Long-Run Effects of Partial Tolling Schemes

EDMOND L. DOUVILLE *

School of Business and Economics, Indiana University Northwest

Abstract

The objective of this paper is to make numerical estimates of the effects of various partial tolling schemes in a long-run setting where road capacities are harmonized with the tolling plan. The results demonstrate that different tolling schemes bring about different sets of equilibrium outcomes in respect to optimal investment in road capacity, travel speeds, numbers of road users, and the division of travel between peak and off-peak periods. These differential effects are relevant where, for political or equity reasons, simple maximization of efficiency may not be feasible or desired. The model presented provides a toll for evaluating these trade-offs.

1 Introduction

The objective of this paper is to make numerical estimates of the effects of various partial (less then first-best) tolling schemes. The perspective is the long-run urban commuting environment where road capacity may be optimized for the tolling scheme in question. The basic issue is how efficient are partial tolling schemes compared to the politically unpopular first-best solution. Also of interest are the impacts of the toll regimes on optimal road capacity, the levels of total road use, and the allocation of travel between peak and off-peak periods. The relative size of the tolls (or subsidies) and the operating characteristics of the roads (like average speeds) are also considered.

To address these issues, the paper uses a type of traffic congestion model first introduced in McDonald et.al. (1999). A road is idealized as a closed loop with entry and exit points uniformly distributed around the loop. At the beginning of each travel period, all vehicles using the road in that period simultaneously enter the road at points evenly distributed along the road. All road users travel the same distance at the same speed, simultaneously exit the road at the end of the travel period. The distance of the standard commuting trip is fixed, but the duration of the each travel period adjusts to accommodate the number of users, given the road capacity, an idea advocated by Else (1981).

The particular model used in this paper is a two-road, two-period model of commuting trips to or from work. The trips can occur in either of two discrete travel periods called the "peak" and "off-peak" periods respectively. The roads are imperfect substitutes for commuting. One innovative feature of the model is that travel in the peak period imposes a "delay" or "waiting" cost on the users in the off-peak period that is proportional to the

* Indiana University Northwest, 3400 Broadway, Gary, IN 46408. E-mail: edouvill@iun.edu

duration of the peak period. The off-peak users are forced into using the roadway at less desirable times in the early mornings and later at night, and their morning arrival or evening departure times are not optimal and involve waiting.

A second innovative feature relates to the treatment of cost and demand parameters. Benefits and costs are each formulated in terms of three parameters. The objective function is linear in these parameters so that only the relative magnitudes matter in the optimization process. Instead of relying on estimates or forecasts based on various historical datasets, "baseline model" is established for the case of no tolls on either road. In principle, envisioning the expected no toll equilibrium reveals the underlying cost and demand parameters. Here, one demand parameter is set arbitrarily, and the other two are chosen to create the desired or expected "own" price elasticity of demand and the "cross" elasticity of demand. Similarly, the cost parameter for the value of time used in commuting is set arbitrarily and the parameters for capacity costs and delay costs are chosen to obtain the expected or desired capacity utilization ratio and the allocation of road use between peak and off-peak time periods. Coordinating the cost and demand parameters is done iteratively. It turns out that sensitivity tests show the numerical results are quite robust to variations demand elasticities and vary predictably with relative cost variations. The mathematical formulation of the model is presented in the next section. The numerical results and conclusions are contained in the final section.

2 The mathematical formulation of the model

This section can be skipped or skimmed with no loss in continuity. The model assumes that although the individual commuters differ as to which of the two roads is the preferred route for commuting and the benefits of a trip on either road, the peak and off-peak time periods are perfect substitutes on each road insofar as benefits are concerned. The time periods differ only as to the cost of travel. The aggregate benefits of road use are approximated as a quadratic function of the number of trips on each road (Nx and Ny, respectively):

(1) $B = \beta_1 N_x - \beta_2 N_x^2/2 + \beta_1 N_y - \beta_2 N_y^2 - \beta_3 N_x N_y$ ($\beta_1 > 0$, $\beta_2 > \beta_3 > 0$)

In this formulation the model implies that all direct benefits of road use are derived by the road users themselves. Each road user knows the benefit which she derives from a trip and travels if and only if this benefit equals or exceeds the cost incurred by that user. Accordingly the marginal benefits are regarded as inverse demand functions:

(2) $P_x = \beta_1 - \beta_2 N_x - \beta_3 N_y$

(3) $Py = \beta 1 - \beta 3 Nx - \beta 2 Ny$

If N_{xp} and N_{xo} (N_{yp} and N_{yo}) denote the number of trips on road X (Y) in the peak and off-peak period respectively:

(4) $N_x = N_{xp} + N_{xo}$

(5) $N_y = N_{yp} + N_{yo}$

The second derivatives are denoted as follows:

(6) $\partial P_x / \partial N_x = \partial P_y / \partial N_y = - B_2$

(7) $\partial P_x / \partial N_y = \partial P_y / \partial N_x = - B_3$

This model adopts the "production function" approach to modelling the production of production of road trips. Traffic volume (V) or flow is the rate at which aggregate travel is produced (measured for example in units of vehicles per hour). This depends on the road capacity K (measured as the maximum flow in vehicles per hour) and traffic density (D) (measured as vehicles per mile of road length):

(8) $V = V(D, K)$

For any given K, let D* represent the density for which V = K. The function V is assumed to have the following characteristics:

(9) $\partial V/\partial D > 0$ for $D < D^*$

(10) $\partial V/\partial D < 0$ for $D > D^*$

It is assumed that the elasticity of V with respect to D is less than or equal to 1:

(11) $\eta = (\partial V/\partial D)^*(D/V) \leq + 1$

This is true as long as increases in density do not increase the speed of traffic flow. The "uneconomic region" of production where $\partial V/\partial D < 0$ is not uncommon in many urban areas and is referred to as "hypercongestion". The speed of traffic on the highway is given by:

(12) Speed = V/D (miles per hour)

This model assumes all commuting trips are for the same distance L_0. The time required for a single trip of a distance L_0 would be (distance/speed):

(13) Time for Trip = $L_0 D/V$

All travel within each period on a given road is assumed to take place simultaneously so that a fixed relation exists between the number of trips and traffic density:

(14) $D = N/L_1$

L_1 is the length of the road. Substituting for D:

(15) $V= V(N/L_1,K)$

The corresponding trip time would be L_0N/L_1V. The "average cost" of a road trip on each road includes the time cost borne by that user. Accordingly, the average cost on road X (Y) in the peak period is defined as:

(16) $AC_{xp} = C_nL_0N_{xp}/L_1V_{xp}$

(17) $AC_{yp} = C_nL_0N_{yp}/L_1V_{yp}$

Hereinafter the constants L_0 and L_1 which play no role are suppressed. The "average cost" in the off-peak period includes a similar time cost element and in addition a delay or waiting cost that is that is proportional to the duration of the peak period:

(18) $AC_{xo} = C_nN_{xo}/V_{xo} + C_wN_{xp}/V_{xp}$

(19) $AC_{yo} = C_nN_{yo}/V_{yo} + C_wN_{yp}/V_{yp}$

The total "variable" cost for each road is:

(20) $TVC_x = AC_{xp}N_{xp} + AC_{xo}N_{xo}$

(21) $TVC_y = AC_{yp}N_{yp} + AC_{yo}N_{yo}$

The net benefits of road use are:

(22) $W = B - TVC_x - TVC_y - C_kK_x - C_kK_y$

The unit cost of capacity is C_k. W is maximized subject to the constraints that the price on each road must equal the average cost on that road in each time period (peak and off-peak) plus any fixed toll T. Recall that the roads are not perfect substitutes, so their "total" prices in a given time period need not be equal. Also, recall that the average cost of road use includes the schedule delay cost in the off-peak period. Consequently:

(23) $P_x = T_{xp} + AC_{xp} = T_{xo} + AC_{xo}$

(24) $P_y = T_{yp} + AC_{yp} = T_{yo} + AC_{yo}$

Whenever the toll for a road and period is optimized, the constraint is removed for that road and period. Unless otherwise noted it is assumed the fixed (unoptimized) toll equals zero. The maximization is carried out over N_{xp}, N_{xo}, N_{yp}, N_{yo}, K_x and K_y. The constraints are represented by the Lagrange multipliers:

(25) $L_{xp} = \lambda_{xp} (P_x - T_{xp} - AC_{xp})$

(26) $L_{x0} = \lambda_{xo} (P_x - T_{x0} - AC_{x0})$

(27) $L_{yp} = \lambda_{yp} (P_y - T_{yp} - AC_{yp})$

(28) $L_{yo} = \lambda_{yo} (P_y - T_{yo} - AC_{yo})$

The general optimization problem then is to maximize:

(29) $W = B - TVC_x - TVC_y - C_kK_x - C_kK_y - L_{xp} - L_{xo} - L_{yp} - L_{yo}$

The marginal costs of road use on each road in each period are defined as follows:

(30) $MC_{xp} = AC_{xp} + (\partial AC_{xp}/\partial N_{xp})N_{xp} + (\partial AC_{xo}/\partial N_{xp})N_{xo}$

(31) $MC_{xo} = AC_{xo} + (\partial AC_{xo}/\partial N_{xo})N_{xo}$

(32) $MC_{yp} = AC_{yp} + (\partial AC_{yp}/\partial N_{yp})N_{yp} + (\partial AC_{yo}/\partial N_{yp})N_{yo}$

(33) $MC_{yo} = AC_{yo} + (\partial AC_{yo}/\partial N_{yo})N_{yo}$

The marginal benefits of capacity for each road are defined as:

(34) $MBK_x = \partial TC_x/\partial K$

(35) $MBK_y = \partial TC_y/\partial K$

The first order conditions for a maximum are:

(36) $P_x = MC_{xp} - \lambda_{xp}[\beta_2 + \partial AC_{xp}/\partial N_{xp}] - \lambda_{xo}[\beta_2 + \partial AC_{xo}/\partial N_{xp}] - \beta_3[\lambda_{yp} + \lambda_{yo}]$

(37) $P_x = MC_{xo} - \lambda_{xo}[\beta_2 + \partial AC_{xo}/\partial N_{xo}] - \lambda_{xp}\beta_2 - \beta_3[\lambda_{yp} + \lambda_{yo}]$

(38) $P_y = MC_{yp} - \lambda_{yp}[\beta_2 + \partial AC_{yp}/\partial N_{yp}] - \lambda_{yo}[\beta_2 + \partial AC_{yo}/\partial N_{yp}] - \beta_3[\lambda_{xp} + \lambda_{xo}]$

(39) $P_y = MC_{yo} - \lambda_{yo}[\beta_2 + \partial AC_{yo}/\partial N_{yo}] - \lambda_{yp}\beta_2 - \beta_3[\lambda_{xp} + \lambda_{xo}]$

(40) $MBK_x = MBK_y = C_k$

In the first best case, all of the multipliers drop out and price equals marginal cost in all periods on both roads. In all other cases, marginal cost exceeds price and the difference can be thought of as the marginal social loss from added roads trips in each period on each road (Wilson, 1983). By construction the marginal cost equals or exceeds the average cost on both roads in both periods. The coefficients of the multipliers are unambiguously positive but this does not guarantee that the multipliers are necessarily positive. In particular, in the no toll case, the multipliers for the off-peak roads are negative (but less than the positive peak period multipliers) indicating that the welfare effect of a subsidy in the off-peak period would at the margin be positive (other things equal).

The numerical calculations reported in the next section are based on the suggestion of R.G.D. Allen (1938) for a linear homogenous production functions with an "uneconomic" region of production:

(41) $V = SQRT[2\sigma KN - (\sigma N)^2]$

Note that σ is a dimensional constant reconciling the difference in units of measurement between N and K.

3 Numerical results and conclusions

Five cases are considered:

- Optimal tolls on both roads in both periods ("first-best" case)
- Tolls imposed during peak periods only
- Tolls (subsidy) for off-peak periods
- Tolls in both periods on one road only
- No Tolls ("baseline case")

The tolls in each case are optimized to maximize net benefits. In addition the road capacities are also optimized for each tolling scheme. All results are premised on the particular "baseline case" of no tolls which is selected. The results reported here are based on demand parameters chosen so that the own price elasticity of demand is approximately 0.5 (1/2) and the cross price elasticity of demand is approximately 0.25 (1/4). If these elasticities seem somewhat high it should be remembered that these are long-run elasticities. In the long run road users are able to vary both residential and employment locations in response to commuting conditions.

The cost parameters were selected to reflect a very heavily congested situation. The idea is that such situations are the not only the most interesting but also the most critical. In the baseline case, the no toll peak-period equilibrium traffic density is 15% above the density at capacity. In other words the roads are hypercongested, as the empirical results presented in this issue by McDonald for the Eisenhower Expressway indicate. The hypercongestion is not as great as it might seem at first impression. The traffic volume at this density has fallen only a little over 1% from the capacity flow. The peak-period speed is set at 50% of the off-peak average speed. For example, if off-peak speeds are 60 miles per hour, the peak speed would be slowed to 30 miles per hour by the congestion. In these conditions about 70% of the road use occurs during the peak period and 30% in the off-peak period. These statistics characterize the baseline model.

The net benefits are greatest of course for the first best case and the largest welfare loss results under the no tolls alternative. It is interesting to see what percentage of this maximum loss is recouped by the tolling scheme in each of the intermediate cases. Imposing peak period tolls captures 78% of the potential benefit. Subsidizing off-peak travel on both roads recovers only 45% of the potential benefits. Imposing tolls on only one road in both periods does not recover half the losses because of the substitution effect

on the other road. Only 35% of the loss is recouped by tolls on a single road for the baseline case examined.

A common criticism of the first-best tolling scheme is that the tolls required seem excessively high. How do the tolls of the "partial tolling" schemes compare to the first best tolls? To begin with, consider that in the first best case tolls are imposed in both the peak and off-peak periods. The off-peak period in this model is somewhat different from models where the off-peak periods have zero congestion. Here the off-peak period has substantial congestion. The off-peak first best tolls are a full 64% of the first-best peak tolls indicating a substantial disparity between marginal and average costs at the first-best equilibrium even in the off-peak period. Comparisons with the first-best peak toll should be judged in this light.

If tolls are imposed only in the peak period, the required toll is only 43% of the peak first-best toll. If only the off-peak periods are priced, the optimal subsidy is 29% of the first-best optimal toll. Finally if only one road is priced, the optimal peak toll is only 60% of first-best and the complementary off-peak subsidy is 24% of the first-best peak toll. Accordingly viewed in terms of political feasibility, the partial tolling schemes with their much lower toll (or subsidy) levels, merit obvious consideration.

Another way of judging the impact of various tolling regimes is in terms of the level of road use under each. An equity argument is often advanced that optimal tolls work by excluding the road users least able to pay the tolls (more precisely by excluding those whose marginal benefits do not exceed the marginal cost of their road use). This question can be answered in a number of ways. First the effect of tolls on the total number of trips can be examined. Secondly the percentage of trips occurring in the off-peak periods can be considered to assess the shifting effects of tolls. Finally if only one road is priced by tolls, the split between tolled and untolled travel can be compared.

Perhaps a preliminary step before calculating the direct impact on total road use is to calculate the point (own) price elasticity of demand at each equilibrium point. (The cross elasticity is 50% of own elasticity in all cases.) The first-best elasticity is .97 compared to the .48 of the baseline case. If tolls are imposed in peak periods only, the elasticity is .58. For the one road tolling regime the equilibrium point elasticities are .73 for the tolled road and .47 for the free road. Finally for the off-peak subsidy plan, the elasticity is .39 which is of course lower than the baseline case. The following results on total travel should be interpreted in the light of these changing elasticities.

The total use under first-best conditions is only 84% of the no tolls use. Under either the "peak tolls only" scheme or the "one-road only" scheme, the level of total road use is a surprising 96% of the baseline no tolls case. The plan of subsidizing off-peak travel produces an interesting result. This tolling scheme increases road use by 4% over the no tolls case. Obviously the subsidy expands road use in addition to relieving the peak-period congestion. In addition to affecting total road use, the tolling schemes cause major shifts in the timing of travel.

Recall that split between peak and off-peak travel in the baseline case was 70-30. First-best tolls result in 51% of trips occurring in the peak period and 49% in the off-peak. In other words the main improvement of first-best tolls is reflected in the more efficient distribution of trips over time. This is also reflected in the "peak only" tolling scheme where somewhat paradoxically the peak period has only 49% of trips while the off-peak has 51%. The off-peak subsidy does not work quite as well. It leaves 55% of the travel in the peak period. Clearly this results from the expansion of travel which the subsidy brings

about. This expansion limits the amount by which peak congestion can be drawn off into the less congested off-peak period.

The "one-road" tolls case requires special attention because it involves a shift between roads as well as between periods and the timing shift is different for the two roads. Only 45% of total travel takes place on the toll road. This 45 % is divided between the peak and off-peak periods in almost the same proportions as in the first best case, 51% in peak, 49% in off-peak. The 55% of total travel on the freeway, the split is very similar to the baseline split. (70-30). Thus the difference between peak only tolls and one road tolls is clearly contrasted. The peak only operates by shifting travel to the off-peak periods while the one road toll scheme curtails travel on the toll road with little inter-period shifting.

There are also variations in operating characteristics among the tolling regimes. This is reflected in the speed of travel occurring each period. Maximum speeds are attained in the off-peak no tolls case and this speed is used as a standard of comparison. The first best speeds are very nearly equalized at 68% (peak) and 70% (off-peak) of the no toll off-peak speed. In both cases this is somewhat faster than congested speed of the peak no toll case, but perhaps not as large an improvement as might have expected. The "peak tolls only" speeds are similar but slightly slower at 66% (peak) and 64% (off-peak) of the least congested speed. Notable is the reversal in which the peak speed actually exceeds the off-peak speed. The off-peak subsidy results in the slowest peak period speeds of any of the partial tolling schemes at 60% of maximum with the highest off-peak speed of 70%. The "one-road-only" tolls again present a split picture. The toll road is very like the first best with speeds of 68% and 70% for the peak and off-peak periods respectively. The freeway has speeds similar to the 'no toll" case, but slightly slower at 48% versus 50% for peak travel and 97% versus 100% for off-peak travel. In general the various toll schemes improve speed in the peak period, but not as much as they decrease speed in the off-peak period. It should be noted that in all of the previous results the road capacity has been optimized for the case under consideration and therefore is different in each case. In fact the effects of different tolling schemes on the optimal road capacity may be the most important contribution of any long-run model of road congestion.

In general, tolling results in smaller road capacities. The percentage of the optimal baseline capacity is an important consideration in evaluating any of the plans. Recall however that the baseline model was set up with the existence of hypercongestion in which the traffic density was 15% above the density at the capacity flow. It is therefore also interesting to ask whether the tolling plans eliminate hypercongestion. For the baseline data considered here, they do. The first-best road capacity is 83% of the no toll optimum and the road operates at a density only 85% of maximum. Note that it would be very rare for a road to operate at 100% of optimal capacity. The optimal road capacity for the "peak tolls only" is slightly larger at 90% of the baseline capacity operating with a density that is 87% of the maximum. The off-peak subsidy is very similar to the baseline case with no tolls. The capacity size is 97% of the no toll baseline optimal capacity. However the road operates in the untolled peak period with a density that is only 97% of the density at capacity flow. The road is not hypercongested. The "one-road-only" scenario is again a divided case. The optimal capacity of the tollroad is 86% of the baseline case and has an equilibrium density of 85% of density at maximum flow. The free road however has a capacity that is actually 2% larger than the baseline capacity, and the road is even more congested with a density of 118% of the density consistent with maximum traffic volume. Perhaps the appropriate comparison in this case is the comparison of total investment in

road capacity for both roads together. Note that for other scenarios the two roads have equal capacities by model design. On a combined basis, the one road toll system has an overall investment in road capacity that is only 94% of the baseline capacity. Tolling always saves road capacity, at least in this model.

What lessons are revealed by these results the reader may ask? This is a political world where the greater efficiency of the first-best case may not be sufficient to offset other less equitable results. The first-best solution may not even be feasible from a political viewpoint. Consequently all of the tolling regimes considered need to be considered since each has particular advantages not possessed by the other alternatives. It may be that this fact is commonly recognized by practitioners in the field. The model presented here quantifies these features and perhaps clarifies the debate.

4 References

Allen, R. G. D. (1938) *Mathematical Analysis for Economists*. St. Martin's Press: New York, NY.

Else, P. (1981) "A Reformulation of the Theory of Optimal Congestion Taxes," *Journal of Transport Economics and Policy*, 15: 217-232.

McDonald, J. F., E. L. d'Ouville and L. N. Liu (1999) *Economics of Urban Highway Congestion and Pricing*. Kluwer Academic Publishers: Boston, MA.

Wilson, J. (1981) "Optimal Road Capacity in the Presence of Unpriced Congestion," *Journal of Urban Economics*, 13: 337-357.