TRANSPORT REVIEWS, 2003, VOL. 23, NO. 4, 453–469

Towards a general microeconomic model for the operation of public transport

SERGIO R. JARA-DÍAZ and ANTONIO GSCHWENDER

Facultad de Ciencias Fisicas y Matematicas, Departamento de Ingeniera Civil, Universidad de Chile, Santiago, Chile

(*Received 11 April 2002; revised 7 November 2002; accepted 13 November 2002*)

After Vickrey's view, Mohring constructed a microeconomic model to determine the optimal frequency of buses serving a corridor with fixed demand. The main result was that frequency should be proportional to the square root of demand. The role of users' costs was shown to be crucial. This approach has evolved over the past decades, improving our understanding of public transport operations. This paper describes and analyses the evolution of microeconomic models for the analysis of public transport services with parametric demand, leading towards a more comprehensive one. An in-depth review of all the contributions in the academic literature is presented, emphasizing both the treatment of variables and the form of the results mostly in terms of frequency and fleet size. A series of partial new elements is also identified. An extension of Jansson's model for a single period is developed analytically, including the effect of vehicle size on operating costs and the influence of crowding on the value of time. Numerical simulations are used for comparison and analysis. A general model is then proposed where bus operations are optimized accounting for a number of simultaneous relations. Finally, the different models are discussed and compared.

1. Introduction

In the microeconomic analysis of urban public transport, two types of resources have to be taken into account: those provided by the operators, as vehicles, fuel, terminals or labour, and those provided by the users, namely their time, usually divided into waiting, access and in-vehicle time. After Vickrey's view (1955), Mohring (1972, 1976) constructed a microeconomic model to determine optimal frequency of buses serving a corridor with fixed demand. The main result was that frequency should be proportional to the square root of demand, and this happened only because all resources (operators' and users') were considered when finding the minimum cost operation. The role of users' costs was shown to be crucial. This approach has evolved along the last decades, improving our understanding of public transport operations.

In this paper we describe and analyse the evolution of microeconomic models for the analysis of public transport services with parametric demand, leading towards a more comprehensive one. An in-depth review of all the contributions in the literature is presented, emphasizing both the treatment of variables and the form of the results mostly in terms of frequency and fleet size. This permits the formulation of a general model where bus operations are optimized accounting for a number of simultaneous relations.

An extension of Jansson's (1980, 1984) model for a single period is proposed, including the effect of vehicle size on operating costs and the influence of crowding on the

Transport Reviews ISSN 0144-1647 print/ISSN 1464-5327 online © 2003 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0144164032000048922

value of time. In this extended model, vehicle size is optimized in addition to fleet and frequency, including a capacity constraint. Unlike the general model, this can be solved analytically. Numerical simulations are used for comparison and analysis of the resulting frequencies. Finally, a comparative analysis is presented in which the different models are discussed and compared.

2. Mohring's models

In his pioneering work based on Vickrey's view (1955), Mohring (1972, 1976) developed a microeconomic model to optimize bus operation in an isolated bus route. A bus fleet B, with an operating cost c per vehicle-hour produces a frequency f. If Y passengers per hour use the service and t_w, t_v and t_a are the average waiting time in the bus stop, in-vehicle time and access (walking) time, respectively, the total value of the vehicle resources consumed (VRC) by operators and users per hour is:

$$\text{VRC} = \text{B}c + P_w t_w Y + P_v t_v Y + P_a t_a Y, \tag{1}$$

where P_w, P_v and P_a are the prices of time (waiting, in-vehicle and access, respectively). Note that a constant c implies a given bus size, which reveals an implicit assumption on capacity enough to accommodate demand.

Cycle time t_c is assumed constant in this model. It is related with B through frequency, i.e.

$$B = f\, t_c. \tag{2}$$

Replacing (2) in (1), assuming that average waiting time is half the vehicle interval ($h = 1/f$) and assuming that t_v and t_a are constant, the VRC can be written as a function of frequency only:

$$\text{VRC} = f\, t_c c + P_w \frac{1}{2f} Y + P_v\, t_v\, Y + P_a\, t_a Y. \tag{3}$$

This shows that increasing frequency has a double effect. It increases operators' expenses but diminishes waiting time (hence the importance of users' costs), which can be easily observed deriving expression (3).

$$\frac{\partial \text{VRC}}{\partial f} = t_c c – P_w \frac{1}{2f^2} Y. \tag{4}$$

Making it equal to zero, and noting that the second derivative is positive, the optimal frequency is obtained as:

$$f^* = \sqrt{\frac{P_w}{2t_c c} Y}, \tag{5}$$

which is known as the 'square root formula'.

Mohring (1972, 1976) also presented a more complex model in which the number of stops in the corridor is a variable as well, and where both the bus cycle time and users' in-vehicle time depend on the number of passengers boarding and alighting. Moreover, a probability of buses not stopping at a bus stop is also considered. Cycle time now is:

$$t_c = t\frac{Y}{f} + t_r + t_p\, p\, (1 - e^{-n}), \tag{6}$$

where p is the number of bus stops in the route and

$$n = \frac{2Y}{fp} \tag{7}$$

is the average number of passengers per bus that boards or alights at each bus stop.

The ratio Y/f is the number of passengers that board a bus within a cycle, while t is boarding and alighting time. Thus, the first term in the right-hand side of equation (6) is the total time each bus remains at the bus stops along a cycle. On the other hand, t_r is the time a vehicle is in motion during a cycle (including delays due to traffic interaction). Delays due to speed reduction and acceleration at the bus stops are included in the third term of the right- hand side. This is expressed as the product between this type of delay time at a bus stop (t_p), the number of bus stops (p) and the probability of actually stopping, $(1 - e^{-n})$.[1]

On the other hand, if l is the average journey length and L is the route length, the average travel time can be written as a function of cycle time:

$$t_v = t_c \frac{l}{L}\,. \tag{8}$$

The average access time (t_a) enters the model as a function of the number of stops (p) and access or walking speed (v_a). Origins and destinations of passengers are assumed to be homogeneously distributed along the route. On average, passengers will have to walk one-quarter of the distance between stops (ratio between L and p) in their origins and in their destinations. Thus,

$$t_a = \frac{L}{2p\, v_a}\,. \tag{9}$$

Replacing the last two equations in the VRC function (3) yields:

$$\text{VRC} = f\, t_c\, (f, p)\, c + P_w \frac{1}{2f} Y + P_v t_c\, (f, p) \frac{l}{L} Y + P_a \frac{L}{2pv_a} Y. \tag{10}$$

Equations (6), (7) and (10) make VCR a function of f and p. Increasing frequency increases operators' expenditure and diminishes both waiting and in-vehicle travel times of the users. Increasing the number of bus stops increases both operators' expenditure and in-vehicle users' travel time, but diminishes access time. Frequency does not influence access time and the number of bus stops does not affect waiting time. Although the trade-offs are quite clear, explicit optimum expressions for f and p can not be found analytically, which is why Mohring finds them numerically.

3. Jansson's models

Jansson (1980, 1984) simplifies Mohring's model (equations 6–10) by eliminating the number of bus stops as a variable and assuming buses always stop, but keeping the effect of passengers boarding and alighting on travel time. This reduces cycle time to:

$$t_c = t\frac{Y}{f} + T, \tag{11}$$

where T is vehicle time in movement within a cycle, including interactions with other vehicles and time to reduce speed and accelerate at each bus stop. On the other hand, frequency is given by the ratio between fleet size and cycle time (B/t_c), which combined with equation (11) yields:

$$B = f\,T + tY. \tag{12}$$

In this model, access time is a constant because neither route design nor the number of bus stops are considered as variables and, therefore, access cost is not relevant to optimize the service and may not be included in VRC. Recalling equations (1) and (8), the relevant total value of the resources consumed (VRC) per hour is:

$$\text{VRC} = Bc + P_w\frac{1}{2f}Y + P_v\frac{l}{L}t_c\,Y. \tag{13}$$

Using equations (11) and (12), we can write expression (13) as a function of B, i.e.

$$\text{VRC} = Bc + P_w\frac{T}{2(B-tY)}Y + P_v\frac{l}{L}\left(T + \frac{tTY}{B-tY}\right)Y. \tag{14}$$

This expression shows that, *ceteris paribus*, increasing the number of vehicles diminishes users' costs but increases operators' costs. Users' cost reduction occurs because increasing frequency diminishes both waiting and in-vehicle travel times, the latter because fewer individuals board and alight per bus.

Minimizing VRC with respect to B yields the optimal fleet size B^*, given by:

$$B^* = tY + \sqrt{\frac{TY}{c}\left(\frac{1}{2}P_w + P_v tY\frac{l}{L}\right)}. \tag{15}$$

From equation (12), the optimal frequency can be obtained as:

$$f^* = \sqrt{\frac{Y}{cT}\left(\frac{1}{2}P_w + P_v tY\frac{l}{L}\right)}, \tag{16}$$

which represents a modified version of the 'square root formula'. According to this result, optimal frequency increases proportionally to the square root of total demand if the second term in the parenthesis is negligible relative to the first, but it can vary proportionally to demand if the contrary happens.

The average number of passengers aboard each vehicle in this model is given by:

$$k = \frac{Yl}{fL}, \tag{17}$$

which should be less than the vehicle capacity assumed for the calculation of c.

An important limitation of the previous models is that they require homogeneous conditions along time. In fact, variables like Y and T do vary during a day. Using these models for each homogeneous period makes the value of c debatable. If, as a result of the optimization procedure, the same number of buses was obtained for each period, then a common (single) value for c would be appropriate (unless for the observation in the previous paragraph). However, if more buses were needed in the peak period, c would have to be larger for that period, because the capital cost for the off-peak period would be nil (no need for a new bus). Besides, crew cost per hour for the operator would be larger, as the firm would need extra drivers in the peak hour in this case. These reasons made Jansson (1980) estimate that the cost of running an extra bus during one off-peak hour represents between one- and two-thirds the cost of running one extra bus exclusively for the peak hour. Along these lines, Jansson expanded his model considering two periods during the day: peak and off-peak, concluding that in general it would be optimal to consider the same number of buses (fleet size) for both periods. Because of practical considerations regarding scheduling, Jansson stated that a common frequency was the most appropriate operating rule; note that a lower cycle time in the off-peak period would imply that less buses would be required. A further modified version of the square root formula is obtained, i.e.

$$f^* = \sqrt{\frac{\overline{Y}}{cT^P}\left(\frac{1}{2}P_w + P_v t Y^P \frac{l}{L}\right)}, \tag{18}$$

where Y^P and T^P are the values that Y and T have in the peak period, respectively, and $\overline{Y}$ is the weighted average of the hourly demand (considering peak and off-peak periods). Thus, the optimal frequency depends both on the peak and off-peak demands.

Finally, Jansson extended his model to determine the optimal size of vehicles, using a linear relation between c and vehicle size (K) that he found appropriate for both running and standing costs, i.e.

$$c(K) = c_0 + c_1 K, \tag{19}$$

where c_0 and c_1 are constants. Vehicle size is equal to the number of passengers aboard each vehicle (k) in the peak period, given by equation (17) with $Y = Y^P$. This was introduced in the previous model, obtaining a new optimal frequency rule:

$$f^* = \sqrt{\frac{1}{c_0 T^P}\left[\overline{Y}\left(\frac{1}{2}P_w + P_v t Y^P \frac{l}{L}\right) + \frac{c_1 E t (Y^P)^2 l}{L}\right]}, \tag{20}$$

where E is the length of the operation period. Two differences emerge in this new expression: first, in the denominator of the square root c is replaced by c_0, and, second, a new term related with c_1 appears. Both differences work to rise f^* in comparison with the situation where the bus size is given (equation 18).

4. Other models

The optimization of isolated bus corridors has received further attention from other viewpoints as well. Kraus (1991) developed a pricing model that considered an interesting new aspect, namely the effect of crowding inside the vehicle by means of a time price related with the passenger/capacity ratio. He considered a bus line that ran between the

CBD and the periphery; the users board along the route and alight at the end. The effect of crowding on the time price is set such that:

$$P_v = \begin{cases} \overline{P}_v \text{ if the passenger boards before stop } i_0 \\ \overline{P}_v + P_h \text{ if the passenger boards after stop } i_0 \end{cases} \quad (21)$$

where i_0 is the bus stop where all seats are occupied. Thus, $P_h > 0$ is the extra time price perceived by those that have to travel standing because all seats are occupied. Therefore, the users that board at the first bunch of bus stops and seat cause a larger marginal (user) cost (through P_h on those that have to stand) than those who board when there are no empty seats left. This is in essence why the pricing model of Kraus (1991) yields higher optimal fares for those that board first. The frequency is an exogenous parameter in this model.

Using elastic demand, Oldfield and Bly (1988) developed a model to determine optimal bus size considering passenger congestion, making waiting time dependant on vehicle occupancy. Demand depends on the generalized cost, under a constant elasticity form. The treatment of waiting time is particularly interesting as it depends not only on the service frequency, but also on the vehicles occupancy rate $ø = k/K$, which affects the probability of a passenger not being able to board the bus, increasing waiting time. Thus,

$$t_w = t_w (f, ø). \quad (22)$$

The model includes three alternative functional forms for the average waiting time: constant occupancy rate, product form and additive form. In the first case:

$$t_w = \frac{\varepsilon}{f}, \quad (23)$$

where ε is a constant that depends on the level at which ø is set. If it is low and passengers arrive randomly, ε will be close to 0.5. The larger ø, the larger will be ε. The product and additive functions are:

$$t_w = \frac{\varepsilon}{f}. (1 - z ø^{\gamma})^{-1} \text{ and} \quad (24)$$

$$t_w = \frac{\varepsilon}{f} + z ø^{\gamma}, \quad (25)$$

respectively, where ε, z and γ are parameters. In both cases, increasing ø increases average waiting time and, for sufficiently low values of ø (depending on the values of both z and γ), the average waiting time collapses to ε/f. As in Jansson (1980, 1984), the operating cost is assumed proportional to vehicle size.

The model includes two type of externalities. On one hand, increasing the number of bus users decreases the number of car users, reducing vehicle congestion and travel time for all users. On the other hand, increasing bus frequency increases congestion and travel time.

The objective is to find vehicle size (K) and frequency (f) as to maximize social welfare, which in this model has three components: consumers' surplus, externalities

and producers' surplus. Both externalities are assumed linear (with respect to demand and frequency, respectively). This permits to find an implicit function for vehicle size.

Using elastic demand as well, Evans and Morrison (1997) included two new variables, namely disruption or non-scheduled delay (d) and accident risk (r). The latter is measured through the level of risk perceived by the users. Both characteristics are inversely valued by the users (they prefer low risk and little delay) and both mean a cost for the operators. Thus, users want them small and operators would like to spend little on them. Demand depends exponentially on generalized cost GC, which is specified as:

$$\mathrm{GC} = P + P_{\mathrm{w}}\, t_{\mathrm{w}}\,(f) + P_{\mathrm{v}} t_{\mathrm{v}} + P_{\mathrm{r}} r + P_{\mathrm{d}} d, \tag{26}$$

where t_{v} is constant and waiting time depends on frequency. Users value risk and delays at P_r and P_d, respectively. P is the service fare.

Operators expenses have two additional terms, c_r and c_d, which represent the cost of achieving a certain risk level r and delays d, respectively. The risk-reduction cost function c_r is empirically specified as:

$$c_r = -\theta_1 \, ln\left(1 + \frac{\theta_2 r - \theta_3}{\theta_4}\right), \tag{27}$$

such that reducing risk from a base level ($c_r = 0$) has a positive cost. The cost of reducing non-scheduled delay is simply specified as:

$$c_d = \frac{\theta_5}{d}, \tag{28}$$

where the θ_i ($\theta_i > 0$) represent different parameters.

Social benefit is maximized, and it includes consumers' surplus, producers' surplus and one negative externality, namely accidents induced by the bus service on the non-users. This externality is expressed as the number of equivalent fatalities (m), which depends on service frequency (f) and on the operators' expenses on safety (c_r). Deaths and injuries weighted by severity are included in m, which is valued in the social benefit expression at a price P_m. Various financial constraints are considered in a numerical maximization.

Economies of scale are present in public transport essentially due to the reduction in waiting time as demand increases. An interesting result from Evans and Morrison (1997) is that these economies are increased after risk and delays are taken into account. Among other things, this is due to the larger expenses in safety that are allowed after a demand increase, which benefits all users.

As seen, isolated corridor analysis played an important role in establishing a rigorous microeconomic approach to public transport analysis. Nevertheless, the spatial dimension is indeed important. Along this line, Kocur and Hendrickson (1982) and Chang and Schonfeld (1991) considered an area served by parallel, equally spaced, bus lines. Both frequency and lines spacing were considered as optimization variables. Bus cycle time and passenger travel time were regarded as independent of the number of users, which made it possible to obtain analytical solutions. In Chang and Schonfeld (1991) the optimal frequency for a single period is:

$$f^* = 3\sqrt{\frac{P_w^2 v_a Y}{W t_c c P_a}}, \tag{29}$$

where W is the width of the area served by parallel bus lines (an extension is developed for multiple periods). Both papers conclude that both the optimal interval between buses and the optimal spacing between lines are inversely proportional to the cubic root of demand. This result shows that when it is possible to act on the bus lines density, the optimal reaction to demand increases has two dimensions: increase lines density and increase bus frequency. Because of this, in this case (cubic root) optimal frequency grows less than in the isolated corridor case (square root). On the other hand, the spatial distribution of demand has been taken into account by Jara-Diaz and Gschwender (2003), who analysed the optimal size and assignment of a bus fleet to a set of bus lines serving either along non-overlapping corridors (transfers are needed) or as a set of direct (point-to-point) services; a new version of the 'square root formula' for fleet size was found, including some coefficients that affect waiting and travel time averages, representing both the spatial structure of demand and the bus lines structure. In another field of spatial investigation (optimization of the physical location of bus routes), Chien *et al.* (2001) proposed a genetic algorithm approach that finds an approximation to the optimal bus route location and frequency for a feeder route. In this recent contribution user and supplier costs are considered and, not surprisingly, the optimal frequency happens to follow a 'square root formula'.

5. New model to optimize frequency and capacity including crowding

In what follows, an extension of Jansson's (1980, 1984) model for a single period is shown, including the effect of vehicle size on operating costs and the influence of crowding on the value of in-vehicle time. In this model, vehicle size is optimized in addition to fleet and frequency, including a capacity constraint. The problem to be solved is:

$$\underset{f,K}{\text{Min}}\ \text{VRC} = f t_c(f) c(K) + P_w \frac{1}{2f} Y + P_v(\text{ø}) \frac{l}{L} t_c(f) Y, \tag{30}$$

subject to

$$k(f) \le K, \tag{31}$$

where

$$\text{ø} = \frac{k(f)}{K} \tag{32}$$

is the occupancy rate that affects the value of in-vehicle time. A linear form will be used for $P_v(\text{ø})$, i.e.

$$P_v(\text{ø}) = P_{v0} + P_{v1}\,\text{ø}. \tag{33}$$

A linear form, given by equation (19), is used for $c(K)$ as well. Cycle time is given by equation (11), and the number of passengers aboard each vehicle (k) is given by equation (17).

The analytical solution of this problem presents two cases, depending on the relative values of c_1 and P_{v1}. If the former is smaller than the latter then constraint (31) is not active, otherwise it is. For the inactive capacity restriction we have:

$$c_1 < P_{v1} \Rightarrow \begin{cases} f^* = \sqrt{\dfrac{Y}{c_0 T}\left[\dfrac{P_w}{2} + tY\dfrac{l}{L}(2\sqrt{c_1 P_{v1}} + P_{v0})\right]} \\ K^* = \sqrt{\dfrac{P_{v1}}{c_1}}\, k(f^*) \end{cases} \tag{34}$$

In the case where constraint (31) is active we get:

$$c_1 \geq P_{v1} \Rightarrow \begin{cases} f^* = \sqrt{\dfrac{Y}{c_0 T}\left[\dfrac{P_w}{2} + tY\dfrac{l}{L}(c_1 + P_{v1} + P_{v0})\right]} \\ K^* = k(f^*) \end{cases} \tag{35}$$

In both cases $k(f^*)$ is given by equation (17). The optimal fleet size (B^*) can be obtained using equation (12).

It is interesting to note that in both solutions the optimal occupancy rate (the ratio between $k(f^*)$ and K^*) does not depend on Y. On the other hand, whether or not constraint (31) is active is also independent of Y. When parameters c_1 and P_{v1} have the same value, both cases yield identical results, i.e. the solution is continuous on the ratio between these two parameters. If $c_1 = P_{v1} = 0$, Jansson's (1980, 1984) optimal solution for frequency is obtained (equation 16), as expected.

Expression (34) shows that when the marginal cost of the vehicle capacity (c_1) is lower than the marginal value of the occupancy rate (P_{v1}), it is optimal to have excess capacity. If the opposite occurs ($c_1 \geq P_{v1}$), the optimal vehicle capacity equals the number of passengers per vehicle (expression 35). Using equations (17), (34) and (35), explicit expressions can be obtained for the optimal vehicle size for each case:

$$c_1 < P_{v1} \Rightarrow K^* = \frac{l}{L}\sqrt{\frac{P_{v1} c_0 T}{c_1}\left[\frac{P_w}{2Y} + t\frac{l}{L}(2\sqrt{c_1 P_{v1}} + P_{v0})\right]^{-1}}, \tag{36}$$

$$c_1 \geq P_{v1} \Rightarrow K^* = \frac{l}{L}\sqrt{c_0 T\left[\frac{P_w}{2Y} + t\frac{l}{L}(c_1 + P_{v1} + P_{v0})\right]^{-1}}. \tag{37}$$

These results show that the optimal vehicle size increases with demand at a decreasing rate (concave in Y) and that K^* is asymptotic to different values depending on each case.

Using equations (32), (33) and the values for the optimal bus size in (34) and (35), the optimal frequencies can be rewritten as:

$$f^* = \sqrt{\frac{Y}{c_0 T}\left[\frac{P_w}{2} + tY\frac{l}{L}(P_v + \sqrt{c_1 P_{v1}})\right]} \text{ and} \tag{38}$$

$$f^* = \sqrt{\frac{Y}{c_0 T}\left[\frac{P_w}{2} + tY\frac{l}{L}(P_v + c_1)\right]}, \tag{39}$$

for cases (34) and (35), respectively. These equations can be compared against expression (16) of Jansson's (1980, 1984) single period and fixed capacity model. Two differences can be seen: a new term related with c_1 appears and, in the denominator of the square root, c is replaced by c_0 ($c_0 \leq c$). Thus, when vehicle capacity is optimized optimal frequency is larger than in the fixed capacity case, as happened in the comparison between equations (20) and (18).

Finally, it is worth noting that equations (35) and (39) correspond to an optimal vehicle size equal to the load size k. As explained in Section 3, this is what Jansson assumes in his model for two periods with vehicle cost depending on vehicle size. Then, it is not surprising that the optimal frequency in this latter case (equation 20) collapses into equation (39) if equal demands are considered for both periods.

6. Illustration and comparison

To have an idea of the differences among models, we have simulated the variation of optimal frequency with total demand for four comparable cases, namely Mohring's model (equation 5), Jansson's first model (equation 16) and ours (equations 34 and 35). Most (but not all) of the parameters used for comparison were taken as those that are typically observed in Santiago, Chile. These are shown in table 1. Note that as $c_1 < P_{v1}$ the relevant equation for the new model is (34). Nevertheless, equation (35) or its equivalent (39), can be represented as well as it does not require a value for P_{v1} but only for P_v for the numerical experiments; note that a smaller value for P_{v1} had to be assumed implicitly in this case.

Figure 1 shows the simulated frequencies for each of the four cases (bold continuous lines) as a function of total demand, which varies from 200 to 9200 passengers per hour. Assuming the value of the parameters given in table 1, this implies passenger flows between 33 and 1533 pax/h, which are low for a city like Santiago, where at least seven corridors have passenger flows higher than 10 000 pax/h (in fact, as large as 30 000 in the

Table 1. Parameters for simulation

Parameter	Value	Units
c	12.07	US$/hr
c_0	4.95	US$/hr
c_1	0.09	US$/hr
l	10	Km
L	60	Km
P_v	2.70	US$/hr
P_{v0}	2.16	US$/hr
P_{vl}	0.54	US$/hr
P_w	8.11	US$/hr
t	5	Sec
T	2.72	hr
t_c	3	hr

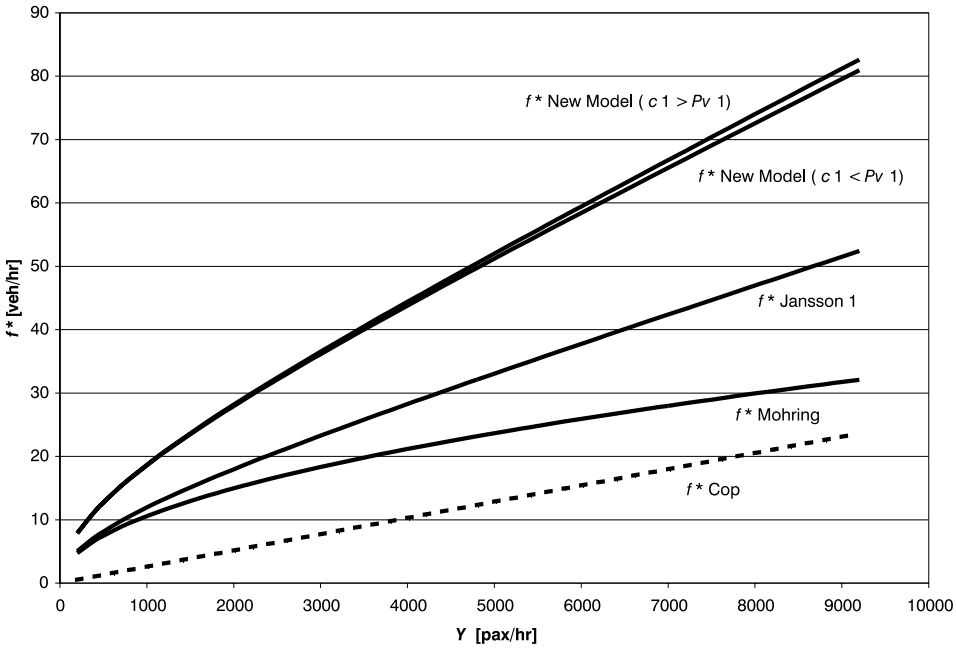

Figure 1. Optimal frequencies in different models.

main corridor, considering buses only!). The passenger flow (F) can be determined in these models by:

$$F = Y\frac{l}{L}. \tag{40}$$

The ratio F/f represents the average number of passengers aboard each vehicle (k), as can be seen from equation (17).

The simulated results show that the three models tend to coincide for the lowest levels of demand, particularly for Mohring's and Jansson's first model. The divergence is evident when Y increases. The four curves can be analysed using equations (5), (16), (34) and (35). First, recall that cycle time t_c is considered constant in Mohring's model while the other three add boarding and alighting time (demand dependent) to the vehicle time in motion T. Thus, for low levels of demand, time at the bus stops is less important such that the models tend to coincide. However, T is always smaller than t_c, which makes optimal frequency always smaller for Mohring's model. This is reinforced when demand increases, as clearly seen through the second term in square roots (16), (34) and (35).

On the other hand, the new model considers cycle time as in Jansson's first, but adds two additional features, namely, sensitivity of the value of travel time to the occupancy rate of the vehicles and operating costs depending on vehicle size. Both features make the solution dependent on K, which has to be optimized along frequency and fleet size. As predicted in the discussion in Section 5, the resulting frequencies are larger for the new model. Optimal vehicle size in the new model varies between four and 47 passengers per

vehicle for the ranges of demand in figure 1. From equations (36) and (37), K^* can be shown to grow asymptotically to:

$$\lim_{y \to \infty} K^* = \frac{l}{L}\sqrt{\frac{P_{v1}c_0T}{c_1}\left[\frac{l}{tL}(2\sqrt{c_1 P_{v1}} + P_{v0})\right]^{-1}} \quad \text{if } c_1 < P_{v1} \text{ and} \tag{41}$$

$$\lim_{y \to \infty} K^* = \frac{l}{L}\sqrt{c_0T\left[\frac{l}{tL}(c_1 + P_{v1} + P_{v0})\right]^{-1}} \quad \text{if } c_1 \geq P_{v1}, \tag{42}$$

which implies, for the values of the parameters considered, maximum vehicle sizes of 61 passengers per vehicle for the former case and 24 for the latter one. It is worth noting that when $c_1 < P_{v1}$, K^* directly depends on the square root of P_{v1}/c_1, which make the optimal vehicle size highly influenced by these parameters. On the other hand, when $c_1 \geq P_{v1}$, the optimal vehicle size simply coincides with the number of passengers aboard the vehicle (k). Optimal operation at capacity suggests the comparison of all cases against 'optimal' frequency when only operators costs are considered, as in this case the minimum frequency which satisfies demand will be used, i.e. $f = F/K$, which is the operating rule that Mohring (1972) calls 'make service frequency proportional to patronage' (p. 594). Assuming a capacity of 65 passengers per vehicle observed in Santiago, Chile, the dotted line in figure 1 represents this rule, which happens to yield the lowest frequency for all levels of total demand, as expected. Thus, all four cases in figure 1 generate optimal frequencies that are larger than the minimum necessary, as required.

In all models but Mohring's, the larger Y, the more important the second term in frequency (the one that includes the value of in-vehicle time) relative to the first term (that includes the value of waiting time). Thus, for large demand values optimal frequency varies linearly with demand. Nevertheless, although the curves look linear after $Y = 3000$, the first term is never negligible relative to the second within the range analysed. For example, in the case of Jansson 1, the second term is only 42% larger than the first for $Y = 9200$.

For synthesis, the numerical simulation clearly shows that optimal frequencies are model sensitive indeed. When cycle time is considered constant and no provision is made for the variation of operating costs with vehicle size, i.e. Mohring's model, the resulting optimal frequencies are the lowest for any given level of demand. On the other extreme, when cycle time is sensitive to demand because of passengers boarding and alighting, and when vehicle size is optimized, the largest optimal frequencies are obtained. The consideration of users' costs makes frequencies larger than the minimum necessary if only operators' costs were considered. Note, however, that when high frequencies are obtained, the assumption of a constant travel time between stops becomes debatable and a dependence of it on frequency seems necessary.

7. Towards a general model

Consider a route of length L and one period. All passengers are evenly distributed along the route and travel an average distance l. Demand is represented parametrically by Y, the number of passengers that require the system in 1 h. The problem to be solved is:

$$\underset{f,K,c_r,c_d,p,c_x}{\text{Min}} \quad \text{VRC}_\text{T} = \text{VRC}_\text{op} + \text{VRC}_\text{U} + \text{VRC}_\text{x}, \tag{43}$$

i.e. to minimize total value of the resources consumed (VRC_T), which includes those perceived by the operators (VRC_op), those perceived by the users (VRC_U) and those incurred by society as externalities (VRC_X). This is subject to a capacity restriction:

$$ø_{\max} K \geq k, \tag{44}$$

where $ø_{\max}$ is the maximum occupancy rate allowed. This could be less than 1 to leave a capacity margin to account for random variations of demand. The actual occupancy rate is evidently given by the ratio between the average number of passengers on each vehicle (k) and their capacity (equation 32) with k given by equation (17).

The objective function is minimized with respect to five design variables: frequency (f), vehicle capacity (K), expenses to reduce accident risk (c_r), expenses to diminish schedule delay (c_d),[3] the number of bus stops (p) and expenses to reduce externalities (c_x). As the number of vehicles (B) is given by frequency times cycle time, B is implicitly optimized through f. Operators expenses are defined as:

$$\text{VRC}_\text{op} = c(K) f t_\text{c}(f, p, ø) + c_\text{r} + c_\text{d} + c_\text{x}. \tag{46}$$

Following Mohring (1972, 1976), cycle time is specified as:

$$t_\text{c} = t_\text{r}(f) + t_\text{p}\, p\, [1 - \text{e}^{-n}] + t(ø,f)\frac{Y}{f}, \tag{47}$$

with n defined as in (7).

Vehicle time in motion (t_r) corresponds to the vehicle cycle time less time spent due to bus stops. We put this as a function of f, which makes it possible to consider congestion. In the case of buses, this function would be increasing beyond a certain frequency (the model does not include car–bus interactions). The second term represents the extra time a bus has to spend due to speed reduction and acceleration at bus stops within each cycle, with t_p the extra time due to one bus stop. The term in brackets is the probability of stopping, which depends on the average number of passengers that wants to board or alight at each bus stop (n), as explained in Section 2. This is more appropriate for bus schedules than for underground services, which might be irrelevant in practice because n is usually sufficiently high to make the term in brackets equal to 1. In any case, this term multiplied times p yields the actual average number of stops a vehicle has to make in a cycle.

The fraction of the third term is the number of passengers that board the bus within each cycle. Once this is multiplied times the time each passenger takes boarding and alighting (t), the total time a vehicle has to stop to permit passenger board and alight within each cycle is obtained. The model accepts that this time can increase due to passenger congestion. If the vehicle is carrying too many passengers, boarding and alighting gets difficult. Also, this time can be affected by vehicle congestion at the bus stop if, for instance, there is a queue of vehicles and the users have to walk to the place where the desired bus is waiting. One has to take into account that t also depends on vehicle design factors (i.e. number and size of the doors), on the fare collection system, etc.

Users' costs can be written as:

$$\mathrm{VRC_U} = P_w t_w (f, ø, \mathrm{d}(c_d))\, Y + P_v (ø)\, t_v (f, p, ø) Y +$$

$$+ P_a t_a (\mathrm{p}) Y + P_m r\left(f, c_r, K, \frac{L}{t_r(f)}, \mathrm{d}(c_d)\right). \quad (48)$$

Average waiting time can be expressed as:

$$t_w = \frac{\varepsilon}{f} + z ø^{\gamma} + d(c_d). \quad (49)$$

The first and second terms follow Oldfield and Bly (1988). The first term represents a general expression for the effect of frequency on waiting time, while the second takes care of the congestion effect caused by demand randomness. The fuller the vehicles (the larger ø), the larger the probability of a passenger not boarding the first vehicle arriving. Thus, average waiting time increases with the occupancy rate. Finally, the function $d(c_d)$ corresponds to the average schedule delay. Including this in the average waiting time assumes that vehicles circulate according to a pre-established schedule or keeping regular intervals.

In-vehicle time is given by expression (8) and, following Kraus (1991), its price depends on the occupancy rate to represent the effect of crowding inside the vehicle. Regarding average access time, in this case of an isolated route it depends on the number of stops according to expression (9).

Finally, there is a last term related with the effect of accidents on the users. The function r represents the risk faced by the users due to potential accidents. As in Evans and Morrison (1997), deaths and injuries weighted by severity are included in r, which is valued at a price P_m. Function r depends positively on frequency (more vehicles circulating) and negatively on c_r (operators' expenses to diminish accident risk). Dependence on K is because the heavier the vehicle, the less severe the consequences of an accident for the passengers. It also depends on the average cruise speed (L over time in movement), as the larger that speed the larger the probability and severity of an accident. Lastly, r also depends on schedule delay, because higher exigencies in the fulfilment of the schedule increase the probability of accidents.

The externalities cost (loss of resources) can be written as:

$$\mathrm{VRC_x} = P_m m\left(f, c_r, K, \frac{L}{t_r(f)}, d(c_d)\right) + P_x x\left(f, K, \frac{L}{t_r(f)}\right). \quad (50)$$

Function m is very similar to function r just described, but refers to the risk faced by non-users, external to the system, as pedestrians and users of other modes. In this case, vehicle size increases that risk, while the other variables follow the same type of relation as in r. Function x represents a vector of externality levels of other sorts, as acoustic effects, gas emissions, and so on. These depend on f, K and speed. Externality prices are contained in the vector P_x.

The model can be easily extended to any number of periods, as in Chang and Schonfeld (1991), taking into account that some variables are common to all periods (e.g. K) while others are period specific (e.g. f).

The treatment of vehicle congestion in this model is somewhat limited. High frequencies can cause delays because of interaction among vehicles serving the route only (i.e. cars do not interact with public transport). This means that the model ignores two effects. First, public transport vehicles could in fact share lanes with cars, causing congestion and delays on car users. Second, if car users are attracted to public transport, there would be a reduction in congestion.

8. Synthesis and conclusions

After a revision of most relevant microeconomic models for public transport operations, a number of partial contributions to the original formulations by Mohring (1972, 1976) and Jansson (1980, 1984) were detected. Some have been incorporated here into a general model where travel times depend on the number of passengers boarding and alighting, operating costs depend on vehicle size, access time depends on the number of bus stops, accident risk and punctuality are interrelated, etc. Road congestion is included, making speed dependent on frequency, and waiting time is affected by vehicle occupancy when it approaches vehicle capacity. Safety, punctuality and externalities are considered in the optimization of frequency, vehicle size and number of bus stops.

An extension of Jansson's (1980, 1984) model for a single period has been developed analytically, including the effect of vehicle size on operating costs and the influence of crowding on the value of time. In this model, vehicle size is optimized in addition to fleet and frequency, including a capacity constraint. The solution has two cases, depending on whether or not the capacity constraint is active. A new richer version of the 'square root formula' was obtained for optimal frequency. Comparing the expressions for the optimal frequency in the different models (table 2), some interesting results arise.

When the spacing between lines and the frequency are optimized altogether, the optimal frequency grows less (proportional to the cubic root of the demand) than in the optimization of the frequency alone (proportional to the square root). This is explained by the possibility of increasing both the frequency and the bus lines' density, as a response to demand growths. On the other hand, allowing the vehicle capacity to be optimized increases the optimal frequency, both in the single period and in the two period's cases. Further, when two periods with different demands and travel times are considered, the optimal frequency (equal for both periods) depends on the demands of both peak and off-peak periods. Finally, when the occupancy rate is taken into account through the value of in-vehicle time, the frequency for a given demand tends to be larger than in other cases because there is yet another positive effect on users, diminishing their cost although operators' cost increases. Note, however, that there is also a positive effect on vehicle size, which softens the effect on frequency.

Numerical simulations confirm that optimal frequencies get larger as users' costs are better represented. Differences can be substantial across models for the same total demand. The effect of users' costs is quite relevant, particularly in the new model where optimal frequencies can get as large as three to four times that of the 'minimum necessary' rule. On the other hand, optimal vehicle size in the new model is larger when the marginal effect of the occupancy rate on the value of time is larger than the marginal cost of size. These make a clear case for a better understanding of the role of users' costs and its components on the optimal design of public transport systems.

Table 2. Optimal frequencies in different models

Model	Optimal frequency	Assumptions
Mohring	$f^* = \sqrt{\frac{P_w}{2t_c c}\, Y}$	Constant cycle and travel times. One period.
Jansson 1	$f^* = \sqrt{\frac{Y}{cT}\left(\frac{1}{2}P_w + P_v t Y \frac{l}{L}\right)}$	Cycle and travel times depend on frequency and demand. One period.
Jansson 2	$f^* = \sqrt{\frac{Y}{cT^P}\left(\frac{1}{2}P_w + P_v t Y^P \frac{l}{L}\right)}$	Cycle and travel times depend on frequency and demand. Two periods.
Jonsson 3	$f^* = \sqrt{\frac{1}{c_0 T^P}\left[\bar{Y}\left(\frac{1}{2}P_w + P_v t Y^P \frac{l}{L}\right) + \frac{c_1 E t (Y^P)^2 l}{L}\right]}$	Cycle and travel times depend on frequency and demand. Two periods. Operating cost depends on vehicle size.
Chang and Schonfeld	$f^* = \sqrt{\frac{P_w^2 v_a Y}{W t_c c P_a}}$	Constant cycle and travel times. One period. Route spacing is optimised.
New Model	$c_1 < P_{vt} \Rightarrow \begin{cases} f^* = \sqrt{\frac{Y}{c_0 T}\left[\frac{P_w}{2} + tY\frac{l}{L}\left(2\sqrt{c_1 P_{v1}} + P_{v0}\right)\right]} \\ K^* = \sqrt{\frac{P_{v1}}{c_1}k(f^*)} \end{cases}$	Cycle and travel times depend on frequency and demand. One period. Vehicle size is optimised. Operating cost depends on vehicle size. Value of travel time depends on load factor.
	$c_1 \geq P_{v1} \Rightarrow \begin{cases} f^* = \sqrt{\frac{Y}{c_0 T}\left[\frac{P_w}{2} + tY\frac{l}{L}(c_1 + P_{v1} + P_{v0})\right]} \\ K^* = k(f^*) \end{cases}$	

Acknowledgements

This research was partially financed by Grant No. 1010687 from Fondecyt, Chile, and the Millennium Nucleus 'Complex Engineering Systems'. The valuable comments of two anonymous referees are appreciated.

Notes

1. This probability is obtained assuming that the number of passengers to be served at a bus stop is a random variable that follows a Poisson distribution with mean n. Thus, the probability of x passengers wanting to board or alight a vehicle at a bus stop is given by:

$$P(x) = \frac{e^{-n} n^{x}}{x!}.$$

   The probability of a bus actually stopping is equal to he probability of somebody wanting to board or alight, which in turn is equal to one minus the probability of nobody wanting to be served, $P(0)$. This yields:

$$1-P(0) = 1 - e^{-n}.$$

2. It is worth noting that both variables are somehow interrelated, as risk could be diminished allowing for some delays, or little delays regarding a set schedule could be achieved increasing accident risk. This is not taken into account in the model.
3. This treatment of risk and schedule delays is an alternative to Evans and Morrison (1997) seen above.

References

CHANG, S. K. and SCHONFELD, P. M., 1991, Multiple period optimization of bus transit systems. *Transportation Research*, **25B**, 453–478.

CHIEN, S., YANG, Z. and HOU, E., 2001, Genetic algorithm approach for transit route planning and design. *Journal of Transportation Engineering*, **127**, 200–207.

EVANS, A. W. and MORRISON, A. D., 1997, Incorporating accident risk and disruption in economic models of public transport. *Journal of Transport Economics and Policy*, **31**, 117–146.

JANSSON, J. O., 1980, A simple bus line model for optimization of service frequency and bus size. *Journal of Transport Economics and Policy*, **14**, 53–80.

JANSSON, J. O., 1984, *Transport System Optimization and Pricing* (Chichester: Wiley).

JARA-DÍAZ, S. R. and GSCHWENDER, A., 2003, From the single line model to the spatial structure of transit services: corridors or direct? *Journal of Transport Economics and Policy*, forthcoming.

KOCUR, G. and HENDRICKSON, C., 1982, Design of local bus service with demand equilibration. *Transportation Science*, **16**, 149–170.

KRAUS, M., 1991, Discomfort externalities and marginal cost transit fares. *Journal of Urban Economics*, **29**, 249–259.

MOHRING, H., 1972, Optimization and scale economies in urban bus transportation. *American Economic Review*, **62**, 591–604.

MOHRING, H., 1976, *Transportation Economics* (Cambridge, MA: Ballinger).

OLDFIELD, R. H. and BLY, P. H., 1988, An analytic investigation of optimal bus size. *Transportation Research*, **22B**, 319–337.

VICKREY, W., 1955, Some implications of marginal cost pricing for public utilities. *American Economic Review*, **45**, 605–620.