCTION

There is an emerging discussion regarding the financial aspects, property and contracts in the provision of public transport services. By the early eighties it had been shown that the optimal operation of a public transport system is linked to an optimal price that falls below average cost, which induces an optimal subsidy (Jansson, 1979, 1984). This has received recent criticism; for example, van Reeven (2008) develops a model aimed at showing that a profit-maximizing operator allowed to take into account the demand effects of its pricing would offer a frequency at least as high as a welfare-maximizing one with no welfare losses; later on, Basso and Jara-Díaz (2010) showed that this result depends crucially on demand inelasticity. By the same period Parry and Small (2009) concluded that in most of the real cases they analyze, increasing transit subsidies would increase welfare although subsidies already cover a large proportion of operators' cost. Presently, the financial aspects of public transport seem to dominate over optimal pricing and welfare, which makes Jansson's (2005) question relevant: "Why is optimal bus transport pricing applied in hardly any urban area of the world?" To this we add that the link between the financial aspects and the design of the public transport system has been absent from the debate.

What has been observed in transit systems is that fares and subsidies are usually determined outside the technical domain, not always accounting for the impact on the main design variables: frequency of services, vehicle sizes and spatial coverage. This translates into a financial constraint on the design of a public transport service, which has been analyzed by Jara-Díaz and Gschwender (2009) by means of a microeconomic analysis of a single transit line. They showed that imposing such a constraint leads to a decrease in the relative weight of users' time in the cost function through the hidden reduction in the weight given to their time values in the associated optimization problem. Analytically those time values get divided by one plus the multiplier of the constraint, which makes users' cost weigh less relative to

operators', causing lower frequencies and larger buses in comparison to the optimal values in the absence of a financial constraint. This was offered as a theoretical explanation for the resulting fleet reduction and use of larger vehicles in the redesign of the bus services in Santiago, Chile, where a self-financial constraint was imposed while keeping the previous average fare; this had a very negative impact on service quality and users' costs. In this paper we want to examine the theoretical effect of a stringent financial constraint on a third most important component from a strategic viewpoint: the spatial structure of transit lines.

This spatial aspect of the design cannot be studied using a single line model and requires extension to a network. In real (urban) cases, the transit network design problem has been based mostly on heuristics (Kepaptsoglou and Karlaftis, 2009), such that a generic design of routes is usually adapted incrementally following reasonable procedures. Here we will deal with the generic design at a strategic level for policy analysis; from this viewpoint, the spatial dimension has been sometimes introduced as a continuous design variable - subject to optimization - in the form of some measure of the (regular) spacing between consecutive lines, as done by Chang and Shonfeld (1991), who considered the distance on a rectangular grid, or Tirachini et al (2010), who considered the angle on a circular city. Analyzing lines structure, though, requires a departure from these continuous approaches where each line operates similarly. There are two meaningful alternative spatial designs that can be used to represent real generic structures. In the first structure users are served mostly with direct lines that follow closely the spatial pattern of demand, which makes transfers on the main OD pairs unnecessary but present route overlapping along the main corridors; this structure has been present in many capital cities in South America. An alternative option is to design a set of bus lines such that users can make the necessary transfers to reach the corresponding destinations; this type of bus lines structures relying on transfers and avoiding overlapping are typically observed in European metropolitan areas. However, it is not evident which one is better.

Departing from Mohring's (1972) and Jansson's (1984) single line transit cost analysis, Jara-Díaz and Gschwender (2003b) considered several lines in a network, introducing the choice between direct services – without transfers – and corridors where transshipments are necessary. They studied these alternative structures aimed at minimizing total costs (users and operators), showing that the outcome might depend on patronage. When patronage is relatively low, the ''full coverage'' of direct lines may be neither in the interest of the bus company nor in that of the passengers because of the low frequencies that would very probably result. However, if patronage is large enough it may well happen that direct services can operate with sufficiently high frequencies and avoiding transfer time.

As explained earlier, we want to study the effect of a stringent financial constraint on the spatial structure of services. This will be done by comparing the total cost function, i.e. the minimization of users' plus operators' costs for exogenously given patronage levels (optimal design benchmark), against the operators' cost function, i.e. the minimum of operators' costs only, which has been shown to be equivalent to the extreme case caused by a stringent financial constraint. The question, then, is how sensitive the optimal spatial structure of lines is – along with frequency and bus size - to the consideration of users' costs (time). To answer this we analyze a spatial demand structure on a simple but representative network, searching not only for frequencies but also for the lines structure and vehicle sizes that minimize a) total cost (users and operators) and b) operators' cost only. Results are comparatively presented, including service structures, fleet sizes needed, in-vehicle travel times and waiting times. It is shown that the best structure differs depending on the inclusion of users' costs in the objective and varies with the demand level.[1]

[1] The issue of service structure has also been analyzed by Jara-Díaz and Basso (2003) in a three nodes network in relation with economies of spatial scope, showing that for the case of equal flows between each of the six origin-destination pairs and equal distances, direct services are less costly for an operator than a hub-and-spoke structure. This type of discussion resembles that in air transport regarding the use of hubs (inducing transfers) versus fully connected networks (direct services; no transfers needed) for profit maximizing and socially optimal airlines. For example, Hendricks et al (1995) show that an unregulated airline might choose either structure depending on various elements including demand level. Using a simple network structure Brueckner (2004) shows that a monopolistic airline would be biased in favor of the

As background, in Section 2 we explain the essence of how a financial constraint operates diminishing the importance of users´ costs in a single line case. Then in Section 3 we add the spatial aspect of design in a representative network where the alternative lines structures are two corridor lines with transfers or four direct lines without transfers. As the (optimal) unconstrained case is a benchmark, it is developed there in order to show the general results. The case with the financial constraint is presented and discussed in Section 4 in order to emphasize that there is an effect on lines structures (also on frequency and bus size). Section 5 concludes.

2. BACKGROUND: FINANCIAL CONSTRAINT IN THE ONE LINE CASE

Following Jansson (1980, 1984), Jara-Díaz and Gschwender (2009) analyzed total cost minimization (i.e. users' plus operators' costs) for a public transport corridor used by a total of Y passengers per hour homogeneously distributed along the corridor, all of them traveling a fraction β of the corridor's length. Vehicles operate at a frequency f. Defining T as the time in motion of the vehicle in a cycle[2] and t as the time that a passenger needs to board or alight, cycle time t_c is given by $t_c = T + 2t(Y/f)$. As frequency is the ratio between fleet size (B) and cycle time then $B = fT + 2tY$. The cost per vehicle-hour for the operator is given by $c = c_0 + c_1K$, where c_0 and c_1 are constants and K is vehicle size. The users' values of in-vehicle and waiting times are P_v and P_w respectively. They impose a financial constraint that restrains the operators' cost to a maximum of A, exogenously given because of, say, budgetary reasons or general policy (e.g. an exogenously imposed fare and no subsidies). As the total value of the resources consumed VRC increases with K (the derivative of VRC with respect to K is

hub-and-spoke structure and would choose lower than optimal frequencies and aircraft size. Pels et al (2000) conclude that "a fully connected network will be more profitable if the level of demand is relatively high, fixed costs are low and economies of density are low".

[2] Time in motion T includes time for acceleration/deceleration, time to open and close the doors and any other component of the cycle time different from the time at stop for boarding and alighting purposes.

positive), K is equal to the resulting load size, which depends on the optimisation variable f, i.e.

$$K = k(f) = \frac{Y}{f}\beta \ . \tag{1}$$

Then the restricted social optimisation problem is (Jara-Díaz and Gschwender, 2009)

$$\begin{aligned} &\underset{f}{\text{Min}} VRC = \left(fT + 2tY\right)\left(c_0 + c_1\frac{Y}{f}\beta\right) + P_w\frac{1}{2f}Y + P_v\left(T + 2t\frac{Y}{f}\right)\beta Y \\ &\text{subject to} \quad \left(fT + 2tY\right)\left(c_0 + c_1\frac{Y}{f}\beta\right) - A \le 0 \end{aligned} \tag{2}$$

As the service is assumed to have predetermined bus stops location, access time cannot be optimized and is not included in equation (2).

If μ is the multiplier of the financial constraint, then the frequency $\tilde{f}$ and bus size $\tilde{K}$ resulting from problem (2) obtained by Jara-Díaz and Gschwender (2009) are

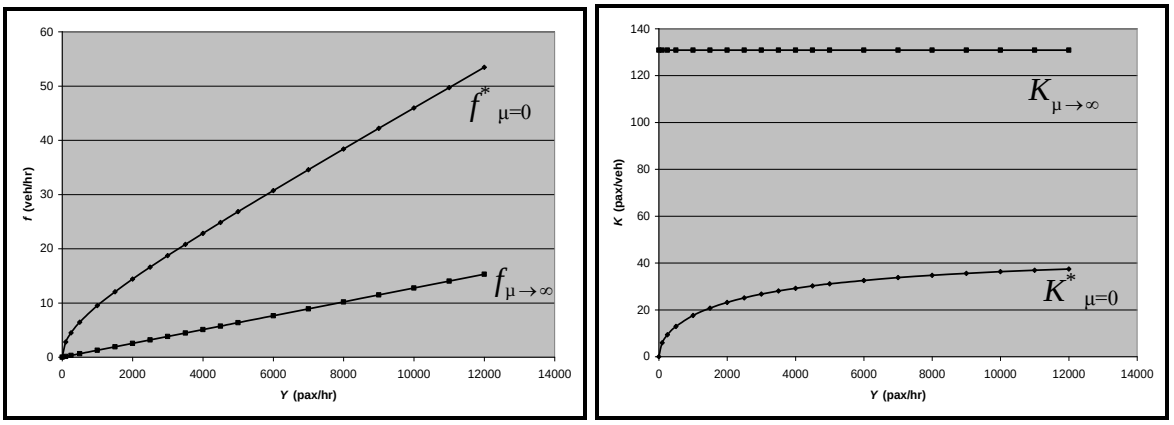
$$\tilde{f} = \sqrt{\frac{Y}{Tc_0}\left(\frac{1}{2}\frac{P_w}{(1+\mu)} + 2tY\beta\left[\frac{P_v}{(1+\mu)} + c_1\right]\right)} \ , \tag{3}$$

$$\tilde{K} = \frac{l}{L}\sqrt{Tc_0Y\Big/\left(\frac{1}{2}\frac{P_w}{(1+\mu)} + 2tY\beta\left[\frac{P_v}{(1+\mu)} + c_1\right]\right)} \ . \tag{4}$$

As shown in Jara-Díaz and Gschwender (2009), the multiplier μ increases as A diminishes. This means that the tighter the budget, the larger is μ, diminishing the role of time values on both frequency and bus size. Two extreme cases can be identified. First, when the financial constraint is not active (μ=0) the unconstraint optimal frequency f* and optimal vehicle size K* are obtained. Second, for $\mu \to \infty$ (which occurs when A is set exactly at the minimum operators' cost for each Y level), the frequency and bus size obtained corresponds to the minimization of operators cost only because all terms with values of time disappear. Figure 1

shows the frequency and bus size for both extreme cases. An intuitive interpretation is that "any given passenger volume can be served with different combinations of frequency and vehicle size, but users' costs would be lower for high frequency-small vehicles combinations while operators' costs are favoured by low frequency-large vehicles combinations, up to a limit" (Jara-Díaz and Gschwender, 2009, p69).

Figure 1: Frequency and vehicle size as a function of the number of passengers (*Y*), for both extreme cases of the financial constraint



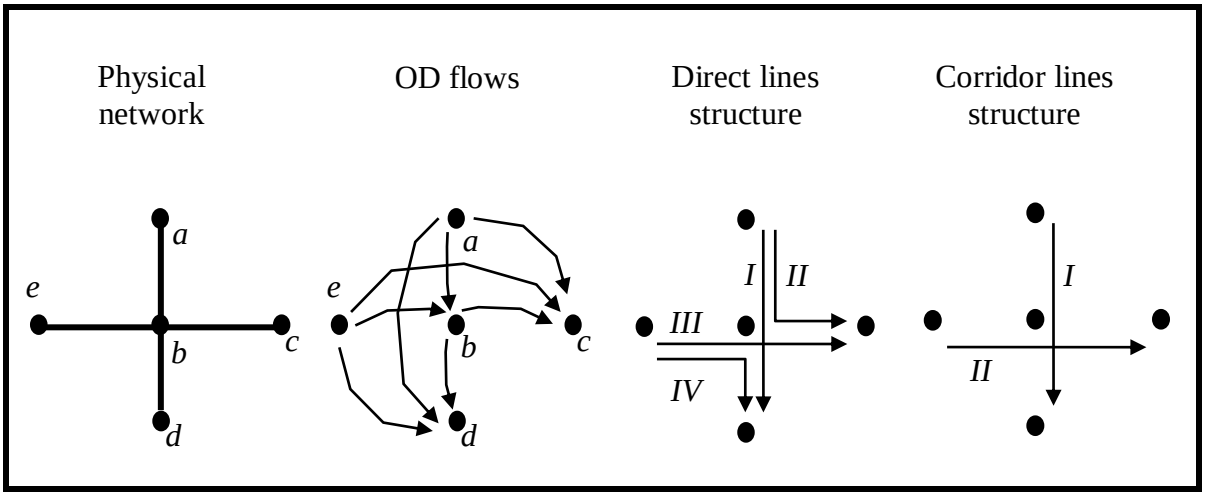
Source: Jara-Díaz and Gschwender, 2009.

In summary, imposing a financial constraint acts on the optimal design diminishing frequency and increasing bus size for all levels of demand. This happens because the constraint operates in such a way that it is equivalent to reduce the importance of users' time in the design problem. Does a financial constraint also affect the spatial structure of transit lines? In order to study this, we will analyze the convenience of different lines structures to serve a given demand pattern on a network for the two extreme cases presented above. First, we will find the best lines structure for a total cost minimization objective as an extreme case in which there is no financial restriction at all and $\mu=0$; second, we will obtain the lines structure that minimizes operators cost only as the case in which the financial constraint is extreme such that $\mu \to \infty$ and users cost are ignored.

3. DESIGN OF LINES STRUCTURES ON A NETWORK: METHODOLOGY AND APPLICATION TO A CASE WITHOUT FINANCIAL CONSTRAINT

The spatial structure of transit lines will be analyzed solving the design problem for two basic lines structures on the simple but representative network presented in Figure 2. Following Jara-Díaz and Gschwender (2003b), we will consider the *direct lines* structure, which links every OD pair such that users need no transfers, and the *corridor lines* structure, which tries to minimize the total length of the lines, forcing transfers in some OD pairs. Both line structures cover the same network and therefore do not affect access time, which is then irrelevant in the optimization. Unlike Jara-Díaz and Gschwender (2003b), operators' cost will depend linearly on vehicle sizes, which becomes a design variable that adds to frequencies and lines structure.

Figure 2: Direct and corridor lines structures.



Let us begin with the unconstrained case of total cost minimization (users and operators); the case of operators' cost only will be presented in the next section. The procedure is as follows: first we search for the optimal fleets and vehicle sizes for each and every line conditional on a lines structure; second, we compare the minimum conditional costs, obtaining the overall optimal lines structure. To do this, we consider that:

- The total demand entering the system is Y passengers per hour distributed equally among the OD pairs, i.e. $Y/8$ passengers on each pair.
- Cycle and in-vehicle times are affected by passengers boarding and alighting times.
- Boarding and alighting occurs sequentially at all available doors.
- Operators' cost depends linearly on vehicle size.
- Every line uses only one type of vehicles (equal vehicle size within each line).
- Waiting time is a proportion ε of the headway (ε=0.5 if buses and passengers arrive regularly, which is assumed in the numerical simulations of the appendixes).

Let $T_0/2$ be the vehicle travel time between two consecutive nodes of the network in one direction without boarding and alighting times, and let t be the time that a user needs to board or alight. The cycle time t_c for each line has two components: time in motion - given by $2T_0$ - and the time at the stops where users board and alight – given by $2t$ times the number of users that board (and alight) a vehicle in a cycle. The number of users that board is composed by two groups: those that board at the origin and those that board in transfers. In the system, the number of transfers is given by the combination of the demand structure and the lines structure, which is 0 for direct lines and $Y/4$ for corridors, as only two OD pairs need a transfer (pairs *ac* and *ed*). Therefore the total number of passengers boarding is $Y(1+\tau)$, where τ is the average number of transfers per trip in each structure (0 and ¼ for direct and corridors, respectively). As the problem is symmetric regarding both the demand pattern and the lines structure, the frequency of service of each line, f_i (and the fleet sizes B_i), will be the same for all lines within a structure (but different among structures) and passengers boarding are equally distributed for each structure among lines. Then the number of passengers that board and alight on each vehicle cycle is $Y(1+\tau)/f_iN$, where N is the number of lines for each structure (2 for corridor, 4 for direct).

Therefore, cycle time for each line i is

$$t_{ci} = 2T_0 + 2t \cdot \frac{Y(1+\tau)}{f_i N} \tag{5}$$

Frequency of line i is the ratio between the fleet of the line and its cycle time:

$$f_i = \frac{B_i}{2T_0 + 2t \cdot \frac{Y(1+\tau)}{f_i N}} \quad , \tag{6}$$

which yields

$$f_i = \frac{1}{2T_0 N}\left[NB_i - 2tY(1+\tau) \right] \quad \text{and} \quad B_i = 2T_0 f_i + 2t \cdot \frac{Y(1+\tau)}{N}. \tag{7}$$

$$B = NB_i = 2T_0 N f_i + 2tY(1+\tau) \quad . \tag{8}$$

The total value of the resources consumed is the sum of operators and users costs. Operators' cost is ($c_0 + c_1K$) times the total fleet size from equation (8). Note that K is the ratio between the maximum load on a line $(3Y/8)/(N/2)$ and f_i .Therefore

$$VRC = (c_0 + c_1 \frac{3Y}{4Nf_i})\left[2T_0 N f_i + 2tY(1+\tau) \right] + P_w t_w + P_v t_v \quad . \tag{9}$$

Let us obtain users' costs. Waiting time has two components: waiting at the origins and waiting at transfers. From the origins $Y/2$ passengers move two nodes and $Y/2$ move one node; long distance passengers wait ε/f_i in either lines structure, but short distance passengers can use $N/2$ lines to move one arc such that their waiting time is $\varepsilon/(N/2)f_i$. Passengers that transfer are τY and each waits $\varepsilon/(N/2)f_i$ at the transfer point. Then total waiting time t_w is

$$t_w = \underbrace{\frac{Y}{2}\frac{\varepsilon}{f_i}}_{\substack{\text{at origins,} \\ \text{long distance}}} + \underbrace{\frac{Y}{2}\frac{\varepsilon}{(N/2)\, f_i}}_{\substack{\text{at origins,} \\ \text{short distance}}} + \underbrace{\tau Y \frac{\varepsilon}{(N/2)\, f_i}}_{\text{at transfer points}} = \frac{Y\varepsilon}{f_i}\left(\frac{1}{2} + \frac{1}{N} + \frac{2\tau}{N} \right) \quad . \tag{10}$$

In-vehicle time for a passenger has three components: time in motion, time in the vehicle due to boarding and alighting of other passengers and time alighting. The first one is always T_0 for

$Y/2$ (long distance) passengers and $T_0/2$ for the remaining half, irrespective of the lines structure; this makes a total of $(3/4)T_0Y$.

Regarding time alighting we have two cases; the six OD flows that end at *c* or *d* and the remaining two OD flows that end at *b*. Each of the $(3/4)Y$ passengers that end their trip at *c* or *d* have to alight as part of a group of $(3/8)Y/(N/2)\ f_i$ passengers per vehicle, because $(3/8)Y$ passengers arrive and alight at either *c* or *d* using $N/2$ lines that operate at a frequency f_i. As the first passenger alights immediately and the last has to wait $t(3/8)Y/(N/2)\ f_i$, the average alighting time is half this total. Therefore, the total alighting time at *c* and *d* is

$$t_{Ac-d} = \frac{t}{2}\frac{3Y}{4}\frac{3Y}{8}/(N/2)f_i = \frac{9}{32}\frac{tY^2}{Nf_i} \quad , \tag{11}$$

Passengers alighting at *b* are of two types: those that end the trip there and those that transfer, such that the total is $(Y/4)+\tau Y$. This total comes from two origins (*a* and *e*) using $N/2$ lines from each, operating at a frequency f_i such that the average alighting time is $(t/2)((Y/8)+\tau Y/2)/(N/2)\ f_i$. Then the total alighting time at *b* is

$$t_{Ab} = \left(\frac{Y}{4}+\tau Y\right)\frac{t}{2}\left(\frac{Y}{8}+\frac{\tau Y}{2}\right)/\left(\frac{N}{2}\right)f_i = \frac{tY^2(1/4+\tau)^2}{2Nf_i} \quad . \tag{12}$$

The final component of in-vehicle time is the time at stops due to the boarding and alighting of other passengers. Note first that this delay is experienced neither by the $Y/2$ short-distance travelers nor by the τY that have to transfer, and this two groups are the ones that actually cause the delay on the remaining $Y-(Y/2)-\tau Y$ (all of them long distance) travelers.

The best way to understand this type of delays is to look at a flow like *a-d* in both structures. In the **direct lines structure**, these passengers use line *I* and suffer the alighting of the short distance travelers *a-b* that split into the $N/2$ lines that serve that link (*I* and *II*), which makes $(Y/8)/(N/2)$. The same number of passengers board at *b* in line *I* to go to *d*. Therefore, a total of $(Y/2N)/f_i$ board and alight each vehicle of line *I* at *b*, causing a total delay $t(Y/2N)/f_i$. This

same analysis holds for the other three long-distance OD flows in the direct lines structure. In the **corridors lines structure** short distance passengers do not split ($N/2=1$) and passengers in line *I* experience the (additional) alighting of passengers that go from *a* to *c* and the boarding of those traveling from *e* to *d*, which adds up to all passengers that make a transfer, τY.

Therefore, the delay experienced by each of the $(Y/2)$-τY passengers identified above can be expressed as $t[(Y/2N)+ \tau Y]/f_i$ for both structures. Then the total delay for passengers in-vehicle due to other passengers boarding and alighting, t_D, is given by

$$t_D = \frac{tY^2}{f_i}\left(\frac{1}{2}-\tau\right)\left(\frac{1}{2N}+\tau\right) \quad . \tag{13}$$

The total in vehicle time t_v is obtained adding time in motion $(3/4)T_0Y$ plus the results (11), (12) and (13), which yields

$$\begin{aligned} t_v = & \underbrace{\frac{3}{4}T_0Y}_{\text{time in motion}} + \underbrace{\frac{tY^2}{2Nf_i}\left[\frac{9}{16}+\left(\frac{1}{4}+\tau\right)^2\right]}_{\text{time alighting}} + \underbrace{\frac{tY^2}{f_i}\left(\frac{1}{2}-\tau\right)\left(\frac{1}{2N}+\tau\right)}_{\substack{\text{delays due to other passengers} \\ \text{boarding and alighting}}} \\ = & \frac{3}{4}T_0Y + \frac{tY^2}{2Nf_i}\left[\frac{9}{8}+\frac{\tau}{2}\left(2N-1\right)\left(1-2\tau\right)\right] \end{aligned} \quad , \tag{14}$$

Note that alighting time is explicitly included in t_v, while boarding time is implicitly included in the waiting time as it is taken as a proportion ε of the headway between buses, and this headway includes time at the bus stop. Nevertheless, as explained above, boarding time of other users affecting passengers that boarded in a previous stop are considered, because they do impact on travel time. Finally, note that equations (10) and (14) for t_w and t_v respectively are general expressions for waiting and in-vehicle times as functions of any given frequency (optimal or not), such that we observe both the effect caused by the parameters that define a line structure (N and τ) and the effect of frequency. This will be shown to be relevant in the discussion on the best line structures.

Now we have all the elements to minimize *VRC* with respect to f_i. Replacing equations (10) and (14) in (9), first order conditions yield optimal frequency as

$$f_i^* = \frac{1}{2N}\sqrt{\frac{Y}{c_0 T_0}}\sqrt{tY\left[3c_1(1+\tau)+P_v\left(\frac{9}{8}+\frac{\tau}{2}(2N-1)(1-2\tau)\right)\right]+P_w\varepsilon\left[N+2(1+2\tau)\right]} \quad . \quad (15)$$

Equation (15) is a general expression for the optimal frequency of all lines in either structure. From this one can see directly that f_i^* decreases with *N* and increases with τ within the range analyzed, which unambiguously show that frequency is lower for each of the four direct lines than each of the two corridor lines, as expected. This property does not translate into the optimal fleet size or the optimal bus size. The former is obtained by replacing (15) in (8), which yields

$$B^* = 2tY(1+\tau)+\sqrt{\frac{T_0 Y}{c_0}}\sqrt{tY\left[3c_1(1+\tau)+P_v\left(\frac{9}{8}+\frac{\tau}{2}(2N-1)(1-2\tau)\right)\right]+P_w\varepsilon\left[N+2(1+2\tau)\right]} \quad (16)$$

while the optimal bus size is given by

$$K_i^* = \frac{3Y}{4Nf_i^*} = \frac{\frac{3}{2}\sqrt{c_0 T_0 Y}}{\sqrt{tY\left[3c_1(1+\tau)+P_v\left(\frac{9}{8}+\frac{\tau}{2}(2N-1)(1-2\tau)\right)\right]+P_w\varepsilon\left[N+2(1+2\tau)\right]}} \quad . \quad (17)$$

The *VRC* can be written as a function of f_i replacing t_w (10) and t_v (14) in (9), which yields

$$VRC = 2c_0 tY(1+\tau)+4f_i N c_0 T_0+\frac{3T_0 Y}{2}\left(c_1+\frac{P_v}{2}\right) \quad . \quad (18)$$

Replacing optimal frequency (15) yields the cost function as

$$C^* = 2c_0 tY(1+\tau)+2\sqrt{c_0 T_0 Y}\sqrt{tY\left[3c_1(1+\tau)+P_v\left(\frac{9}{8}+\frac{\tau}{2}(2N-1)(1-2\tau)\right)\right]+P_w\varepsilon\left[N+2(1+2\tau)\right]} + \frac{3T_0 Y}{2}\left(c_1+\frac{P_v}{2}\right) \quad (19)$$

From the total generic cost function (19) and the values of τ and N, the optimal structure can be found by comparison. The third term is equal for both direct and corridor lines and cancels out. As shown numerically in Appendix 1 the first term of equation (19) is negligible with respect to the second term, which allows analytical comparison using only the square root. This yields that **the total cost of direct lines is lower when**[3]:

$$\frac{tY}{\varepsilon P_w}\left(c_1 + 0.25P_v\right) > \frac{4}{3}. \tag{20}$$

The probability of direct lines being the more convenient structure increases with the size of tY and with the ratios $c_1/\varepsilon P_w$ and $P_v/\varepsilon P_w$ (which is consistent with Jara-Díaz and Gschwender, 2003b). The intuition behind this is related with the relative importance of waiting and in-vehicle times in each structure (including their prices) and with operators' costs. To discuss this, it is convenient to examine first the lines structure that emerges when the financial constraint is stringent, i.e. when users' costs are (implicitly) dismissed, which we present in the next section. Using this as the point of departure to understand the role of users' time will be proved to be particularly useful.

4. DOES A STRINGENT FINANCIAL CONSTRAINT AFFECT THE SELECTION OF LINES STRUCTURE?

As explained above, in order to analyze if a financial constraint has an impact in the selection of the spatial structure of transit lines, we will consider now the other extreme case of the constraint multiplier, $\mu \to \infty$, which brings to zero the contribution of the values of time, i.e. only operators costs are minimized. Solving the new problem, the frequency, fleet and vehicle size that minimize operators' expenses only are

[3] For the numerical example in Appendix 1, the level of Y that makes the total cost of both structures equal using the approximation behind eq. (20) is 13.8% larger than the exact value when eq. (19) is used.

$$f_i^{\bullet} = \frac{Y}{2N}\sqrt{\frac{3c_1}{c_0 T_0}t\left(1+\tau\right)} \quad . \tag{21}$$

$$B^{\bullet} = Y\left[2t\left(1+\tau\right)+\sqrt{\frac{3T_0 c_1}{c_0}t\left(1+\tau\right)}\right] \tag{22}$$

$$K_i^{\bullet} = \sqrt{\frac{3c_0 T_0}{4c_1 t\left(1+\tau\right)}} \tag{23}$$

which can be also obtained imposing zero time values ($P_w = P_v = 0$) in equations (15) to (17). Then the minimum operators' expense for each line structure is

$$C^{\bullet} = Y\left[2c_0 t\left(1+\tau\right)+2\sqrt{3c_0 c_1 T_0 t\left(1+\tau\right)}+\frac{3T_0 c_1}{2}\right] \tag{24}$$

The comparison of expression (24) for both line structures yields that **direct lines are always better than corridors in this extreme case in which users costs are ignored**. Unlike the no-financial-constraint case, where the sum of users and operators cost is minimized, now corridor lines are never the best structure. This implies that a financial constraint does affect the selection of the best spatial structure of transit lines.

Interestingly, the result of direct lines being always better than corridor lines when only operators costs are taken into account contradicts the intuition of Jara-Díaz and Gschwender (2003b, p276), who stated: "What would be the best spatial structure of services if users' costs were not taken into account? Clearly, in that case direct services would never be an undoubtedly superior solution (at most, a tie)." Our new analyses correct this erroneous intuition. What happens is that avoiding transshipments - represented by τ - diminishes not only cycle time but also fleet size. This is clearly shown by expression (22), where fleet size increases with τ, which is nil for direct lines. Moreover, expression (23) shows that vehicle size decreases with τ. For short, direct lines with no transshipments imply a lower fleet of larger buses, which reinforces the result represented by Figure 1 for the single line case and,

as discussed earlier, induces a lower cost for every demand level because large vehicles are cheaper per place and capacity is adjusted to demand (vehicles always full). This is what lies behind the lower operators' cost for direct lines in equation (24) where transshipments play the key role.

What would be the impact on users' cost - which has been ignored in this design – of implementing what is best for the operators only? Waiting time and in-vehicle time - generically shown in equations (10) and (14) respectively - can be now evaluated at the frequency that minimizes operators' cost in equation (21), which yields

$$t_w^{\bullet} = \varepsilon\left(N + 2 + 4\tau\right)\sqrt{\frac{c_0 T_0}{3c_1 t\left(1+\tau\right)}} \tag{25}$$

$$t_v^{\bullet} = \frac{3}{4}T_0 Y + Y\sqrt{\frac{t c_0 T_0}{3c_1\left(1+\tau\right)}}\left[\frac{9}{8} + \frac{\tau}{2}\left(2N-1\right)\left(1-2\tau\right)\right] \tag{26}$$

Note that total waiting time is, in this case, independent of patronage; this evidently occurs because waiting time is inversely related with frequency which in turn increases linearly with *Y*. Evaluating equation (25) yields that, surprisingly, **waiting time is always lower in corridors**, which means that the effect of a larger frequency (eq. 21) dominates over the effect of transshipments, a very interesting result indeed. These two effects can be seen in the generic waiting time equation (10), where the line structure effect ($1/2 + 1/N + 2\tau/N$) is larger for corridors (5/4) than for direct lines (3/4) because of the mandatory transfers, but the frequency effect reverses the result of the comparison. Analogously, evaluating equation (26) at the corresponding values of τ and N we confirm that **when only operators' cost are minimized, in-vehicle time is always lower in direct lines**, which was expected as t_v is directly linked with cycle time.

So, for synthesis, when only operators´ costs matter because of a financial constraint, the preferred design corresponds to a direct lines structure with a (relatively) small fleet of

(relatively) large buses, with a negative impact on users' waiting time. Let us analyze how this changes when users' costs enter the picture.

Let us take this case of an extreme financial constraint - where direct lines are always the best - as the point of departure to understand why introducing users' cost affects the best line structure (and the other design variables). As waiting time is lower for the corridor structure, one might think that when users' costs are considered corridors could become the best structure whenever waiting time dominates over in-vehicle time (which is larger in corridors) plus operators' costs. So a relevant question is how in-vehicle and waiting times vary when the design follows total cost minimization. Is waiting time still lower (and in-vehicle time larger) for corridors under this objective? Let us examine this.

The expressions for the waiting and in-vehicle times are obtained replacing optimal frequency from equation (15) into equations (10) and (14) respectively. Evaluating these expressions for direct and corridor lines, it can be shown (Appendix 2) that in this case without financial constraint **corridor lines always have the lowest waiting time** and **the largest in-vehicle time**, just as in the case of the extreme financial constraint and for the same reason**:** in spite of the transfers needed in the corridor structure (lines structure effect) total waiting time is lower than in direct lines because the frequencies that passengers observe are higher (frequency effect).[4] On the other hand, in-vehicle time is larger for corridor lines, because transfers imply a larger number of passengers boarding and alighting, increasing time at bus stops and cycle times. This explains the role played by the waiting time value P_w in the total cost minimizing condition (20) for the best lines structure. Larger P_w values decrease the probability of direct lines being the best ones, and this happens because, as we have shown, waiting times are always lower for corridors. Note that the contrary happens with the size of P_v.

[4] It is worth noting that in our model transfers produce only additional waiting time. Neither the negative perception of transfers nor additional walking time is considered.

Corridor lines can be superior for low levels of demand when there is no financial constraint (total cost is minimized), according to result (20). Why is this? We do know what happens with waiting time (lower for corridors) and in-vehicle time (lower for direct). Regarding operators' costs - obtained by replacing f^* in the first term of equation (9) – it can be shown that neither direct nor corridor lines structure are systematically superior. As shown numerically in Appendix 3, for low values of Y the difference in waiting time dominates over the differences in the other two components, even in a region where operators' cost is lower for direct lines. For short, for low demands each direct line results in low frequencies yielding large waiting times, which changes the optimal structure towards corridor lines when users' costs are taken into account: the waiting time effect dominates.

For completeness, let us analyze optimal fleet and vehicle size under each lines structure. The comparison of the fleet sizes using equation (16) is similar to the one made in the comparison of the total cost (19): the first term is much smaller than the second one and therefore the square root can be used for analytical comparison, yielding the same conditions described in (20): increasing tY, $P_v/\varepsilon P_w$ or $c_1/\varepsilon P_w$ increases the probability of direct lines having the lowest fleet. Regarding vehicle size in equation (17) the conclusion is that increasing tY, $P_v/\varepsilon P_w$ or $c_1/\varepsilon P_w$ increases the probability of direct lines having larger vehicles than corridor lines. Finally, when moving from $\mu \to \infty$ to $\mu = 0$ it is evident that operators cost increases and that users costs must decrease by a larger amount. Behind this, of course, lies the variation of fleet and vehicle size. Numerical simulation with data from Santiago, Chile, replicates what was shown in Figure 1 for the one-line case: optimal fleet more than double the operators' cost minimizing fleet and optimal vehicle size is less than half.

Table 1 summarizes the design variables and level of service for both extreme cases of the financial constraint.

Table 1: Summary of results

Financial Constraint Multiplier	Equivalent objective	Best structure	Fleet size	Vehicle size	Average waiting time	Average in-vehicle time
$\mu = 0$	Min $C_U + C_O$	The probability of direct increases with tY, $P_v/\varepsilon P_w$ or $c_1/\varepsilon P_w$	Lower for best structure	The probability of $K_D > K_C$ increases with tY, $P_v/\varepsilon P_w$ or $c_1/\varepsilon P_w$	Lower for corridors	Lower for direct
$\mu \to \infty$	Min C_O	Direct	Lower for direct	Larger for direct		

As said earlier, the analysis of a stringent financial constraint on the design of a public transport system was used by Jara-Díaz and Gschwender (2009) to explain the fleet reduction and vehicle size increase that was part of the complete redesign of the bus system in Santiago, Chile, with dramatic consequences for the users. Looking at the bottom row of columns 3, 4 and 5 in Table 1, it seems that in Santiago only fleet and vehicle sizes had been impacted by the self-financial constraint, namely, smaller and larger than optimal fleet and vehicle size respectively. Regarding the third element, whose analysis was the aim of this research, the pre-existing direct lines structure was changed towards a mix of feeder and corridor trunk lines. According to our results, minimizing operators costs only should have resulted into a system of direct trunk lines; however, corridors were preferred. This cannot be explained in terms of a financial constraint that was not stringent enough, because minimizing total cost for a system with large bus patronage as in Santiago would also yield a direct trunk lines system. We believe that this was due to an important difference between the design process behind fleet and vehicle sizes - which are the result of large scale optimization problems - and the design of a lines structure, which is mostly based on heuristics and intuition. Moreover, the strategic model used to design the public transport routes, frequencies and vehicle sizes was not sensitive to the effect of boarding and alighting times on cycle times (that extends to fleet and, eventually, to costs). Nevertheless, it is quite interesting to note that after the evident initial difficulties, the transit system in Santiago is changing in the three design dimensions analyzed here in the direction suggested by our results: fleet size has increased by

some 30% with smaller than average vehicles, and some services have been either extended or complementarily merged, increasing direct connectivity and inducing some overlapping.

5. CONCLUSIONS

By extending the single line model (Jara-Díaz and Gschwender, 2009) to a representative network, we have shown that a financial constraint does not only affect transit design in terms of frequency and vehicle sizes, but also in terms of the spatial structure of lines. If no financial constraint exists, the optimal structure - corridor or direct lines - will depend on the level of demand Y and on the values of some key parameters: values of time and the marginal cost of providing vehicle capacity. But a stringent financial constraint - which is shown to reduce the importance of users' costs - changes the unconstrained result because what happens to matter is the reduction of the fleet that can be induced by diminishing cycle time through the elimination of transfers, making direct lines with a smaller fleet of larger buses always the best (sub-optimal) option.

The network analyzed includes one central and several peripheral nodes. When users' and operators' cost is minimized, we have shown that it becomes more likely that direct lines are the most convenient as demand increases or the relative value of waiting time decreases (everything else constant); this happens mainly because, in spite of the mandatory transfers, waiting times are lower in the corridors as a result of higher frequencies. When only operators' cost is minimized, direct lines are always more convenient because they avoid transfers, diminishing boarding and alighting time, thus reducing cycle times and fleet size which, finally, reduces operators' cost. Nevertheless, when demands are low, each direct line (specialized in one OD pair) may result in low frequencies yielding large waiting times. This is the reason why the inclusion of users' cost (time) in the optimization changes the optimal structure towards corridor lines for low levels of demand. It was found that for both extreme

cases of the financial constraint, corridor lines yield always lower total waiting times and larger in-vehicle times than the direct lines, but the waiting time effect dominates. The fact that total in-vehicle time is larger in corridors than in direct lines is explained by the transfers, which imply higher in-vehicle times for some passengers.

In summary, for a system with given technical characteristics direct lines are the best structure for the operators for all demand levels[5]. Interestingly, direct lines are also the optimal structure for users and operators when demand is sufficiently high[6]. However, the fleet size is lower in the first case (with larger vehicles) negatively affecting users through the waiting time. It is worth noting that the optimal structure is influenced by the term tY, i.e. demand acts through the boarding and alighting time of passengers. Therefore, the demand effect is reduced when boarding and alighting is made easier for large groups of passengers, for example using several doors simultaneously (as in metro systems), favoring the corridors structure. On the other hand, if a transfer penalty and/or transfer walking time were considered, the probability of direct lines being the best ones would increase. Nevertheless, no relevant change in the qualitative results would occur.

Note that the network model presented here was built to analyze the choice between direct and corridor services for the main lines in an urban setting, the so-called trunk lines. To be able to analyze the convenience of a feeder-trunk system as a whole, our approach could be extended to consider unbalanced demands and shorter services in the extreme points of the network. It would be interesting as well to include crowding, expressed as the ratio between load size k and vehicle size K, which would affect waiting time through the probability of not being able to board a vehicle and could be used to capture discomfort making the in-vehicle

[5] This coincides with the result obtained by Jara-Díaz and Basso (2003) for their simplest case (equal distances, equal flows) in a three nodes network.

[6] This resembles the results obtained in the air transport literature for a socially optimal service structure that depends on demand (Brueckner, 2004), if one associates hub and spoke with corridors (both have transfers) and fully connected with direct lines (no transfers).

time value an increasing function of that ratio. This could yield an optimal vehicle size larger than the maximum load, as obtained by Jara-Díaz and Gschwender (2003a) for a single line model.

ACKNOWLEDGEMENTS

We thank funding by Fondecyt, grant 1120316, and the Institute for Complex Engineering Systems, grants ICM:P-05-004-F and CONICYT:FBO16.

REFERENCES

Basso, L.J. and Jara-Díaz S.R. (2010) The case for subsidisation of urban public transport and the Mohring effect. **Journal of Transport Economics and Policy, 44**, 365–372.

Brueckner, J. K. (2004). Network structure and airline scheduling. **The Journal of Industrial Economics, 52**, 291-312.

Chang, S.K. and P.M. Schonfeld (1991). Multiple period optimization of bus transit systems. **Transportation Research Part B: Methodological, 25**, 453-478.

Hendricks, K., M. Piccione and G. Tan (1995). The economics of hubs: the case of monopoly. **Review of Economic Studies, 62,** 83-99.

Jansson, J. O. (1979). Marginal Cost Pricing of Scheduled Transport Services. **Journal of Transport Economics and Policy, 13**, 268–94.

Jansson, J. O. (1980). A simple bus line model for optimisation of service frequency and bus size. **Journal of Transport Economics and Policy, 14**, 53-80.

Jansson, J. O. (1984). **Transport System Optimization and Pricing**. John Wiley & Sons, Chichester.

Jansson, J.O. (2005). Bus transport system optimization and pricing. In: **Ninth Conference on Competition and Ownership in Land Transport** (Thredbo9), Lisbon, Portugal.

Jara-Díaz, S. R. and L. Basso (2003). Transport cost functions, network expansion and economies of scope. **Transportation Research E, 39**, 269-286.

Jara-Díaz, S. R. and A. Gschwender (2003a). Towards a general microeconomic model for the operation of public transport. **Transport Reviews, 23**, 453-469.

Jara-Díaz, S. R. and A. Gschwender (2003b). From the single line model to the spatial structure of transit services: corridors or direct? **Journal of Transport Economics and Policy, 37**, 261-277.

Jara-Díaz, S. R. and A. Gschwender (2009). The Effect of Financial Constraints on the Optimal Design of Public Transport Services. **Transportation, 36** (1), 65-75**.**

Kepaptsoglou, K. and M. Karlaftis (2009). Transit Route Network Design Problem: Review. **Journal of Transportation Engineering, ASCE, 135**, 491-505.

MIDEPLAN (2007) **Precios sociales para la evaluación social de proyectos** (Social prices for project appraisal). Chilean Government.

Mohring, H. (1972) Optimization and scale economies in urban bus transportation. **American Economic Review, 62**, 591-604.

Pels. E., P. Nijkamp and P. Rietveld (2000). A note on the optimality of airline networks. **Economics Letters, 69**, 429-434.

Parry, I., Small, K.A. (2009) Should urban transit be reduced? **The American Economic Review 99**, 700–724.

SECTRA (2004) **Análisis modernización transporte público, VI etapa** (Public transport cost analysis). Report Stage VI, Chilean Government.

Tirachini, A., Hensher D.A. and S.R. Jara-Díaz (2010). Comparing operator and users costs of light rail, heavy rail and bus rapid transit over a radial public transport network. **Research in Transportation Economics 29**, 231-242.

Van Reeven (2008) Subsidisation of Urban Public Transport and the Mohring Effect, **Journal of Transport Economics and Policy**, **42**(2), 349-359.

Appendix 1: Numerical comparison of total cost from equation (19)

Table A1.1: Values of the parameters used in the numerical evaluation*

Parameter	Value	Units
c_0	10.65	US$/hr
c_1	0.203	US$/hr
t	2.5	Sec
T_0	2,72	Hr
P_w	4.44	US$/hr
P_v	1.48	US$/hr
ε	0.5	

* Bus costs parameters calculated from SECTRA (2004). P_v taken from MIDEPLAN (2007); P_w set to three times P_v.

Table A1.2: Numerical comparison of total cost without and with approximation

Y (pax/hr)	4,000	6,536	7,439	10,000
Total Cost (US$/hr)				
(A) Direct	18,465.8	29,474.9	33,379.5	44,430.3
(B) Corridor	18,400.7	29,474.9	33,407.0	44,543.0
(A)-(B)	65.1	0.0	-27.5	-112.7
First term				
(C) Direct	59.2	96.7	110.0	147.9
(D) Corridor	74.0	120.8	137.5	184.9
Second term				
(E) Direct	3,016.8	4,231.4	4,648.4	5,807.9
(F) Corridor	2,937.0	4,207.2	4,648.4	5,883.7
(E)-(F)	79.8	24.2	0.0	-75.8

Total cost is equal for direct and corridor structures for Y = 6,536 passengers per hour. The third term of eq. (19) is equal for both structures so it cancels out. The first term is never larger than 3.1% of the second term for both structures. When only this latter is used, the difference in cost becomes nil for Y = 7,439 pax/hr, 13.8% larger than the exact value.

Appendix 2. Optimal waiting and in-vehicle times.

Replacing optimal frequency from equation (15) into the expressions for the waiting time (10) and in-vehicle time (14) yields

$$t_w^* = \frac{\varepsilon\left(N+2+4\tau\right)\sqrt{c_0 T_0}}{\sqrt{t\left[3c_1\left(1+\tau\right)+P_v\left(\frac{9}{8}+\frac{\tau}{2}(2N-1)(1-2\tau)\right)\right]+\frac{P_w\varepsilon}{Y}\left(N+2+4\tau\right)}} \quad \text{(A2.1)}$$

$$t_v^* = \frac{3}{4}T_0 Y + \frac{tY\left[\frac{9}{8}+\frac{\tau}{2}\left(2N-1\right)\left(1-2\tau\right)\right]\sqrt{c_0 T_0}}{\sqrt{t\left[3c_1\left(1+\tau\right)+P_v\left(\frac{9}{8}+\frac{\tau}{2}\left(2N-1\right)\left(1-2\tau\right)\right)\right]+\frac{P_w\varepsilon}{Y}\left[N+2+4\tau\right)}} \quad \text{(A2.2)}$$

The waiting times for each structure are obtained replacing (N, τ) by (2, 1/4) for corridors and (4, 0) for direct lines. This yield that total waiting time for corridors is lower than for direct lines when

$$75tc_1 + 28.125tP_v + 150\frac{P_w\varepsilon}{Y} < 135tc_1 + 47.25tP_v + 180\frac{P_w\varepsilon}{Y} \quad \text{(A2.3)}$$

which is always true. Analogously, in-vehicle time for corridors is larger than for direct lines when (A2.4) is valid, which is always true.

$$147tc_1 + 55.125tP_v + 294\frac{P_w\varepsilon}{Y} > 135tc_1 + 47.25tP_v + 180\frac{P_w\varepsilon}{Y} \quad \text{(A2.4)}$$

Appendix 3: Differences in total cost components as a function of *Y*.

Y (pax/hr)	1,000	4,000	6,536	10,000
Operators' cost (US$/hr)				
(A) Direct	1,518.6	5,010.5	7,873.1	11,756.2
(B) Corridors	1,484.8	5,022.3	7,948.8	11,929. 7
(A)-(B)	33.8	-11.8	-75.7	-173.5
Waiting time cost (US$/hr)				
(A) Direct	587.3	1,023.2	1,192.0	1,328.7
(B) Corridors	524.4	875.9	999. 0	1,093.0
(A)-(B)	62.9	147.3	193.0	235.7
In-vehicle time cost (US$/hr)				
(A) Direct	3,070.2	12,432.1	20,409.8	31,345. 4
(B) Corridors	3,082.9	12,502.6	20,527.1	31,520.3
(A)-(B)	-12.7	-70.5	-117.3	-174.9