OPTIMAL BUS FARES

By Ralph Turvey* and Herbert Mohring†

This paper† is about optimal fares on buses under first-best conditions. Traffic congestion and the non-optimal pricing of petrol, of parking, of railway journeys and so on are thus simply ignored. They are, of course, enormously important, but they are better understood than the problem on which this paper concentrates.

This problem is to find out what marginal cost pricing of bus services consists of. One apparently obvious answer, which is wrong, is that if a bus run costs \(X\) and carries \(Y\) passengers, and if all these passengers travel the same distance, the marginal cost per passenger and hence the optimal fare is \(X/Y\). This ignores the point that \(X\) is a joint cost. Demand, unlike costs, is a matter of individual passenger journeys; but it is a function not only of fares but also of travel times. These, once the passengers have got to the bus stops, depend on frequency and speed.

PASSENGERS' TIME COSTS

The right approach is to escape the implicit notion that the only costs which are relevant to optimisation are those of the bus operator. The time-costs of the passengers must be included too, and fares must be equated with marginal social costs. This approach has already been put forward in a paper by one of us [1, p. 591]. The present paper differs by not including in the analysis the time cost of getting to and from bus stops, since we here assume their number and location to be given.

As a preliminary, let it be noted that the overall average bus speed depends on the following factors:

(a) Average traffic speed, which influences the speed of the bus between stops.

(b) The rates of deceleration to a stop and acceleration from it.

(c) The time spent at each stop. This depends linearly upon the number of passengers boarding and alighting (in the case of single-doorway buses) or upon either the one or the other (in the case of two-doorway buses). The values of the constant and of boarding and alighting times per passenger depend upon the type of bus and the fare-collecting system. On all these matters see the authoritative Report LR 521 [2].

Take (a) and (b) as given and assume that single-doorway buses are used. Then the extra time added to a bus trip by a stop can be written as:

\[ T + b \cdot B + a \cdot A \]

---

*Economic Adviser, Scicon, when this paper was written. Now Economic Adviser, Bureau of Economic and Social Policies, International Labour Office, Geneva.
†Professor, University of Minnesota.
‡We are indebted to E. V. Morgan, J. K. Welsby, Geoffry Mole, Tony Churchill, and T. M. Coburn for comment on earlier drafts.
That is, the sum of the extra time caused by deceleration and acceleration and of the
deadtime, plus (boarding-time times number boarding) plus (alighting-time times
number alighting). The time spent by any particular passenger from arriving at the
bus stop until he alights at his destination will be the sum of the following items:

(1) Waiting time, which varies inversely with frequency;
(2) Length of journey divided by mean speed between stops;
(3) A time depending on the number of stops made during his journey and the
numbers of other passengers boarding and alighting at each.

This total time is a cost which is borne by the passenger himself. It is the excess of
marginal social costs over this which ought to be reflected in the fare he pays, if the
fare is to help to make his individual journey decision the socially optimal one. This
excess is the increase in journey times which he imposes on other passengers plus any
increase in the bus operator’s costs necessary to preserve other services. If, for
example, the slowing down of a bus would shorten the turn-round time available
before the scheduled departure time of the subsequent run, adding to crew costs, or
would raise fuel costs, then these additional costs of traversing the bus route without
adversely affecting other services constitute part of marginal cost. Thus the excess
marginal social cost consists of the following items:

(I) $b$ times the number of passengers already on the bus when he boards and
boarding behind him at his stop (or $T + b$ times the former when the
bus stops specially for him) plus $a$ times the number of passengers on the
bus when he alights (or $T + a$ times this number when the bus stops
specially for him).

(II) The interval between buses if the bus leaves some one behind when he is
on it (since he then makes that passenger wait for the next bus).\footnote{This is a simplification. If $A$ on the first bus causes $B$ to wait for the second bus, $B$ in turn may make $C$ wait for the third bus.}

(III) The additional cost, if any, to the operator of preserving the quality of his
services for all other passengers.

By boarding or alighting, the passenger will cause some people further along the
route to wait longer. But he will also cause other people to wait for a shorter time,
since they would have had to wait for the next bus if this one had not been delayed.
So, if passengers’ arrivals at bus stops are random in relation to bus timings, these
two effects cancel out on average. (On infrequent services where timetables mean
something to passengers as well as to the operators, arrivals are not random and the
analysis would be different. This paper, however, concentrates entirely on the case
where passengers do not know when the next bus will come.)

FREQUENCY

We are now in a position to look at optimal fares. First let us look at optimal fare
level in relation to the bus operator’s costs, and repeat an important point made in
the earlier paper [1]. This requires us to turn from the marginal condition for the
optimal number of passengers, given the number of bus runs, to the marginal
condition for the optimal number of bus runs, given the number of passengers. One
bus run less will involve a certain gross marginal cost saving which must be equated
with the following marginal effects:

(i) Increased travel times due to greater stop-times and a larger number of
stops caused by more passengers boarding and alighting from each
remaining bus on average.

(ii) Increased probability of each remaining bus being full, so that some
waiting passengers have to wait for the next bus.

(iii) Increased costs of maintaining the services provided by the remaining
buses as a result of their increased stop times and increased number of
stops.

(iv) Increased waiting time. (Waiting time on average equals half the interval
between buses, and this interval is increased by the subtraction of one
run.)

Now suppose both that the fares are optimal, so that the number of passengers is
optimal, and that the number of bus runs is optimal. Then we can say that the
value of \([ (I) + (II) + (III) ] \times \) (number of passengers per run) is equal to the fare
revenue per bus run. We can also say that the value of (i) to (iv) is equal to the
marginal cost of a bus run. So if the value of \([ (I) + (II) + (III) ] \) times the number
of passengers per run were the same as the value of (i) to (iv), the fare revenue from
a bus trip would just equal the marginal cost of a bus trip. But they are not the same.
There is an asymmetry between the disbenefits of more passengers with the same
number of bus runs and the disbenefits of fewer bus runs with the same total number
of passengers. Fewer buses mean increased waiting time because of a longer average
interval between buses. Items (i), (ii) and (iii) correspond with (I), (II) and (III),
but item (iv) corresponds to nothing. Hence \([ (I) + (II) + (III) ] \) times the number
of passengers falls short of (i) to (iv), and this means that the optimal fare revenue
from a bus run falls short of marginal cost.

The point then is that, even if there are constant private costs to scale of bus
operation, there are decreasing social costs. If the number of bus runs and the number
of passengers both went up by \( x \) per cent, total waiting time would not go up by
\( x \) per cent. So marginal social costs are below average social costs. We have a classical
case for subsidy in order to achieve optimal resource allocation. This case has nothing
to do with congestion. It is just coincidence that considerations arising from con-
gestion also point in the same direction, to subsidising buses or to taxing or restrain-
ing private car use. Conversely, a failure to subsidise buses or tax car-use sufficiently
accelerates a shift from buses to cars for two quite separate reasons. One is a matter
of congestion costs. The other, which is the subject of this paper, is that a decline in
the total number of bus journeys which results in fewer runs causes a further decline,
frequency being an important factor in determining demand. Less means worse and
worse means less.

In explaining marginal social less private cost of a bus passenger in time terms as
the sum of (I), (II) and (III), we are looking at marginal costs when the size of the
system is given. If we wish instead to look at costs when the system size is adjustable,
i.e. to examine long-run rather than short-run marginal costs, we immediately meet
the difficulty that the system cannot be adjusted to provide for just one passenger

282
more or less. The correct way round this is to say that optimal fares should always be determined by “short-run” marginal costs and that this pricing rule has to be supplemented by an investment rule. Our investment rule was formulated above in terms of equating the value of (i) to (iv) with the marginal costs of an extra bus run along this route. Point (iv) is the effect of the increase in frequency in decreasing waiting times, so we can say that the value of this decrease must be subtracted from the cost of the extra buses, crews, fuel, tyres, etc., in order to get the marginal social cost. It is the failure to make this subtraction which vitiates the proposition referred to at the beginning of this paper.

The distinction between the (short-run) pricing rule and the (long-run) investment rule may help to explain a difficulty about operating costs which bothered a reader of an earlier version of this paper. By “investment” we have meant decisions about system size, including how many runs are made along each route during each period, so the term “long run” is rather inappropriate. But, terminology aside, a rise in, say, fuel costs does not of itself dictate a rise in fares according to the pricing rule. This may seem baffling. Yet the fact is that one passenger more or less will cause no noticeable change in fuel consumption, so that the price of fuel is not directly relevant to the optimal fares for influencing passengers’ individual decisions. However, the rise in fuel costs may, by the operation of the “investment” rule, make some runs on some routes extra-marginal, so that they are suspended. If the consequence is that bus loadings increase, the operation of the pricing rule will then dictate some fare increases. So a rise in fuel costs will result in a rise in optimal fares after all. Fuel costs, crew wages and the cost of new buses all affect system size; this helps to determine bus loadings, and it is bus loadings which are relevant in setting fares. We now consider how fares should be set in more detail.

TIME COSTS IN OPTIMAL FARES

It must be recognised that a difficulty of the pricing rule is that it requires valuations of passengers’ time, both waiting at bus stops and when the buses are in are delayed. For cost-benefit analysis in transport it is customary to use standard uniform values, waiting time being treated as, say, three times more valuable than time spent in a bus. While this is a sensible pragmatic procedure, we have to recognise that it is a lack of uniformity, that is to say a dispersion round the average, which partly generates the problem in the first place! An important reason why reduction in fares or increases in service frequency will attract new passengers is that new passengers implicitly attach higher values to time than the existing passengers.

Having noted this complication, we have to ignore it in discussing the structure of optimal bus fares in terms of the time costs. These (to remind the reader) are:

(I) Delay caused to passengers on the bus by boarding and alighting;
(II) Delay caused to people waiting at bus stops because the bus is full.

An idea of the possible magnitude of these items can be obtained by using data from the admirable investigation by Cundill and Watt of boarding and alighting times [2].

A similar point applies fairly generally to all optimal pricing by public enterprises, as explained in [3] chapter 7.
Consider first the ordinary two-man red London buses with a single open entrance at the rear. In the peak hours these buses carry 38 passengers on average and have a typical service frequency of 4 minutes. Their stop-time in seconds is 0.95 plus 1.15 times number boarding plus 1.0 times number alighting. If to this dead-time of 0.95 seconds we add a numerically convenient guess of 20.05 seconds for the extra time involved in decelerating to a halt and accelerating from rest instead of carrying on, then, in terms of the symbols used earlier, we have $T = 21$ seconds, $b = 1.15$ seconds and $a = 1.0$ second. Thus the marginal time-costs imposed on others by a passenger are:

(I) $38 \times (1.15 + 1.0) = 81.7$ seconds for other people in the bus plus

$38 \times (0.95 + 20.05) = 798$ seconds if the bus stopped specially for him.

(II) 4 minutes’ extra wait for any person he prevents from getting on the bus by being on it when it is full.

Different results are naturally obtained with other types of buses and other fare systems. The one-man-operated flat-fare Red Arrow in London, for example, has a stop-time in peak hours equal to whichever is the greater of the following:

$5.65 + 3.3$ times number boarding;

$8.2 + 1.4$ times number alighting.

**HIGHER FARES ON FULL BUSES**

Since optimal fares depend on time-costs, which depend partly on fare-collection systems, which both determine what fare structures are possible and help to determine costs which affect the optimal system size, it is clear that complicated interdependencies are involved. But, whatever bus type is used, one primary feature of optimal fare systems is obvious enough. That is that the fare should be higher:

(α) The greater the expected number of people on the bus when passengers board and alight;

(β) The greater the probability that the bus is full when there are people at bus stops who want to get on it.

Although this is a simple principle, it is a fruitful one to apply. Firstly, it means that fares should be higher in the crowded direction, i.e., inward to work and outward in the evenings, than in the reverse direction. It is, incidentally, interesting to consider how cities would have developed if the rule had always been applied. Secondly, it means that peak fares should exceed off-peak fares. It is, however, less clear whether weekend and evening fares should be low, since higher wage costs may cause these services to have such high operating costs that the operation of the investment rule causes them to be few and hence crowded. Thirdly, and this is the most novel conclusion, fares should be positively related to distance only when the probability of buses being full is non-negligible along the whole route. This is because item (α) above has nothing to do with the distance travelled; only (β) has. A passenger travelling two crowded miles is twice as likely to prevent a waiting passenger from boarding as a passenger travelling only one crowded mile.

---

4There are possible exceptions to this. The probability of an extra passenger causing a bus to stop specially for him may be greater when flows are low than when they are high, as pointed out on page 599 of Mohring’s paper [1].

284
To illustrate all this, consider the example of the main direction of travel during the peak. We can draw a curve showing the average number of passengers on the inward buses during the morning rush-hour on the route segments bounded by seven stops. The fares from 1, 2, 3, 4, and 5 to 7 should be successively higher because at each of them the boarding time of an additional passenger will delay successively more existing passengers. The boarding delay imposed at 6, however, is below that at 5, so the fare from 6 to 7 should be below that from 5 to 7.

All journeys, including segment 5 to 6, involve a probability of imposing cost type (β), i.e. more waiting delay.

Taking both (α) and (β) costs together, therefore, optimal fares are not positively related to distance in this example. Not only should the fare from 3 to 6 fall below that for the shorter distance 5 to 6; both should exceed the fare for the longer distance from 1 to 5. What matters is not distance, but how full the buses are. The highest fares should be for journeys going past those stops where queues sometimes last longer than the interval between the buses.

For the outward journey in the morning (α) is lower than on the inward journey and (β) is negligible. Hence a very low fare is justified. In the evening peak everything is the other way round.

The example relates only to a radial feeder route. However, at least along the routes studied by Cundill and Watts [2], no such systematic variations were observed in the numbers of passengers on buses. As explained above, a positively distance-related element does then come into optimal fares. But the fares should vary less than proportionately with distance, since the (α) element is not distance-related.

**REVENUE CONSTRAINTS**

In practice, no bus undertakings receive general subsidies calculated along the lines which our analysis would suggest. Many, indeed, are supposed at least to break even. Hence in practical terms their aim to maximise welfare, i.e. (under our first-best assumptions) willingness to pay less costs, is constrained by a net revenue requirement. In such cases, the pricing rule involves departing from the optimal tariffs in the way that brings in the extra net revenue with the least effect on resource allocation. So fares could be designed by first constructing an optimal fare structure and then modifying it so as to meet the financial constraint. Alternatively, and more realistically, one might begin with an existing non-optimal fare structure which does meet the financial constraint, and then improve it by putting some fares up and others down in a way which does not spoil financial performance but as far as possible reflects the principle suggested here.
In either case, some of the common differences between actual fare structures and unconstrained optimal ones appear to be capable of rationalisation in simple "what the traffic will bear" terms. In particular, it seems plausible that passengers going in the opposite direction to the main stream are able to pay as much as the majority of passengers. Off-peak passengers, on the other hand, may be more sensitive to fare levels than commuters are; so off-peak fare concessions are much more common. The conflict between distance-related fares and flat fares, finally, involves additional complications in the choice of fare system, which can have a very large influence on costs. It is interesting to recall that many years ago London tramways, which had distance-related fares, also had a flat twopenny fare between the hours of 10 in the morning and 4 in the afternoon. This fitted in very neatly with the second and third implications of the principle put forward above.

Finally, let us revert to the investment rule. The number of bus runs on a route constitutes only one of many types of decisions about the service to be provided. The other types include answers to such questions as:

What time in the morning should the service start?
What time in the evening should it cease?
Should the route start nearer to the centre of town?

The effect on total passenger time-costs of different answers to one of these questions will vary from case to case. To determine the optimal extent of the service is thus as difficult as determining the optimal frequency and fares for a given extent of the service. Furthermore, the other aspect of the investment rule has been passed over rather lightly in this paper: that is, the analysis of the bus operator's costs. In discussing the optimal number of bus runs along a route during a certain period, we assumed that the marginal cost of one extra bus run was known. In fact, indivisibilities in hiring of crews and, still more, in the purchase of buses mean that crew costs and annual bus costs are joint over a large number of runs. Investment decisions are thus far more complicated than they have been represented as being. The marginal cost of a single run (per-run costs such as extra fuel apart) largely consists of the shadow-values of crew-time and bus-time, which can only emerge from system-wide optimisation. But failure to explore these matters does not impugn the suggested pricing rule. The fuller the bus, the more should a passenger pay for getting on it, being on it and alighting from it.

REFERENCES


---

3It is what Littlechild calls a mutatis mutandis marginal cost, as explained in [3], chapter 7.

286