Threadbo conference in Molde 25-28 June 2001

EFFICIENT MODAL SPLIT

by Jan Owen Jansson, EKI
Linköpings University, Sweden

Contents

1. Background and purpose 2
2. Modal split and basic characteristics of public and individual transport 3
2.1 Present modal split by distance and by volume 3
2.2 Economies of density of demand for scheduled public transport 5
2.3 Diseconomies of density of demand for urban car transport 6
2.4 The reform potential 7
3. From car to bus in the central city 9
3.1 General theory of the price-relevant marginal cost of scheduled public transport 10
3.2 Applications to urban bus services 11
3.3 Separate track for buses 12
4. Peak-load pricing of scheduled public transport: purpose and potential 14
4.1 Peaking problems in the time dimension 14
4.2 Spatial peaking problems 15
5. Solution to daily peaking problem – peak-load pricing of commuter transport by bus 16
5.1 The price-relevant marginal cost of peak trips 17
5.2 The price-relevant marginal cost of off-peak trips 18
5.3 Numerical example of the optimal structure of bus fares 20
6. Solution to weekly peaking problem – peak-load pricing of interurban train services 21
6.1 The price-relevant cost of passenger train services 21
6.2 The optimal structure of train fares 22
6.3 Numerical example of three different lines 24
7. Refutation of the main objections to public transport subsidization I: “The cost of public funds” 26
7.1 The excess burden of different taxes including prices exceeding the marginal costs 26
7.2 Implications for cost-benefit analysis and pricing policy in transport 30
7.2.1 The cost of public funds for the purpose of public production 30
7.2.2 The cost of public funds for the purpose of transfer payments to households 31
7.2.3 The cost of public funds for the purpose of subsidizing decreasing-cost industries 31
8. Refutation of the main objections to public transport subsidization II: towards achievement of both allocative- and X-efficiency in public transport 32
REFERENCES 33

EFFICIENT MODAL SPLIT

1. Background and purpose

A lot is said at this conference about ways and means of improving the cost-efficiency of public transport. Two main possibilities for improvement are to stimulate competition, and to enhance the motivation and creativity of operators by introducing the profit motive into a traditional "public service".

The question is, if the present *allocative inefficiency* in transport markets will also be improved in the process? This paper is meant to serve as a counterbalance to the main preoccupation of the conference, by looking at the present allocation of total travel between in the first place individual car and public transport services, and identifying the main reform potential for beneficial changes of the modal split. And it is argued that these changes will not be brought about by the increased reliance on market forces. On the contrary, better planning of public transport systems, and, I dare say, continued or increased subsidization are two necessary conditions for realizing the potential improvement of the resource allocation. A complementary, significant point is, however, that there is no inevitable conflict between the ambition to increase cost-efficiency in public transport, and a transport policy towards an efficient modal split. The paper ends by pointing out a fruitful, strategic new area of research: how to design a system of subsidization of decreasing-cost public transport, which makes profit maximization on the part of the operators and social surplus maximization coincide?

2. Modal split and basic characteristics of public and individual transport

In this introductionary section the salient features of the present modal split in passenger transport is held up, and the cost characteristics of, on one hand, scheduled public transport (SPT), and, on the other, individual car transport, which are the main explanatory factors for the observed great differences in the modal split on different routes.

2.1 Present modal split by distance and by volume

In a country like Sweden car transport is dominant, but there are important niches for other modes of transport at either end of the spectrum of trip distances. For long-distance transport *speed* is obviously the main quality of service that makes a difference, which can break the dominance of car transport.

But speed has typically a price in terms of ready access to the transport facility. Airports and high-speed railways should not be too close to where people live, or spend their leisure time. Therefore the feeder transports (by another mode of transport) are relatively long for the fast modes, which means that a natural division of the non-urban transport markets by distance has come about. For personal transport this is illustrated in fig 1, where the model split of personal transport longer than 10 km in Sweden is depicted. Road transport is completely dominating in the distance range between 10 and 100 km. For longer distances the railway transport share is steadily growing with distance. Around 300 km, air transport starts to make an impact. For very long distances, in the region of 1000 km, air transport is market leader both for travel time reasons and geographical necessity.

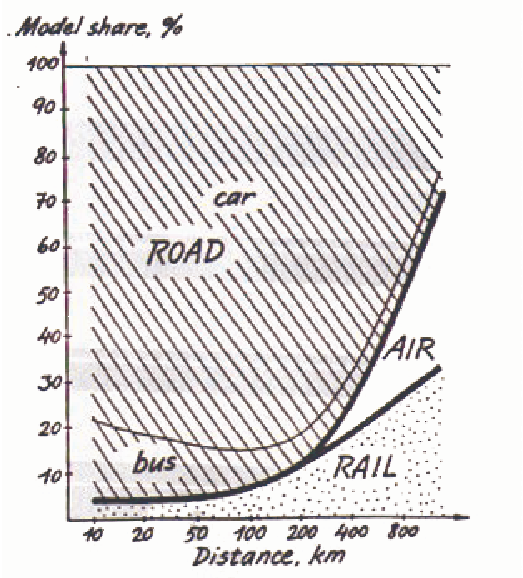


Figure 1 **The modal split of personal transport by distance**

With the introduction of high-speed trains (after the diagram was originally drawn), rail transport seems to make inroads in the long-distance markets for especially business travel, where air transport has been dominating.

Besides the technical division of the transport sector by mode, i.e. by the bearer of transport vehicles, a complementary economic outlook, focusing on the organizational form of transport production, shed further light on the competitive conditions in the transport sector, and, in particular, urban transport which is not covered by the diagram of figure 1.

The three main forms of organizing transport are (1) to use your own vehicle, (2) to hire a (whole) transport vehicles and (3) to buy a share of the space of a vehicle in scheduled public transport (SPT) service. The most significant front line, as far as competition is concerned, is between do-it-yourself car transport and SPT by road, rail and air.

For distances where the SPT share is as low as around 20%, which is the case in a wide middle-distance range, it can be concluded that mainly so called "captive riders" are using the SPT alternatives to private car travel. As long as there is an appreciable number of travellers without a car at their disposal, railway and/or bus transport services are viable, but the generalized cost (GC) of car travel is definitely lower, which is very obvious when the travellers make up of group of at least two persons per car.

For short-distance urban travel commuter train services can be very competitive both on account of their own strengths, and of weaknesses of its rival: trains are faster, and less expensive than the private car for commuters with monthly passes, by which the average monetary cost per trip is typically below 1 euro, irrespective of distance, and the marginal monetary cost is zero. The general competitiveness of short-distance urban rail transport is maintained by considerable subsidies (by local taxpayers) to operators of commuter train services. However, in case the car travel alternative would involve the payment of a parking charge adjusted to the conditions of the market in the central city, that alternative would be inferior in terms of generalized costs at almost any distance, also without subsidization of public transport.

In table 1 below the car share in the total travel by motor vehicles is given for different segments of the total transport market of Stockholm. As seen the car share range from 28% for commuting between the inner suburbs and the central city in the rush hours to 72% for travel between the outer suburbs. The latter value is definitely a corner solution, i.e. it is determined by the rate of car ownership and car disposal rather than the relative GC for alternative modes of transport.

Table 1 **Car share of total motorized trips by road and rail in Stockholm**

MARKET SEGMENT	Time period	Percentage
Travel within central city	All day	50%
Inner suburbs to central city	Rush hours	28%
Outer suburbs to central city	"	31%
Inner suburbs to inner suburbs	All day	52%
Outer suburbs to outer suburbs	"	72%

In urban areas it is the volume of travel rather than the travel distance, which is the main modal split determinant. It is common to consider the relationship between the generalized cost and *route volume*, but a wider perspective is obtained by relating GC to the *density of transport demand* of an area, as will be demonstrated presently.

The point is that an urban car transport system has a much lower capacity than a public transport system. This means that urban car transport is eventually an increasing-cost activity, while an urban public transport system is a decreasing-cost activity almost indefinitely. This

is the main explanation for the wide variations in the modal split of different urban transport market segments exemplified in table 1 above.

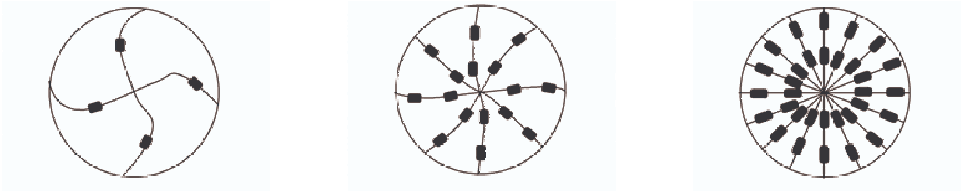
2.2 Economies of density of demand for scheduled public transport

Given the transport infrastructure for a particular SPT system, the following two cost relationships give rise to marked traffic volume economies in SPT-service production.

1) The SPT-service producer cost per passenger (or freight ton) is steeply falling with vehicle *size*. The main reason for this is that the driver, or crew cost per vehicle is either constant, or is increasing markedly degressively with vehicle size, and in the second place the vehicle capital and running costs per passenger (or freight ton) kilometer is decreasing with increases in vehicle size.

2) The SPT-service user cost per trip (in the case of passenger transport) is falling with vehicle *number*: the more vehicles there are in the SPT system, the less waiting time and/or access time (walking time in local transport) are required per trip.

The "Mohring effect" has, in principle, both a time and space dimension. Therefore it is instructive to consider a system, or network of SPT-services, and not just a line. The best, simple illustration of what it is all about is the Circletown model. (Jansson, 1997)



(1) four small buses ***(2) Sixteen mediumsize buses*** ***(3) Sixty-four big buses***

Figure 2 **Buses and bus lines in Circletown for three different levels of the density of demand**

The economies of density of demand takes a number of expressions:

- more bus lines means less access time
- when bus lines are denser, each line can be straighter to save travel time
- more buses on each line means less waiting-time at bus stops
- buses should be successively bigger as the density of demand is increasing, which will reduce the bus operator's cost per trip.
- bigger buses can have a higher rate of occupancy with impunity as to queuing time for passengers.

A numerical example of the decreasing cost character of urban bus transport is calculated by the Circletown model (Jansson 1997). Comparing the top and bottom rows in table, it is seen that as the density of demand is increasing from a very low level (100) to a high level (15.000), obtainable when the majority of all trips longer than "walking distance" are made by bus, the total producer and user cost per trip goes down to a third of the initial level. The frequency of service goes up from one bus every twentieth minute to one bus every fourth minute, and the line density is also multiplied by five. It is notable that AC_{prod} falls as much as AC_{user}. This is mainly

achieved by increasing the bus size. A minibus of a maximum of 20 passengers is optimal in a situation with extremely low density of demand, and large buses taking more than one hundred passengers is optimal in the opposite extreme case.

Table 2 **Optimal design and generalized costs of bus transport in Circletown at different levels of demand**

DENSITY OF DEMAND:	OPTIMAL DESIGN OF BUS SERVICES			AVERAGE TRIP COST, Euro		
Number of trips generated per km^2 and day	*Bus size*: number of seats	*Frequency*: buses per hour	Average walking distance to/from stops (meter)	Producer cost including externality charges of 4 km bus trip, AC_{prod}	User cost of walking to/-from bus stops, waiting, and riding 4 km by bus, AC_{user}	Total cost of 4 km bus trip, AC
100	*20*	*3*	*350*	1.57	3.62	**5.19**
250	*27*	*4*	*260*	1.20	2.88	**4.08**
500	*34*	*5*	*210*	0.98	2.45	**3.43**
750	*39*	*6*	*180*	0.88	2.24	**3.12**
1 000	*43*	*6*	*160*	0.81	2.11	**2.92**
2 000	*54*	*8*	*130*	0.68	1.84	**2.52**
4 000	*68*	*10*	*100*	0.57	1.62	**2.19**
8 000	*86*	*12*	*80*	0.49	1.45	**1.94**
15 000	*105*	*15*	*60*	0.43	1.33	**1.76**

Source: Jansson, 1997

It is worth emphasizing again that the Mohring effect, i.e. the economies of number in the user costs are not enough to give rise to the markedly decreasing costs. Economies of vehicle size in the producer costs is another necessary condition, because without these economies it would be possible to nullify the Mohring effect by employing very small buses in such a large number that walking and waiting time would become trivial.

2.3 Diseconomies of density of demand for urban car transport

Car transport in a given urban road network will sooner or later show decreasing returns. In many urban relations the potential demand is low enough for making it possible to carry out all the trips demanded by car, if everyone had a car at his/her disposal: car cost with respect to the route volume is constant for all practical purposes. On the main routes into the central city, on the other hand, only a fraction of the total travel demand could be carried out by individual car transport. The limited capacity will be manifest by a steeply rising GC_{car} well before all demand is met. On the assumption that total demand along a particular route is (by and large) given, the GC_{car} function represents the demand for the alternative mode of transport, say bus, on the route concerned.

In the three diagrams of figure 3, the values of the generalized cost of bus trips in Circletown in table 2 are the basis for the falling GC_{bus} curves. In diagram (a), where the trip volume is measured in thousands, the whole range of values in table 2 is represented: the total travel goes from 0 to 30.000 trips per km^2 and day. In diagram (b) where the trip volume is

measured in hundreds, the end point of the GC_{bus} curve is thus at a level of trip volume of 3000, and in diagram (c) where the trip volume is measured in tens, the end point is at a level of trip volume of 300.

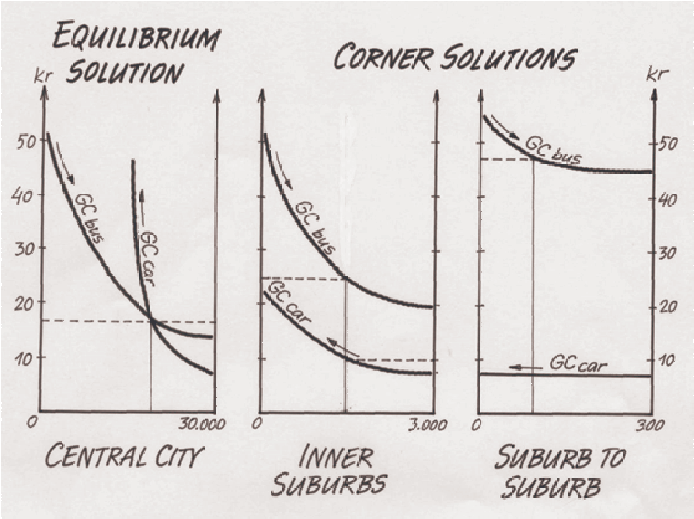


Figure 3 **Modal split and density of transport demand**

The curves for GC_{car} with respect to trip volume are going in the wrong direction, from right to left in the diagrams. In diagram (a) GC_{car} is steeply rising, because the road capacity is insufficient to accommodate a car trip volume above 15 000 trips per km^2 and day. Diagrams (b) and (c) represent urban areas of much sparser population, where it would be physically possible to make all trips by car. As long as the parking is free or cheap, GC_{car} will be below GC_{bus} in the whole trip volume range represented by diagrams (b) and (c), and the modal split is simply determined by car disposal as a "corner solution".

One important market segment is not represented among the three diagrams above: the commuter traffic between the suburbs and the central city. Commuter traffic by car to the central city is high in absolute terms, but low relative to the public transport volume. The equilibrium mechanism is more complicated than is indicated by diagram (a) of figure X above, because there are two capacity limitations which each can be the most telling one for different categories of car commuters: the capacity of the "entrances" to the central city (rather than the roads as such between the suburbs and the central city), and the parking space in the central city. For commuters spending 8 hours at their place of work, the real cost of the parking space for a car used just for the work trip can be of the same order of magnitude in big cities as the capital cost of the car itself.

2.4 The reform potential

Allocative inefficiency in the sense that price is more or less different from the price-relevant marginal cost abound in the transport sector. This is true about individual car transport as well as public transport. This does not mean, however, that a very inefficient modal split is omnipresent. In the wide middle-distance market segment price adjustments to fulfil strict efficiency conditions would have only a small effect on the modal split.

The main potential for policy changes that would make a difference as regards allocative efficiency is to be found at both ends of the "spectrum" of figure 1, where the distribution of

trips between individual car transport and various modes of scheduled public transport (SPT) is already relatively even.

The even split is a sign that an “interior solution” applies, where $GC_{car} \approx GC_{SPT}$, rather than a corner solution where $GC_{car} \ll GC_{SPT}$.

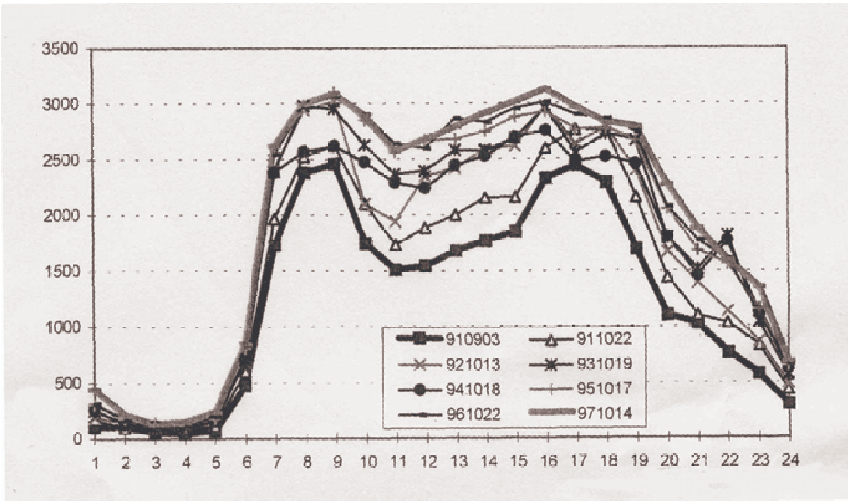
In what follows three market segments are discussed more in detail, where the potential seems high for substantial increases in the SPT share with consequent welfare gains.

1) Travel in the central city all day
2) Urban commuter transport
3) Interurban railway passenger transport

3. From car to bus in the central city

Of the three market segments pointed out above, travel within the central city differs from the other two trip markets in so far that peakiness of demand is a minor problem. In chapters 4 and 5 peak-load pricing is the main policy measure discussed. Here a combination of pricing policy and regulation of street space is the key to the goal of social surplus maximization.

The traffic structure in the central city during the work-day is markedly different from the rush hour traffic between the suburbs and the central city. The latter is now the main task for the public transport. Outside rush hours the level of demand is much lower for the public transport system, which is a main reason for its financial problems. The total demand for travel in central city during office hours, on the other hand, is fairly evenly spread over the day. This is well recorded as regards car traffic. The time-profile for the car traffic in the central city has lately become rather flat from early morning to late afternoon, which is illustrated in figure 4 below. This is partly explained by peak traffic spreading into the time period between the morning and afternoon peaks, and partly by the fact that the commercial traffic has its peak during the workday rather than before and after. In "Trafiken i Regionplan 2000" it is established that the share of taxis, LDVs, service vehicles, and cars for business trips is very high in the central city road traffic during office hours, up to 50%.



Source: Transek (2000)

Figure 4 **Hourly time-profile of main road traffic flow in the central city of Stockholm on an autumn day of eight different years (1991-1997)**

Personal transport involving the carriage of heavy tools or bulky parcels, etc, are "captive" car traffic, but there is a potential for bus traffic in the central city to win over other kinds of business trips as well as most private travel. If proper road pricing is introduced, this would, of course, help a lot. An even more important step to take in the market for car travel is to tighten up the parking policy. A car trip within the central city requires two different parking spaces: one may be the "base" for the car, so to speak, at the owner's place of work or residence. This would often belong to the "reserved" parking market segment. The other parking place, however, would normally have to be in the open market segment, in the street,

or in a commercial parking lot (outside, or inside a multi-storey car park). If the commercial level of parking prices could be ruling everywhere, an incentive at least as strong as proper road pricing to discourage car use in the central city would exist.

For the public transport system serving the central city, the Circletown model above indicates the potential for lowering GC_{bus} towards the level of GC_{car}, provided that (1) the goal of social surplus maximization is adopted, and (2) that in case proper road pricing is lacking the buses have their own exclusive lanes, because if the buses and cars are crowding together, it is the former which are the main losers.

A necessary condition for social surplus maximization is that the bus fare is set equal to the price-relevant marginal cost. The theory of optimal pricing of scheduled passenger transport is outlined in the next two sections. First come a more general discussion, and then the specific case of urban bus services is addressed.

3.1 General theory of the price-relevant marginal cost of scheduled public transport

The basic formula for the price-relevant marginal cost of SPT-services is written like this:

$$MC = MC_{prod} + B\frac{dAC_{user}}{dB} \qquad (1)$$

The number of trips by the public transport system concerned, is denoted B. The external marginal cost appears as charges on the vehicles payable by the public transport company. In case the public transport vehicles are buses in an urban road network, they should pay congestion tolls and externality charges just as cars and other motor vehicles for the costs they cause by making use of road space. These charges are part of MC_{prod} in (1), and will consequently be passed on to the public transport passengers included in the fares.

Although the two pioneering classics in the field of SPT pricing principles – Mohring (1972), and Turvey and Mohring (1975) – were couched in what the authors chose to call short-run marignal cost (SRMC) pricing terms, it is here argued that in order to gain substantial additional insight, one should make a departure from the road staked out in those classic works, by regarding the number of transport vehicles as variable in the costing and pricing analysis. Whether or not this is a departure from the golden SRMC-pricing rule is a matter of semantics, which we shall not go into here. The important point is anyway that, no matter which design variable, or which combination of design variables are adjusted to meet an increase in demand, the price-relevant marginal cost should come out the same. (Jansson, 1984)

The two price-relevant cost components of (1) above will take quite different values depending on in which way additional passengers are taken on. However, in optimum the sum of the two components is the same irrespective of how capacity is augmented. The symbolic production function below can be used as a basis for discussing this central aspect.

$$B = f(N, S, V, H, \phi) \qquad (2)$$

B = number of trips
N = number of equisized SPT-vehicles
S = vehicle size in terms of holding capacity (i.e. the maximum number of passengers)

V = speed
H = handling capacity, i.e. the number of passengers boarding and/or alighting per unit of time
ϕ = occupancy rate (= holding capacity utilization)

The expansion path is defined by such combinations of the design variables in (2) which minimize the total cost TCprod + TCuser for every level of output, B. Along the expansion path the price-relevant cost, MC is independent of which (single) design variable, or combination of design variables are changed when calculating this cost. The relative order of magnitude of its two components can, however, be very different, which should be carefully noted, in view of the fact that only one of them is a producer marginal cost. Let us illustrate this point by some examples:

(1) Additional passengers (ΔB) can normally be accomodated (almost) without any additional producer inputs, simply by increasing the occupancy rate, Φ. However, in particular in peak periods this cannot be done with impunity as regards the user costs. Hence MC will in this case solely consist of an occasionally high user cost component, representing quening and/or crowding costs of the passengers.

(2) Another, more regular way of accomodating additional passengers is to increase the number of vehicles (N). In this case the MCprod component will be fairly substantial, while the user cost component becomes negative due to general economies of vehicle number in the user costs.

(3) A third possibility is to increase the size (S) of vehicles in order to take on more passengers or freight. This would leave the user cost component in the price-relevant cost by and large unchanged, and only MCprod contributes to MC.

3.2 Applications to urban bus services

In the case of urban bus transport the number of vehicles is the most practical factor to increase. Together with the old marginal cost proxy "the average cost of the marginal plant" a simple, and yet robust MC-expression is obtained, at least for an urban bus transport system containing a good number of vehicles. The two terms of (1) above can consequently be specified like this in the case of urban bus services:

$$MC = \frac{\Delta C}{\Delta B} - \frac{vB\Delta t}{\Delta B} \qquad (3)$$

ΔC = incremental cost of the bus company for putting in another bus in operation
B = existing number of passengers
ΔB = number of new passengers carried by the additional bus
Δt = waiting-time saving per trip by existing passenger
v = value of one minute waiting-time saving

The price-relevant cost and the optimal fare should be differentiated in the first place between peak and off-peak periods. However, as was pointed out before, in the central city trip market, this feature is not very prominent, so this exercise will wait till later (section 5) where public transport for commuters is discussed.

As seen in (3), the first term in the two-term expression for MC is a proxy for MCprod, which would be fairly close to the average cost of the bus operator, ACprod. In addition the "Mohring effect" has to be taken into account. The second term of (3) representing the Mohring effect is a negative cost, i.e. a benefit, which makes the optimal fare level fall well below ACprod. In section 2 the results of the Circletown bus service optimization model were presented. In this model it was found that along the expansion path, which inter alia implies that bus size is optimally adjusted to different levels of the density of demand, MC is roughly constant. Since ACprod is steadily falling along the expansion path, the financial result of optimal pricing is that total cost recovery is very low for low demand densities, and grows successively to a maximum of about 50% for a very high demand density.

The principal problem of optimal bus transport pricing has to do with vehicle size rigidity. If an existing bus fleet consists of buses of markedly inoptimal sizes, it can be difficult to make the right adjustments quickly, and the application of formula (3) may give some odd results like negative fares.

Another complication not mentioned so far is the fact that passengers put two different demands on transport vehicle capacity – demand for space in the vehicle (a seat), and demand for vehicle time during the act of boarding and alighting. In principle, the optimal fare has two components:

(1) the space occupancy charge
(2) the boarding/alighting charge,

of which the latter is normally the least important, but computationally the most complicated item.

3.3 Separate track for buses

The ideal situation to aim at in the central city travel market is something like that illustrated in the left-hand diagram of figure 3 in section 2: an equilibrium solution where $GC_{bus} \approx GC_{car}$. If such a position is obtained, and in addition the part of each GC made up of the monetary price is equal to the price-relevant cots, it would be perfect. The opposite, worst case is illustrated below, which unfortunately prevails in many cities in reality. Buses and cars are crowding together in the streets in the absense of any form of reasonable road pricing. The situation is bad for both, but the buses come out worst, because they are relatively large and clumsy, and have to move in and out of the slow-moving traffic to let out and pick up passengers at stops. Under these circumstances GC_{car} is all the time well below GC_{bus}: only those without a car at their disposal take the bus.

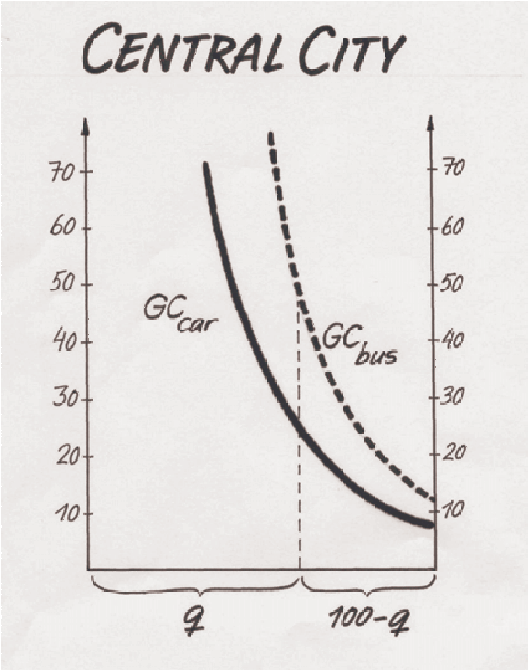


Figure 5 **Cars and buses are crowding together**

In the absense of proper road pricing, which could eliminate the congestion, a second-best solution is to provide separate track for the buses. This would take away a substantial amount of the precious road space for the cars, but it is in line with the first-best solution, where substantially more buses would run, and a good part of the present car traffic would be gone.

4. Peak-load pricing of scheduled public transport: purpose and potential

Apart from travel in the central city, the transport market segments where SPT is most important are characterized by big demand peaking problems both in time and space. The basic idea of peak-load pricing is to level out the peaks and troughs in the demand profile in each particular market in order to save capacity costs, and raise the rate of capacity utilization in off-peak periods. To an economist the rationality of peak-load pricing is self-explanatory, but to other people, who look at the ebb and flow of traffic almost as a natural phenomenon, it is not at all obviously rationale. If the demand peaks are due to a "higher order" of work and leisure organization it is pointless to try to change it; the main result would be a wholesale redistribution of income from peak travellers to off-peak travellers.

The elasticity of demand is apparently very important – the own-price as well as the cross-elasticities. The problem is that only some short-run elasticities are reasonably well known from transport operators' experiences and econometric studies of public transport demand. (Comprehensive surveys of transport demand elasticities are Goodwin, 1992, and Waters and Oum, 1992). The long-run elasticities are what we would like to have. These would tell us whether there is significant potential for adjustment of work starting times, times for vacation, leisure travel habits etc.

4.1 Peaking problems in the time dimension

The greatest reform potential of the pricing of SPT-services is in the differentiation of the fares structure by time of the day (urban commuter transport), day of the week (interurban transport), and season of the year. A useful rule-of-thumb roughly valid in many places is that interurban rail fares in off-peak (Monday, Tuesday, Wednesday, Thursday, and Saturday) should be so low relative to the Friday and Sunday fares level that the daily demand is the same all week.

It is also interesting to note that the marked weekly peak of interurban travel is reversed so far as intraurban travel is concerned. In big cities the vast public transport systems – bus, over- and underground train services – is working well below capacity during weekends. Bearing in mind that the interurban transport sytems are strained to the utmost in different critical hours during Friday and Sunday an interesting possibility is that much lower fares for intraurban weekend travel and higher fares for interurban travel in connection to weekends would make people stay more often in their hometowns, also at weekends and during holidays.

A similar difference is to be found in the seasonal time profile between urban and interurban travel demand. In summertime, and particularly during the general vacation weeks, urban traffic – car traffic as well as public transport – is ebbing. (Relative exceptions to this rule are some unique tourist cities like Paris and London). Summertime is instead high tide in the non-urban transport systems, with a possible exception for domestic airlines, where business travellers make up almost two thirds of the total patronage during autumn, winter and spring. Airlines rightly try to compensate the large drop in the travel of their main customer category, the business travellers, from mid-June to the end of August (in Sweden) by very substantial discounts off economy class fares. The Swedish Railways (SJ) fears that airlines would fill the empty chairs largely by former train passengers, unless SJ responds by a similar offer of substantial discounts off the ordinary rail fares. It should be remembered that train travel

demand would be at its highest during the Summer months even in the absence of fare discounts. The result is now that trains are overfull. This is peak-load pricing in the reverse. It resembles anomalies like sale in December during the Christmas rush.

4.2 Spatial peaking problems

The geographical peakiness of SPT demand is another reason why the optimal price structure is at least as important as the optimal price level.

The cause of the spatial peaking problem is the multi-product character of SPT-services. For a number of good reasons a busline, or train service does not only produce transport from A to B, but also from B to A, and in the normal case, where one or more stops are made underway, between a good number of places on the route between A and B. It is very unusual that the structure of demand is such that the passenger flow is constant all the way from A to B. The normal pattern is instead that after setting out from A an accumulation of passengers on the bus or train occurs to begin with up to "the critical section", where the expected passenger flow is at a maximum. The critical section is often relatively short. Sooner or later towards the end of the line, it is common that the number of alighting passengers starts to exceed the number of boarding passengers, and the occupancy rate is falling. An example of a typical profile of the bus occupancy along a diametrical, urban bus line in the morning peak is given in fig 6 below.

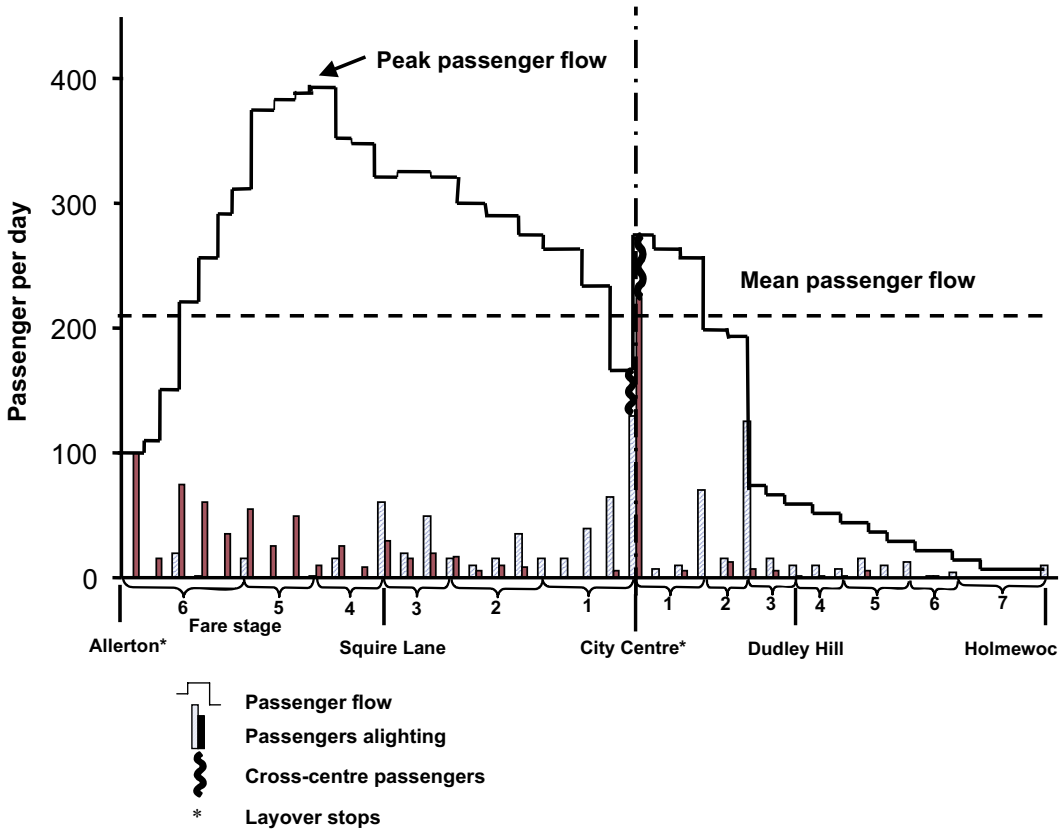


Figure 6 **The number of boarding and alighting passengers, and the resultant number of passengers on the bus of a cross-centre service during the morning peak**

When the bus turns around in the centre of town, or continues in the same direction on a diametrical course towards a neighbourhood at the opposite side of town – it makes little difference so far as the pattern of the rates of boarding and alighting is concerned – it will have a lot of excess capacity all the way to the terminus. Consequently, even in the peak period the average occupancy rate from start to end on a diametrical route, or on a complete round in the case of a radial route, is seldon much higher than about one half.

To raise this figure by spatial peak-load pricing would be fine, if a corresponding price differentiation in respect of where along the line, and in which direction trips are made, could be practicable. One must not expect a very large effect, at least not in the short run, because the work-place concentration to the central city would probably not change. Neither would a great many people change the location of their homes as a result of spatial peak-load pricing of urban SPT-services, at least not in the short run. The long-run effects are, however, strictly unknown.

In the following two sections two peaking problems are taken up for which demand management by peak-load pricing would be both practicable and beneficial: the daily peaks in urban commuter transport, and the weekly peaks in interurban train transport.

5. Solution to daily peaking problem – peak-load pricing of commuter transport by bus

In urban transport between the suburbs and the central city, as well as between different suburbs, journeys to/from work is the most important travel purpose. Therefore the morning and afternoon peaks stand out in the time-profile of demand. In addition, the spatial peaks of each round voyage is just as marked. It is not only on the backhaul that the occupancy rate is low, but also on the main haul the bus is fully occupied only in the "critical section", which may constitute just a fraction of the whole route as indicated by figure 6 above.

In OECD 1985 a model was designed by the present author with a view to determining the optimal differentiation of peak and off-peak fares in urban commuter transport by bus. In that model the question of optimal bus size was kept out of consideration. The purpose was to pinpoint the basic elements of peak-load pricing. Since it can be assumed that all producer costs are linear, the pricing-relevant marginal costs can be calculated without specifying the peak and off-peak demand functions. The marginal conditions for social surplus maximization are sufficient to obtain the optimal price structure. A summary of the model results is given below.

Consider a bus route between a suburb and the central city served by N buses in the peak period. If these buses were in operation all day, *Case a*, the total cost of the bus company can be roughly expressed like this:

$$TC_a^{prod} = \beta N \qquad (6)$$

In the case where off-peak capacity is less than peak capacity, *Case b*, two categories of buses are used - "peak-only buses" and "all-day buses". As the names suggest, a peak-only bus is in operation only in the morning and afternoon peaks manned by a driver on a "split shift", or by two half-day working drivers. An all-day bus is in operation during two straight shifts. The total producer cost is in this case written:

$$TC_b^{prod} = \beta N_{ad} + \beta_1 N_{po} \quad (7)$$

where

N_{ad} = number of all-day buses
N_{po} = number of peak-only buses
$N = N_{ad} + N_{po}$ = total peak vehicle requirement)

The total time costs of the users of the bus service are also a function of ***N*** in the first place. The more buses there are, the higher the frequency of service with consequent reduction in waiting times. In *Case a* where the number of buses in peak and off-peak periods is the same, we can write the total user costs like this, denoting the total number of bus trips per day by ***B***:

$$TC_a^{user} = f(N)B \quad (8)$$

In *Case b* where the number of buses are different in peak and off-peak, the total user cost comes to:

$$TC_b^{user} = f(N) \cdot B_{peak} + f(N_{ad}) \cdot B_{off\text{-}peak} \quad (9)$$

We now have a complete, very simple expression for the total social cost of bus services per (work)day:

$$TC_a = \beta N + f(N)B \quad (10)$$

$$TC_b = \beta N_{ad} + \beta_1 N_{po} + f(N) B_{peak} + f(N_{ad}) B_{off\text{-}peak} \quad (11)$$

5.1 The price-relevant marginal cost of peak trips

The price-relevant cost of bus traffic should be calculated per bus in the first step. In the second step we arrive at a cost per bus trip simply by dividing the pricing-relevant cost per bus by the number of trips made on a marginal bus.

For *Case a* the cost and benefit (= negative cost) of another bus in the system is easily obtained as:

$$\frac{dTC_a}{dN} = \beta + B\frac{\delta f}{\delta N} \quad (12)$$

In *Case b* we have, in principle, two costs of additional buses, depending on whether peak only or all-day bus is added, although only the former alternative would be relevant when it comes to peak-price calculations.

The incremental cost difference between adding an all-day bus and withdrawing a peak-only bus has some interest in that it represents the marginal off-peak capacity cost in *Case b*. If one

wants to increase the number of buses in off-peak, keeping the peak capacity constant, this is the way to go about it: add an all-day bus and take a peak-only bus out of operation, or, which happens in practice, of course, put a peak-only bus into all-day service.

$$\frac{dTC_b}{dN_{po}} = \beta_1 + B_{peak}\frac{\delta f}{\delta N} \tag{13}$$

$$\frac{dTC_b}{dN_{ad}} = \beta + B_{peak}\frac{\delta f}{\delta N} + B_{off\text{-}peak}\frac{\delta f}{\delta N_{ad}} \tag{14}$$

$$\frac{dTC}{dN_{ad}} - \frac{dTC}{dN_{po}} = \beta - \beta_1 + B_{off\text{-}peak}\frac{\delta f}{\delta N_{ad}} - \tag{15}$$

The next question is: by what should the incremental cost of an additional bus be divided to get the pricing-relevant cost per trip? The first thought, that the incremental cost should be shared by all passengers using the additional bus while it is in operation, is wrong. Only those passengers that are on the bus in the "critical section" of the route concerned have "cost responsibility" for an additional bus. This may be only something like half the total number of passengers travelling by the bus; for example, all passenger trips made on the back-haul put hardly any demand on capacity. We assume that a given proportion ***(α)*** of the total peak trips, $\boldsymbol{B_{peak}}$, are capacity-demanding in the sense that they occupy seats and standing space on buses when the buses traverse the sections of each individual route which constitute "spatial peaks".

Dividing the incremental cost of another bus by the number of capacity-demanding peak trips per bus employed in the peak periods, $\boldsymbol{\alpha\, B_{peak}/N}$, we get the pricing-relevant cost, $\boldsymbol{MC_{peak}}$ in two versions:

$$MC^a_{peak} = \frac{\beta N}{\alpha\, B_{peak}} + \frac{Bf(N)}{\alpha\, B_{peak}} \cdot E_{fN} \tag{16a}$$

$$MC^b_{peak} = \frac{\beta N}{\alpha\, B_{peak}} + \frac{1}{\alpha}\, f(N) \cdot E_{fN} \tag{16b}$$

where $E_{fN} = \frac{\delta f}{\delta N} \cdot \frac{N}{f(N)}$

5.2 The price-relevant marginal cost of off-peak trips

The *normal off-peak case* should be that buses practically never run fully occupied, and it is irrelevant to pursue the preceding "average cost of the marginal bus" argument. Increased off-peak patronage should not require additional buses.

The acts of boarding and alighting of additional passengers will reduce overall bus speed. If bus travel were free in off-peak, this pricing-relevant cost would be almost negligible. On the other hand, if a ticket were to be bought from the bus driver, the pricing-relevant cost would be doubled or trebled. However, it is nonsensical to charge a price with the main rationale that the very collection of the price, and nothing else, causes the pricing-relevant cost. Perhaps the

best compromise is to introduce extremely cheap monthly or yearly passes for off-peak travel. Passengers with passes cause hardly any additional cost (over and above the cost of a free rider) and a yearly pass at the cost of, say 30 Euro would not (as it should not) discourage any person in need of bus transport in off-peak periods, and it could have a desirable, preventive effect on children's or others' riding for fun or mischief.

The case of off-peak capacity being scarce: Under certain, not very likely circumstances, it can be right to reduce off-peak capacity so much that the capacity constraint becomes binding in the critical sections also in off-peak. Or in other words, the number of all-day buses, N_{ad} is made just sufficient to meet the off-peak demand in the spatial peaks on individual routes.

In this case, *Case c*, the pricing-relevant cost of off-peak trips in the critical section becomes:

$$PC_{off\text{-}peak} = \frac{\left(\frac{dTC}{dN_{ad}} - \frac{dTC}{dN_{po}}\right) N_{ad}}{\alpha B_{off\text{-}peak}} = \frac{(\beta - \beta_1) N_{ad}}{\alpha B_{off\text{-}peak}} + \frac{1}{\alpha} \cdot \frac{\delta f}{\delta N_{ad}} N_{ad} \qquad (17)$$

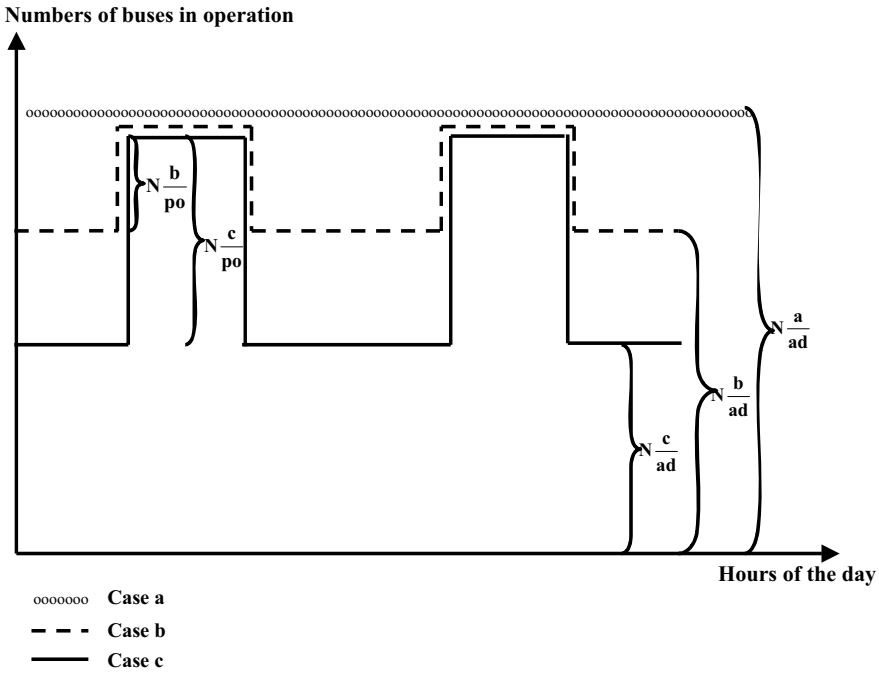


Figure 7 **The bus inputs during the day in three cases**

It is interesting to note that it matters very little for the pricing-relevant cost whether or not off-peak capacity can be assumed to be a binding constraint. The pricing-relevant cost is very low all the same. This is a reflection of the fact that the benefit to all off-peak passengers of increasing the frequency of service is so relatively great (when the frequency is at a low level initially) that the incremental cost of another peak-only bus extending its operation to all-day service is almost offset by cost savings for the original off-peak passengers.

5.3 Numerical example of the optimal structure of bus fares

On the assumption that the average waiting time at stops is equal to half the headway time, and with the parameter values used in the previous Circletown model, the level and structure of optimal fares can be calculated: the results are summarized in table 3.

As seen, peak fares in the critical section should be more than twice the producer average cost, but still at a moderate level in absolute terms. Fares for trips outside the critical section, and off-peak fares generally, should be very low, which indicates that a substantial subsidy is required. It can be observed that it turns out that for peak trips the level of the pricing-relevant cost is substantially higher in *Case a* than in *Case b* in the numerical example. This is not a general characteristic of running only all-day buses versus differentiating the peak and off-peak frequency of service, but a result of the fact that the bus size is held cosntant in all calculations. However, in the given circumstances of the model example the same bus size cannot be optimal in both alternatives.

Table 3 **Examples of price-relevant marginal costs and bus company average costs per trip in peak and off-peak, excluding night, Saturday and Sunday services**

Price-relevant costs in peak periods	**Euro per trip**
Trips in critical section, *Case a*	3.30
Trips in critical section, *Case b* and *c*	2.40
Other trips, *Case a, b,* and *c*	0.20
Price-relevant costs in off-peak periods	**Euro per trip**
Trips in critical section, *Case a* and *b*	0.08
Trips in critical section, *Case c*	0.10
Other trips *Case a, b,* and *c*	0.08
The bus company cost per trip, AC^{prod}	**Euro per trip**
All trips, *Case a*	1.12
All trips, *Case b*	1.04
All trips, *Case c*	0.94

6. Solution to weekly peaking problem – peak-load pricing of interurban train services

The peak-load pricing problem is now weekly rather than daily, when it comes to interurban transport. The peaks and troughs in the daily demand time-profile can, at least partly, be met by varying the length of trains at different departure times without increasing the slack. During off-peak days, on the other hand, the excess capacity of the rolling stock, which will arise if uniform pricing over time is applied, cannot be gainfully used elsewhere, because the weekly time-profile is by and large the same for all lines. In Jansson, et.al 1992 a model of railway transport was prepared for deriving the price-relevant marginal cost of passenger train services, which has been further developed in the ongoing UNITE project by a case study of a particular line. The main results of this work are summarized below.

In long-distance public transport, where time-tables are used by prospective riders, the "Mohring effect" is more difficult to estimate compared to the case of urban travel. Therefore, the right approach for long-distance train transport is to calculate the price-relevant marginal cost by assuming that additional passenger demand is met by vehicle size increases.

6.1 The price-relevant cost of passenger train services

This idea is relatively simple to apply in the case of flexible-formation train transport, where train size (length) is adjustable; carriages can be added to or uncoupled from the train in a marshalling yard during night. Boarding/alighting charges is moreover a very mild complication because unlike in bus transport the number of inlets and outlets is increasing proportionally to vehicle size, and tickets are bought in advance, which means that ticket transaction time is no part of the transport vehicle time.

The only more demanding bit in the calculation of the incremental cost of adding another carriage to a train is to find out how energy cost develops as a train is made successively longer. In a joint study with SJ it was found that, given train speed, energy consumption will increase linearly with train length in the whole range of observations. Then the price-relevant cost can be formulated in a very simple way. The least unit of supply is another carriage carried from the point of departure, say the central station of Stockholm, to the final destination, for example Malmö, and back again. The incremental cost of producing this additional capacity constitutes the numerator of the pricing-relevant cost, and the number of additional passengers thus accommodated constitutes the denominator:

$$MC_{ti}^{train} = \frac{\mu_{ti} + cD}{n} \qquad (18)$$

MC_{ti}^{train} = pricing-relevant marginal cost per occupied seat day t train departure i

(t = 1.....365, and i = 1.....m)

μ_{ti} = opportunity cost day t train departure i of the marginal carriage

c = additional running cost of a train per kilometer caused by coupling up another carriage

D = round voyage distance

n = target number of occupied seats per carriage

This formulation of the price-relevant marginal cost presupposes that the train on the route concerned only makes one round voyage per day. On a shorter route it may be possible to carry out one and a half, or more two rounds, which would mean that the denominator is to be increased by a factor of 1.5, 2, etc.

Note that the price-relevant cost is given per occupied seat of a round voyage. This cost should be shared out among all passengers successively occupying a particular seat during a round voyage. The number of passengers per seat and round voyage could be two, one in each direction, or more than two, since many passengers make shorter trips than the whole distance from start to end.

An efficiency condition is that summed over all departures all days of a year the opportunity cost of a carriage should equal the annual capital cost.

The financial result of optimal pricing of passenger train services is easily imagined. The revenue will cover the capital and operating costs of carriages including guards' wage costs, but no contribution will be made towards covering the costs of engines including engine-drivers' wage costs, nor to the major part of overhead costs which are independent of train length. Only about half the total costs of passenger services will be covered by optimal train fares.

6.2 The optimal structure of train fares

An additional efficiency conditions, which is useful in the derivation of the peak-load pricing structure, can be written like this:

$$\mu_{t1} = \mu_{t2} = ... \mu_{ti} ..... = \mu_{tm} = \mu_t \qquad (19)$$

The rolling stock of a particular line can be assumed as given one particular day. The number of engines and carriages can only be changed from one day to another. An efficiency condition is then that each day the given number of carriages should be distributed between the m trains such that the capacity utilization is nearly constant. This means in turn that the opportunity cost of a carriage is the same every departure a particular day, as shown in (18) above.

The stochastic element in railway travel demand is substantial, so a very high occupancy rate should not be aimed at. At present for SJ the mean occupancy rate of SJs trains is about 1/3, but is systematically rather different in different sub-markets. By eliminating the systematic differences by means of peak-load pricing, aiming basically at equalization of the train occupancy rate in time and space, it should be possible to raise the mean occupancy rate to at least 1/2, which would be a very considerable improvement.

The first demand equalization to aim at should be to make the Monday-Thursday and Saturday (off-peak) level of demand nearly equal to the Friday and Sunday (peak) level. A

representative example of the time profile of train travel demand by day of the week in Sweden is given in figure 18.

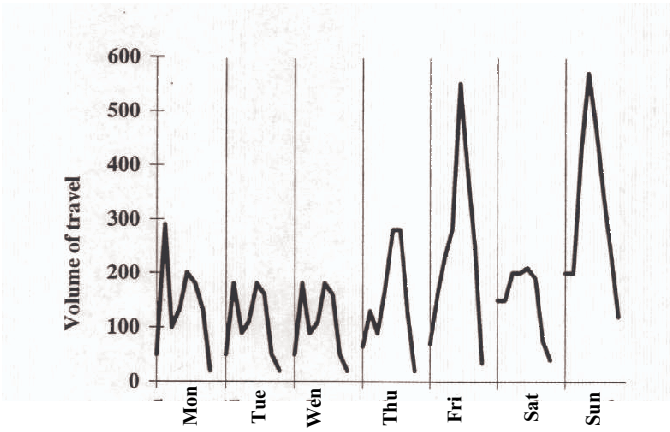


Figure 7 **Rail travel between Stockholm and Gävle on the different days of the week**

It was found in Jansson et.al 1992 that $\mu_i = 0$ in off-peak, i.e. if fares in off-peak were based on just the running cost component cD of the pricing-relevant cost (see equation (4) above), the level of demand would just fall short of the peak level as it would be when peak traffic alone pays the carriage capital costs.

The second demand equalization to aim at should be to level out spatial peaks and troughs. We have not gone into this matter very deeply. A lot remains to be done. Take just as a rather typical example the line between Stockholm and Sundsvall in the north of Sweden. Dividing it into three sections, the daily passenger flow in peak and off-peak, respectively, in the three sections was as follows:

Table 4 **Passenger flow per day in three sections in peak and off-peak relative to the critical section of the Stockholm-Sundsvall line**

Line segment	**Fri, Sun**	**Mon-Thur, Sat**
Stockholm - Gävle	100	53
Gävle – Söderhamn	75	37
Söderhamn - Sundsvall	45	22

The peak/off-peak price differentiation advocated above would equalize the flow figures in each particular row in table 4. To make spatial supply and demand match better, the first step is to make some trains from Stockholm turn around already in Gävle, and some in Söderhamn. The spatial demand equalization would then imply a price differentiation to the end of making the occupancy rate equal in each section of the line.

6.3 Numerical example of three different lines

In the numerical example below of optimal fares according to the principles of peak-load pricing, the focus is on the fare differentiation by day of the week. Fares examples are given for a rather short, a medium-distance, and a fairly long-distance line. Never mind the absolute values of the figures; they are in Swedish currency in the year 1990. It is the structure of fares, which is interesting. As seen in table 5, both with and without a budget constraint, off-peak fares should be only about one third of the peak fares. Since the elasticity of demand for rail travel differs somewhat between routes with and without airline competition, both cases are considered in the illustration.

Table 5 **Optimal rail fares for different days of the week on three different lines in 1990, SEK per second class single trip from start to end**

Line distance	**Day of the week**	**Optimal fares**			
		without budget constraint		**with budget constraint**	
		Air comp	**No air comp**	**Air comp**	**No air comp**
170 km	Fri, Sun	113	113	154	154
	Mon-Thu, Sat	30	30	56	57
335 km	Fri, Sun	187	202	240	303
	Mon-Thu, Sat	50	58	87	123
550 km	Fri, Sun	254	294	292	462
	Mon-Thu, Sat	72	96	104	204

As seen, the route distance makes little difference so far as the peak/off-peak differentiation is concerned. Naturally where air transport is an alternative, fares are more markedly tapering off with respect to distance.

The low off-peak fares in the optimal tariff would apply to 70 % of total travel. This means that the weighted average fare level is substantially lower than the level of SJ's fares at that time. In case the financial result is constrained to be the same as that of SJ, we found that also the level of second-best (Ramsey) fares is lower than SJ’s fares, which were not differentiated between peak and off-peak days.
We made a rough calculation of the likely travel volume increase as a result of changing over to the optimal fares structure. In the unconstrained case the travel volume would double. Most of the increase would, of course, occur in the off-peak period. In the constrained case the price level has to be substantially higher, and as a consequence, the increase in the total volume of

travel is down to 40%. It is interesting to note that the net welfare gain in the latter case is as high as 75% of the net welfare gain of peak-load pricing in the case where no budget constraint is assumed.

7. Refutation of the main objections to public transport subsidization I: "The cost of public funds"

In urban bus transport, in particular, it seems that less than half the revenue should come from fares in order to meet the marginal conditions for social surplus maximization. And in addition, as long as urban road pricing is missing (apart from general fuel taxation), second-best optimal prices of urban public transport should be still lower (Larsen, 1997).

The latter reason for urban public transport subsidization seems together with welfare distributional considerations to be the main, or only, rationale in the minds of politicians and decision-makers for the present large subsidization of urban public transport (which in many towns and cities happens to be of the right order of magnitude from a social-economic point of view). If proper road pricing would at last be introduced, it is consequently likely that many influential persons would argue for abandoning the subsidization of public transport.

A more sofisticated argument to the same effect is that the so called "cost of public funds" may justify the discontinuation of public transport subsidization in case road pricing is introduced. This issue is the main topic of the following discussion. There is also another main objection to public transport subsidization: "Cost efficiency", i.e. carrying out a given task at the least cost, may suffer in a subsidization regime, which could easily outweigh the allocative gain obtained by optimal pricing. I take the stand that both cost efficiency, or "X-efficiency" (Leibenstein, 1966) and allocative efficiency are to be aimed at, and the crux of the matter is to design the institutional framework such that the potential goal conflict is eliminated. This is a new research area, which is just touched upon in this paper.

7.1 The excess burden of different taxes including prices exceeding the marginal costs

The ideal state of affairs is obtained when the prices are equal to the price-relevant marginal costs everywhere in the economy, and public undertakings, including subsidies to decreasing-cost industries, are financed by the surplus from increasing-cost industries, the revenue from externality charges (unless it is earmarked to compensating the sufferers), and individual poll taxes. By differentiating the poll taxes in accordance with the widely different ability to pay of different individuals, e.g. in such a way that everybody pays the same total tax next year as he/she actually did last year, but now (in the first-best situation) in the form of a lump-sum tax which is independent of next year's income, ambitious distributional goals can also be obtained.

The point of lump-sum taxes is that there is no excess burden involved. So in reality, when funds are raised by income and/or commodity taxation there is a "cost of public funds" (CPF) involved. This is a seemingly convenient concept which has recently been introduced in practical cost-benefit analysis. Empirically it is measured as the weighted average of the excess burden, or deadweight loss per crown of tax revenue raised by each particular tax. An efficiency condition, taking only allocative efficiency into account, is that the marginal cost of public funds is equal for each tax. A brief survey of the causes of the excess burden of four different types of taxes follows below. A literature survey of the empirical work with a view to estimating the marginal cost of public funds is given in Brendemoen 1999.

(1) A **specific commodity tax** upsetting the optimality conditions of a first-best economy

imposes an "excess burden" on the economy (over and above the burden on the tax-payers), which can be measured in a diagrammatic, partial analysis of the commodity market concerned by the triangle representing the difference between the consumers' surplus lost and the tax revenue obtained. It should be observed that a price set above the price-relevant cost, for example for public transport services, imposes a corresponding excess burden on the economy.

The measurement of the excess burden should be based on the *compensated* demand curve, i.e. the demand as it would appear in case consumers were compensated by income rises, as the commodity tax is successively raised, to keep them at the same level of utility. When it comes to a specific tax imposed on just one commodity, constituting a small fraction of total consumption, the income effect is small enough to be ignored for practical purposes (Willig 1976), and the more easily observable uncompensated demand curve could be used as a good approximation when calculating the excess burden of a specific commodity tax.

The table below illustrates the relative order of magnitude of the excess burden of a specific commodity tax under the simplifying assumptions of a constant marginal cost of production, and a linear demand function in the relevant range. In this case the excess burden relative to the tax revenue obtained is determined by just the proportional difference between the resulting output volume (after the tax is imposed) and the first-best output volume, irrespective of the elasticity of demand. If the ratio of second-best to first-best output is X, the excess burden (EB) relative to the tax revenue (TR) raised is equal to X/2(1-X). This ratio defines the "average excess burden", and dEB/dTR the "marginal excess burden". When TR reaches its maximum, the marginal excess burden goes to infinity, while the average excess burden in the illustrative example of table 1 is just one half, i.e. half a crown per crown of tax revenue.

Table 6: **The average and marginal excess burden of a specific commodity tax**

Percentage reduction of first-best output volume	Average excess burden, EB/TR	Marginal excess burden, dEB/dTR
10%	0.06	0.13
20%	0.13	0.33
30%	0.21	0.75
40%	0.33	2.00
50%	0.50	∞
.		
.		
100%	∞	-1

A given percentage reduction of the first-best output volume can be the result of rather different tax levels when the elasticity of demand takes different values. The stronger substitutes of the taxed commodity there are, and, consequently, the more elastic the demand is, the greater the effect of a given tax on the quantity demanded will be. The root cause of the excess burden of a specific tax on, for example, oranges is thus that the choice of fruit is distorted.

(2) A **general commodity tax**, e.g. a uniform value-added tax (*vat*), imposes an excess burden on the economy which has a different primary cause. The resource allocation between commodities is not the main problem since all prices are raised equi-proportionelly, but the choice between work and leisure.

In the hypothetical case where the total *vat* revenue is paid back to the households as given lump-sums, a reduction in the quantity demanded of all different commodities will occur only if the total amount of work is reduced. On a free labour market this is likely to happen. The real wage is reduced to the tune of the increase in the price level due to the *vat*, and as the income effect on the choice between work and leisure is eliminated by a hypothetical refund, only the substitution effect on labour supply is operative. The substitution effect is always positive – lower real wages lead to a reduction of the supply of labour when looking at the compensated labour supply curve. Therefore the main excess burden of a uniform *vat* is to be found in the labour market. We shall come to this later when the excess burden of income taxation is discussed at the end of this section. Let us first deal with "Ramsey taxes", since the contribution of Ramsey (1927) has played a main role in the public economics literature since it was "rediscovered" by Baumol and Bradford (1970).

(3) By **differentiated commodity taxation** it is possible to do better than by a flat *vat* rate, it can be argued. The first-best position cannot be reached, because by taxing consumption, however sophisticated it is done, the optimal balance between work time and leisure time is upset. The only "solution" to this problem would be to tax leisure time by the same rate as work time, and in this imaginary case no commodity tax differentiation is called for. Such taxation would be equivalent to a poll tax, because when all 24 hours of a day are taxed by the same rate per hour, the total tax is obviously a given lump-sum.

Ramsey (1927) considered the problem of second-best commodity taxation, on the condition that leisure time could not be taxed, and assuming homogeneous consumers. (The term "second-best" was, to be sure, not yet invented at that time). Ramsey found that the tax rates of different commodities should be inversely proportional to the absolute value of the demand elasticites, or with an alternative formulation that the resulting percentage deviation between the first-best and the actual output should be the same for all commodities.

To bring proportional demand quantity curtailment about by commodity taxation, tax rates which are disproportional to the marginal cost of production are required, because the demands for commodities which are complements to leisure decrease less than the demands for commodities which are substitutes to leisure. Hence the Ramsey Rule implies that complements to leisure should be relatively heavily, and substitutes relatively mildly taxed. When the assumption of homogeneous consumers are relaxed, everything gets much more complicated, which will be discussed in a following section. At this stage it can be concluded that the easing-off of the excess burden of a uniform *vat* that could be obtained by Ramsey taxes, given the total tax revenue required, is likely to be relatively insignificant in a hypothetical economy of homogeneous consumers (compare Stiglitz, 1988). A greater problem is probably the excess burden in the labour market, which is not directly addressed by the original theory of optimal commodity taxation.

(4) A proportional **income tax** has an equivalent effect on the choice between work and leisure as a uniform *vat*. Income taxation is by far the most important source of income for central, and, in particular, local governments (in Sweden among other countries), so in reality the main excess burden of taxation is for double reasons connected to the labour supply function. This is also well certified by the empirical studies of the "cost of public funds". Let us look a little closer at this function:

$$S_L = f\left(\frac{W(1-t)}{MC(1+c)},\quad LST,\quad GC_{com},\ldots\right) \qquad (20)$$

S_L = supply of labour

$w = \frac{W(1-t)}{P}$ = real wage rate

W = wage cost for employers

t = all-inclusive tax rate on labour

P = MC (1+c) = price level, where
MC = level of marginal cost of production
c = average commodity tax rate

LST = lump-sum tax (including negative taxes like state pensions)

GC_{com} = generalized cost of home-work commuting

In the public economics literature it is common – and theoretically adequate in my view – to call every excess of the final price over the marginal cost a ”tax”, no matter whether it is a governmental tax, or just the result of profit-maximizing under imperfect competition, or full-cost price-making. In expression (20) above **c** stand for the average tax rate of commodities in this wider sense.

The explanatory variables included in (20) are not the only arguments, but supposedly the most important ones of the labour supply function. The first-mentioned argument, the real wage rate has both a substitution and an income effect on the choice between work and leisure. The excess burden should be calculated with reference to the *compensated* labour supply function. Although the “ordinary” uncompensated supply function could very well be completely inelastic, or even backward-bending, it is axiomatic that the compensated labour supply function is positively related to the real wage rate, and consequently that there is an excess burden of taxes on labour, irrespective of the actual shape of the uncompensated supply function. Stiglitz (1988) gives the following exceedingly simple formula for the excess burden (EB) relative to the tax revenue (TR) obtained by a proportional income tax rate, **t**:

$$\frac{EB}{TR} = \frac{t}{2}e \qquad (21)$$

In this formula **e** stands for the compensated labour supply elasticity. Unfortunately, **e** is very difficult to estimate. Different approaches to estimating the compensated labour supply elasticity have yielded a wide range of values – from negative values, which are inconsistent with basic theory, to values close to unity, which also are implausible. In the *Handbook of Labour Economics* edited by Ashenfelter and Layard, the results of a number of major

empirical studies are summarized, and it is concluded that “if labour economists had to vote on the best elasticity, the average might be 0.11” (Pencavel, 1986, chapter 1).

Using this figure and the fact that about two thirds of labour cost are taxes gives a relative excess burden of the Swedish tax system of approximately 0.025, which does not seem very high. The *marginal* cost of public funds is, however, higher – at least twice as high. However, the value of **e** is the critical parameter. A rather low value for men is plausible, but a considerably higher value seems to apply to the female labour supply.

7.2 Implications for cost-benefit analysis and pricing policy in transport

So far we have followed the traditional way of discussing the excess burden of taxation by keeping the question of what the taxes are to be used for out of the discussion. It is now time to bring this matter into the picture.

The government needs tax money for a number of reasons: three main purposes can be distinguished in this connection:

(1) To buy resources for the production of *public goods* (which are services in reality, i.e. "immaterial goods") as well as services, which do not have the character of ”public goods”, like medical care and schooling, but which are provided free, or highly subsidized.

(2) To transfer money in the form of various *allowances* to the needy.

(3) To pay *subsidies to decreasing-cost industries* to make marginal cost pricing feasible.

The third purpose is not very prominent in general discussions of public finance, but, of course, relevant for the question of optimal transport pricing.

The point is that the *net* excess burden imposed on the economy by tax-financing different public undertakings depends on the character of the undertaking in question. Bringing home this point, the three, aforementioned main purposes of taxation are taken up in turn:

7.2.1 The cost of public funds for the purpose of public production

A relatively clearcut case is the tax-financing of pure public goods like national defence, and the administration of justice. For these undertakings there are no offsetting effect on labour supply to set against the excess burden of the required taxation. Pure public goods are characterized by non-rivalry and non-excludability, which means that everybody are “free riders”. The appearance of more and/or better public goods on the market gives no extra incentive to work, because the goods are available for free.

What about investments in transport infrastructure? Cost-benefit manuals issued by national road and rail administrations prescribe that “the marginal cost of public funds” should be taken into account so far as tax-financed road and railway investments are concerned. In Sweden the current figure is 0.3 crowns per tax-crown. There are no tollroads in Sweden. It is sometimes argued, however, that if some new roads were instead financed by tolls, the investment costs should not be inflated by the factor 1.3 Is this very logical? When a free road becomes a toll-road the real wage rate expression in the labour supply function (1) above is affected downwards by a denominator increase. In the absence of money illusions this would

create an excess burden comparable to a rise of the income tax rate, **t** in the numerator. In the next step, however, it should be observed that road users will normally get something in return for the tolls paid; they save time, and the accident risk may be reduced. The price level index **P** should not go up, if an increase in a constituent price is set off by a quality rise, which the consumers value at least as much.

A variation of the latter argument can be applied to the main case of free-road investments: the quality of the road services goes up (the generalized cost goes down) at least to the tune of the tax money requirement for bringing it about. The fuel tax can be regarded as the price of the road services of a "free-road" network. As long as the nominal value of the fuel tax stays the same, and the quality of road services is improved, the real price of road services is falling. This should be taken into account when considering the excess burden of the tax money used for state grants to investments in transport infrastructure. The practice of inflating the investment costs by a factor reflecting the "marginal cost of public funds" seems doubtful in this case.

7.2.2 The cost of public funds for the purpose of transfer payments to households

Income transfer from the more wealthy to the less wealthy affects two main arguments in the labour supply function (20) – the real wage rate, **w**, and **LST**.

The former effect gives rise to an excess burden because the real wage rate is decreased. The negative influence on labour supply by the substitution effect of the real wage decrease is reinforced by the latter effect, i.e. the income effect of the lump-sum payments to the needy, which are financed by tax rises.

7.2.3 The cost of public funds for the purpose of subsidizing decreasing-cost industries

Taxation with a view to financing subsidies to decreasing-cost industries has also two effects on the labour supply function: if the income tax rate, **t** is increased, the numerator of the real wage expression in (20) goes down, but so does the denominator of (20) when the price of decreasing-cost industries is lowered down to the marginal cost. As a first approximation it can be assumed that such a restructuring of taxes will have no influence at all on labour supply. If some taxes are raised to finance cuts in other taxes, the real after-tax wage rate will stay the same, and no excess burden will arise. On the other hand, in the case of public transport (as well as road services), it is seen that a possible positive effect on the labour supply may come up from the third argument in (20), $\mathbf{GC_{com}}$. There is a lot of discussion of the secondary effects on employment, and economic growth in general, of reductions in the generalized cost of transport, which we shall not go into here. This would lead too far away from the main line of the present argument. Let it only be said that the possibility that a reduction of $\mathbf{GC_{com}}$ can increase employment reinforces the main conclusion that the taxation necessary to finance a marginal cost pricing policy of local public transport does not result in an excess burden on the economy.

8. Refutation of the main objections to public transport subsidization II: towards achievement of both allocative- and X-efficiency in public transport

The second part of the heading above is the title of a forthcoming article by Johansen et.al (2001), where the basic idea is to design the subsidization system for public transport companies such that profit maximization on the part of the SPT-operator will coincide with social surplus maximization. "Performance contracts" towards this end have in fact already been introduced in some cases in Norway. It would lead too far here, to go into this intriguing issue in depth. Let us only call attention to a promising area for new research, and by some further comments clarify what the basic idea boils down to.

Profit maximization is "producer surplus" (PS) maximization. What is most desirable from a social point of view is that the "social surplus" (SS) is maximized. The difference between these two surpluses is, of course, the consumers' surplus (CS).

$$SS = PS + CS$$

Transport-system-externalities like exhaust fumes from buses can be assumed to be internalized by appropriate externality charges, and need not be further elaborated. However, the basic idea of solving the apparent conflict between the operator's desire to maximize his profit, and social surplus maximization, is related to externality internalization: every change in CS that is a result of whatever step is taken by the operator should ideally be internalised in his profit and loss account. Given that a public transport system is a pronounced decreasing-cost activity, social surplus maximization would create a large financial loss, and is therefore out of the question for the operator. On the other hand, in case every positive change in CS would increase his revenue and every negative change in CS would reduce his revenue correspondingly, the operator would have the necessary incentive to act as a SS-maximizer.

How should this incentive be provided in practice? The most interesting idea is that the price and/or supply regulations can be abolished, which typically are parts of the deal between the principal and the operator, if the subsidization "formula" could take the form of CS internalization. Then the operator has the freedom and right incentive to take steps in pricing policy, service design, manning, etc. which would enhance both cost-efficiency, and allocative efficiency.

The challenge is, obviously, to find "the formula", which would work in practice, that is to say, is sufficiently reliable, as well as intelligible for both regulator and operator.

REFERENCES

Baumol, W and Bradford, D (1970). ”Optimal departure from marginal cost pricing”. *American Economie Review,* June.

Brendemoen, A (1999). “Marginal Cost of Public Funds”. Rapport till LOKTRA. ECON, Oslo.

Goodwin, P (1992), “A review of new demand elasticities with special reference to short and long run effects of price changes”. *Journal of Transport Economics and Policy.*

Jansson, J O (1984), “Transport system optimization and pricing”. Wiley

Jansson J O (1997), ”Theory and practice of transport infrastructure and public transport pricing” in *Recent developments in Transport Economics* (ed. by Nash and de Rus). Aldershot. Gower.

Jansson, J O, Andersson, P, Cardebring, P, and Sonesson, T (1992), “Prissättning och finansiering av järnvägens transporttjänster”. TFB& VTI forskning/reserach, Stockholm

Jansson, J O and Lindberg, G (1998), “Pricing principles” in *Pricing European transport Systems* (PETS), Deliverable D2 to DGVII of the European Commission

Johansen KW, Larsen O and Norheim B (2001), “Towards achievement of both allocative and X-efficiency in public transport”. *Journal of Transport Economics and Policy*. Sept.

*Larsen, O (1997). ”*Kostnadseffektiv rushtrafik”. *TØI rapport 346/1997*. Transportøkonomisk Institutt. Oslo.

Leibenstein, H (1966), “Allocative efficiency and X-efficiency”. *American Economic Review.*

Mohring, H (1972), “Optimization and scale economies in urban bus transportation”. *American Economic Review.* September.

OECD (1985). “*Coordinated urban transport pricing”*. Road Transport Research. Paris.

Oum, T.H, Waters II W.G and Yong, J.S (1992), “Concepts of price elasticities of transport demand and recent empirical estimates”. *Journal of Transport Economics and Policy*.

Pencavel, J (1986). ”Labour supply of men” in *Handbook of Labour Economics*. Vol 1, ed. by Ashenfelter and Layard. Amsterdam, North Holland.

Ramsey, F (1927). ”A contribution to the theory of taxation”. *Economic Journal*, March.

Regionplane- och trafikkontoret (2000), “Trafiken i Regionplan 2000”, Samrådsunderlag. Stockholms Läns Landsting.

Stiglitz, J (1988). “*Economics of the public sector”* 2nd ed. New York. Norton.

Transek (2000), “Förbifart Stockholm, Trafikanalys och samhällsekonomisk kalkyl,” i Vägverket Region Stockholm. RAP 2000:0434. Augusti