OPTIMAL CONGESTION TOLLS
FOR CAR COMMUTERS

A Note on Current Theory

By Jan Owen Jansson

THE PRODUCTION PERIOD FOR PEAK HOUR TRAFFIC

The theory of congestion tolls (or road pricing) is that the toll should equal the
difference between social marginal cost and private cost for the flow which will
prevail after the introduction of the toll.

The deduction of the cost function has normally followed the conventional lines of
general cost theory. This, however, does not hold good at one important point, viz.,
when it is assumed that the production function has to be defined for a given unit of
time.

Ordinary short-run cost theory indicates how costs vary when the quantity of
production per unit of time is changed. The time unit is a fraction of the given
production period. The normal – often implicit – assumption is that the production
period during a day covers either all 24 hours of the day or (perhaps more often) an
agreed working day of, say, 8 hours. Generally speaking, if production per day is to
be increased, the additional production factors must be put in during the stipulated
production period, e.g., the eight-hour working day.

It is possible, of course, that the production period per day may be extended, for
example by working overtime. In that case one can compare the cost of raising
production by lengthening the production period with the alternative cost of stepping
up the rate of production during the same period as before. In practice this might
mean comparing the cost of overtime working, on the one hand, with the cost of
operating above optimal capacity, on the other.

The production function becomes much more complicated when there is this
possibility of extending the production period by overtime. But the normal textbook
assumption is that the production period is fixed. This assumption is also implicit in
the theory of congestion tolls. There the costs are related to the traffic volume on a
given length of road per given unit of time.

This may be fairly realistic when considering urban traffic during working hours.
Most traffic at this time may be engaged in commercial and shopping journeys which
have to be completed during working hours. The ordinary cost theory is then
applicable. But the case of journeys to and from work is quite different; and these
journeys are the crucial ones for road pricing.

The period of production for journeys to work ends obviously when the latest
OPTIMAL CONGESTION TOLLS FOR CAR COMMUTERS

J. O. Jansson

jobs start. But – this is the crucial point – it has no fixed, institutionally given starting point. It starts when the earliest car commuters have to leave their homes; and this depends on their expected travel time, which is a function of the number of car commuters travelling in the morning.

Journeys from work in the evening are governed by the same principle, with the apparent difference that the starting point of the period of production is fixed, while the end point is flexible and depends on the time of travelling by car.

In general, the incremental cost of an increase in output will, in the short run, be less when it is possible to extend the period of production – a more efficient factor combination can be maintained – than when, as in the normal case, the additional output has to be produced within a fixed period of production.

In the theory of congestion tolls it is assumed that the period of production is fixed. If – as will be argued in this paper – this is not a realistic assumption for the production of trips between home and work, congestion tolls for car commuters calculated by the method now used will be too high.

ILLUSTRATION

Consider the morning peak in Stockholm. There is a great inflow of traffic, consisting mainly of car commuters, on the primary distributors from the suburbs to the central area of Stockholm. Those distributors contain the bottlenecks of the network. Traffic counts made at the border of the central area show that the morning peak occurs between 7.00 and 8.30 a.m., Before 7.00 the traffic is comparatively light. The average flow at the border during the peak is about three times as high as during the previous hour.

In order to calculate the optimal congestion toll on the primary distributors in the morning, a relationship between flow and travel time for these roads is deduced.

According to current theory the marginal congestion cost per mile, $X$, on which the toll is based, will be:

$$X = Q = \frac{\frac{dt}{dQ}}{t}$$

where $Q$ is the flow, that is, the number of vehicles crossing the border of the central area per minute during the morning peak, and $t$ is the journey time in minutes per mile.

This is, however, an inadequate simplification in a case of a flexible period of production, as will be clear from the following reasoning.

Suppose that the number of car commuters crossing the border of the central area during the morning peak increases by, say, 20 per cent. The distribution of starting times of the jobs of the additional commuters is assumed to be roughly similar to the original pattern of working times. The traffic counts at the border will reveal two important changes in the morning peak traffic:

1. The average flow per minute during the peak increases by less than 20 per cent.
2. The peak is widened, mainly by an extension of the earlier part of it. The nature of the change in the peak is shown in the diagram:
The reason for the extension of the peak is as follows:

The latest part of the peak is made up of those commuters who start work latest (say 8.30) and whose jobs are situated close to the border of the central area. The earliest part of the peak consists of commuters who begin their jobs early and who have the longest journeys within the central area border.

When traffic increases speed is reduced, so car commuters in general have to start their journeys earlier to be in time for work. The earliest group of commuters have to arrive earlier at the border than before, because they have a comparatively long journey in front of them at a lower average speed. The latest group of commuters, on the other hand, need not be so much earlier at the border, for the obvious reason that they have a shorter journey ahead of them.

The later part of the peak is therefore moved slightly back (i.e., to the left), while the earlier part of the peak is moved a more substantial distance to the left. Thus, a widening of the peak takes place.

MATHEMATICAL FORMULATION

A mathematical formulation of the argument is as follows. The symbols used are:

- \( K \) = the number of cars crossing the border of the central area during the morning peak.
- \( Q \) = average flow during the morning peak, i.e., number of cars crossing the border per minute.
OPTIMAL CONGESTION TOLLS FOR CAR COMMUTERS

J. O. Jansson

\[ T = \text{the length of the morning peak in minutes.} \]

\[ t = \text{journey time per mile on the primary distributors. } T \text{ and } t \text{ are functions of } K. \]

\[ t = t(K), \quad \frac{dt}{dK} > 0 \]

\[ T = T(K), \quad \frac{dT}{dK} > 0 \]

Average flow during the morning peak is:

\[ Q = \frac{K}{T(K)} \]

When the extension of the period of production shown by a widening of the peak at the border is taken into account the congestion toll per mile, \( X_1 \), will be:

\[ X_1 = K \cdot \frac{dt}{dK} \quad (1) \]

This is the time cost per mile which is imposed on the morning peak traffic by an additional commuter trip. It should be noted that the period of production is automatically extended when \( K \) increases.

According to current theory the congestion toll per mile, \( X_2 \) is a function of flow:

\[ X_2 = Q \cdot \frac{dt}{dQ} \]

To make a comparison of \( X_1 \) and \( X_2 \), it is necessary to show \( X_2 \) as a function of \( K \).

\[ X_2 = Q \cdot \frac{dt}{dQ} = \frac{K}{T} \cdot \frac{dt}{dK} \cdot \frac{1}{\frac{dT}{dK}} = \frac{K}{T} \cdot \frac{dt}{dK} \cdot \frac{T^2}{T-K \cdot \frac{dT}{dK}} \]

\[ = K \cdot \frac{dt}{dK} \cdot \frac{1}{T - \frac{K}{T} \cdot \frac{dT}{dK}} \quad (2) \]

The ratio \( X_2/X_1 \) is:

\[ \frac{X_2}{X_1} = \frac{1}{1 - E_{TK}} \quad \text{where } E_{TK} = \frac{K}{T} \cdot \frac{dT}{dK} \]

\( E_{TK} \) is the elasticity of the length of the peak, \( T \), to the number of car commuters, \( K \).

Normally \( E_{TK} \) is greater than zero and less than unity. It is then clear that \( X_2 \) is greater than \( X_1 \). When \( E_{TK} \) approaches unity, the flow on the primary distributors is getting close to capacity. When \( E_{TK} \) is unity, maximum capacity is reached. A further increase in the number of car commuters cannot then increase the flow (per minute); the whole effect of this increase of traffic will – in the most favourable case – be absorbed by an extension of the peak, while flow (per minute) is left unchanged.
THE BACKWARD CURVE

It is, of course, also conceivable that when demand grows flow is actually diminished. We are then on the notorious backward-sloping part of the time cost/flow curve. \( E_{TF} \) is greater than unity in that region. An increase of 1 per cent in the number of car commuters must lead to an extension of more than 1 per cent in the length of the peak, if the identity \( K = T \cdot Q \) is to be fulfilled when flow (per minute) is falling.

Let us look a bit closer at this “inefficient” part of the curve.

According to current theory situations beyond the point of critical density should unconditionally be avoided. It must obviously be inefficient, so it is argued, to produce a flow per unit of time which could be increased by reducing costs.

This is not so evident when applied to car commuting during a flexible period of production. When the critical density is exceeded it is true that an additional commuter trip in the morning may reduce flow per minute, but – this is the point – total output of the period of production will still increase, simply by an extension of the period of production. Marginal time cost is certainly high but still positive, because travel time, \( t \), is treated as a function not of flow, \( Q \), but of the number of work-trips completed in the morning, \( K \), during a period of time which is dependent on \( K \).

In a more detailed analysis it would be necessary to include further variables in the function \( t = t(K) \).

For example, the timing of the departures of car commuters in the morning might affect travel time, \( t \), quite substantially. Perfect timing might mean that total travel time and waiting time at the places of work are minimised for a given \( K \). This could be achieved – at least in theory – by imposing on car commuters some kind of “timetable” for departures in the morning and in the evening. But an adequately differentiated system of congestion tolls might also serve as a means of minimising total time cost for a given output, besides attaining the optimal level of output.

EFFECT ON EARLY MORNING TRAFFIC

The extension of the peak in the morning will not be completely costless. The traffic before the morning peak will be adversely affected. That cost has not been included in the schematic model above. If this could be done the picture would, of course, be more complete.

Some rough estimates indicate, however, that the addition of the cost imposed on the pre-peak traffic will make a quite insignificant contribution to the optimal congestion toll. There are two reasons for this forecast. First, the pre-peak traffic is comparatively light; and secondly, it is usually possible — there are no institutional obstacles — to carry out those car-miles slightly earlier in the morning if a widening of the peak makes that necessary. Very early in the morning the utilisation of the network is quite low.
WHAT DIFFERENCE DOES IT MAKE?

Is the difference between the values of the optimal congestion toll deduced in the current way and the way here advocated ($X_2$ and $X_1$) significant?

When flow is close to capacity the differences are very considerable, because $X_2$ is then approaching infinity, while $X_1$ is finite, even when maximum capacity is reached.

The optimal flow is, however, likely to be well below the maximum flow, and the idea of introducing road pricing is to attain that optimum. Therefore, the part of the flow curve close to maximum capacity is not of such vital importance in this context. The interesting part is the in-between range of levels of flow. It is here that the optimum level probably is to be found, whatever method of calculating marginal time cost is used.

The lower the volume/capacity ratio becomes, the closer to unity the ratio $X_2/X_1$ will be, because $E_{TK}$ is approaching zero.

For a given $V/C$ ratio, however, the value of $E_{TK}$ varies from one situation to another. One important relationship seems to be that the shorter an average commuter trip is in relation to the dispersal of starting times of jobs (or times when work is over), the lower $E_{TK}$ falls. If a representative commuter travels, say, only a few miles from home to work and the spread in time between the earliest starting and the latest starting jobs is considerable, the elasticity $E_{TK}$ is probably low and the difference between $X_2$ and $X_1$ is not likely to be very significant.

However, as a matter of principle, it seems desirable that in the theory of congestion tolls due account be taken of the fact that the period of production is flexible.

_The Economic Research Institute, Stockholm School of Economics_