Road Pricing in Practice and Theory

JOHN F. MCDONALD *

College of Business Administration, University of Illinois at Chicago

Abstract

This introductory article to the special issue discusses briefly both the practice and theory of road pricing. Actual methods of charging for road use are discussed, and the current trend towards greater use of time-of-day congestion tolls is noted. The basic model of traffic congestion and road pricing is presented next, and this is followed by a summary of the extensions to the basic model represented in this issue of the Review of Network Economics. Researchers are now developing models that come very close to presenting realistic congestion toll systems that are reasonably efficient.

1   Introduction

This article is a brief survey of both the practice and theory of road pricing. We begin with a look at road pricing in practice in Section 2, and then turn to the economic theory of efficient road pricing. The article by Philip Blythe discusses the technology of road pricing and its actual and planned use in the U.K. The theoretical discussion in Sections 3 and 4 below starts with a basic model of a single route that is subject to traffic congestion, and then turns to some of the numerous more elaborate models that have been developed in recent years. This last topic serves as an introduction to the other articles in this special issue of the Review of Network Economics. Each of the articles by David Boyce and Qian Xiong, L. N. Liu, Edmond d'Ouville, and Akiyama, Mun and Okushima adds a particular complication to the basic model. As a set, these articles provide the reader with a fine overview of economic models of traffic congestion and road pricing.

2   Road pricing in practice

This discussion shall concentrate primarily on road pricing in the U.S. and the U.K. In the U.S., streets and roads are the responsibility of local governments, while state governments supply state highways. The federal government has provided funds to the states for federal highways (for example, the Interstate Highway System, and so on). The local, state, and federal governments all levy taxes on gasoline, and this is the primary means through

---

* College of Business Administration, University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL, 60607, USA. E-mail: mcdonald@uic.edu

which highway construction is financed. Local streets are partly financed through other local taxation such as the property tax, and in some cases by impact fees levied on new development. Local and state governments also impose license fees on vehicles. In effect, then, the main system of road pricing in the U.S. is the gasoline tax. Many observers have noted that the rate of investment in infrastructure declined in the U.S. starting in the 1970s while traffic congestion in urban areas increased. Winston and Bosworth (1992) explained that the most important reason for falling infrastructure investment, especially in highways and roads, was a lack of revenues dedicated to the purpose. The federal tax on gasoline remained at four cents per gallon from 1960 to 1982, which meant that the real federal gas tax dropped from four cents to 1.3 cents per gallon. The federal gas tax has since been raised and is now 14 cents per gallon. State taxes on gas also remained flat in nominal terms from 1960 to 1980 at about nine cents per gallon. Since then the average state tax on gas has increased to approximately 16 cents per gallon. It seems that the sharp increase in gasoline prices in the 1970s caused by the Organization of Petroleum Exporting Countries deterred governments from raising taxes (or keeping them constant in real terms). The price of gasoline dropped in the 1980s; government officials evidently felt that gas taxes could be increased in an era of declining pump prices. The recovery of spending on highways and roads was the result. Much of that spending was financed by the rising gas taxes that increased monies in the federal and state highway trust funds.

Many states in the U.S. also operate tollways. For example, the State of Illinois operates an extensive system of tollways in the Chicago metropolitan area. The tolls are set to pay off the tax-exempt bonds that were issued to finance construction, and do not vary by time of day (that is, with the level of traffic).

The use of gasoline taxes and tolls that do not vary by time of day (in the case of tollways) to "price" roads ignores the fact that efficient use of a highway – or any other economic good – requires that users pay a price that is equal to the marginal cost of their use. In the case of highway users the marginal cost of use includes the congestion cost imposed on others during times of heavy traffic. Traffic congestion is a classic example of a negative externality; a cost is imposed for which there is no corresponding payment. The standard model of traffic congestion that is discussed in the next section leads to the conclusion that economic efficiency will be improved by a congestion toll that equals the difference between the marginal cost and the average cost of the marginal trip on a highway route. This point has been recognized in the economics literature since the 1920s, but tolls resembling congestion tolls have been imposed only fairly recently.

There is growing interest around the world in implementing congestion tolls in some form. Small and Gomez-Ibanez (1998) categorize congestion pricing methods as follows:

(1)   City center congestion pricing,

(2)   Toll rings,

(3)   Single facility congestion pricing, and

(4)   Area-wide congestion pricing.

City center congestion pricing was pioneered in Singapore in 1975, and their system for 23 years used a paper windshield sticker that permits a vehicle to enter the city center. An electronic "smart card" system was introduced in 1998. A fee for entering the restricted zone applies from 7:30 A.M. to 6:30 P.M., and the highest fee is charged for entry between

8 A.M. and 9 A.M. The amount of traffic entering the restricted zone dropped by 44% after the scheme was put into operation in 1975. However, traffic congestion was created in areas just outside the zone.

The biggest news in recent years is the imposition of city center congestion pricing in London. Those who wish to drive in Central London between the hours of 7:00 A.M. and 6:30 P.M. must pay a fee of 5 pounds. Vehicles are monitored by 230 cameras, and violators are fined 80 pounds. The congestion fee has reduced the level of traffic in Central London by 15% to 20%, and cars are able to move about 40% faster than before. One can consult the Transport for London web site for the details of the system.

Toll ring systems exist in three cities in Norway (Bergen, Oslo, and Trondheim). These tolls are used primarily to raise revenue to pay for transportation facilities and do not vary appreciably by time of day. Economists in Norway are studying alternative road pricing schemes.

Two individual facilities that are subject to congestion tolls are Autoroute A1, the expressway that connects Paris north to Lille, and State Route 91 in southern California. And in New York City it costs more to use certain bridges and tunnels in the rush hour. Autoroute A1 has a toll designed to spread the Sunday evening peak traffic, and has influenced successfully the timing of trips in the desired manner. Since December 1995 State Route 91 has had a system of congestion tolls that are assessed for use of two new lanes (in either direction) via a transponder mounted on the dashboard of the auto. This example is discussed in detail by McDonald, d'Ouville, and Liu (1999). Also, since 1996 a high-occupancy land on an expressway in San Diego has been available for use by low occupancy vehicles (that is, one person) if a permit is purchased. This fee is, in effect, a peak-period toll because there is little or no incentive to buy a permit if one's use of the expressway does not occur in the peak periods. Public acceptance of both California projects has been good, probably because paying the toll can be seen as paying for access to additional road capacity. Proposals for peak-period tolls on other individual facilities have been defeated by public opposition (in Seattle and San Francisco, for example).

Systems for area-wide congestion pricing have been under intensive study in the Randstad region of the Netherlands and in London, but not implemented. Instead the decision in London was to impose the central city congestion toll. As the article by Akiyama, Mun, and Okushima explains, various forms of area-wide pricing are being studied for Osaka and several other metropolitan areas around the world.

The article by Philip Blythe discusses the evolution of road pricing policy in the U.K. An important event in that evolution is the adoption of the Transport Act of 2000, which gives local authorities an incentive to adopt congestion pricing by requiring that such revenues be retained and invested in local transport facilities. Blythe also discusses technical options for road pricing, the city center pricing in Durham and London, and other projects that are being planned.

3   Highway congestion: The basic model

The presentation of the basic highway congestion model follows McDonald, d'Ouville, and Liu (1999). The fundamental equation for traffic volume says that

(1) $V = DS$,

where V is traffic volume (vehicles per hour), D is traffic density (number of vehicles per mile in a photograph of the highway), and S is average speed of the vehicles. Consider a highway of one mile in length and a time period of one hour. Assume that the highway is in a steady state over this unit of time, so that V, D, and S are constant. It is helpful to imagine that the highway is a circle of unit circumference. Density D is the number of vehicles on the "race track", so the total quantity of travel produced and consumed during one hour is given by

(2) $Q = DS = V$.

Each vehicle on the circular highway will pass a given point on the circle S times per hour; the total quantity of travel Q is thus $DS = V$, which establishes the numerical identity between traffic volume V and total quantity of travel. This analysis can be converted into the conventional economics of production by showing that traffic density D is numerically identical to the variable input (that is, labor and vehicle time). Speed is the inverse of average variable cost measured in units of time, and volume is the output of the highway, thus

(3) $AVC = (1/S) = D/V$.

where AVC is average variable time cost per mile of travel. Total variable cost is simply

(4) $VC = (AVC)V = D$.

In this example D vehicles (and their drivers) are employed for one hour to produce V miles of travel. Suppose that there are 50 cars and drivers on the "race track," and that average speed is 30 mph. These 50 drivers produce 1500 miles of travel in one hour using 50 hours of driver and vehicle time. This leads to the specification of a production function that can be written

(5) $V = f(D,K)$,

where K is the capital embodied in the highway.
    The next step is to derive the marginal cost of traffic volume (the output), given the average variable cost. Marginal cost in the short run is defined as the change in total variable cost as output increases by one unit;

(6) $MC = dVC/dV = d(AVC{\times}V)/dV$.

This marginal change consists of two parts, so that

(7) $MC = AVC + V(dAVC/dV)$.

This is the crucial result; it says that the marginal cost of output (traffic volume) is equal to average variable cost plus the change in average variable cost with respect to traffic volume times the traffic volume. It is known that average variable cost rises (that is, average speed falls) as traffic volume rises (up to the maximum traffic volume). However, the individual traveller makes the trip at average variable cost (the inverse of average speed). In the absence of a toll this is the cost that the traveller sees and is the price that the traveller uses to make the decision whether to travel on the highway at that time. The external cost is therefore the difference between marginal cost and average cost, that is,

(8) $MC - AVC = V(dAVC/dV) =$ efficient toll.

The efficient toll is equal to this external cost, traffic volume times the change in average variable cost with respect to an increase in traffic volume.

The argument for the use of this toll is that an efficient allocation of resources is created. The efficient toll produces an allocation of resources that is Pareto efficient. A reallocation of resources away from the efficient level of output (that is, traffic volume) cannot be accomplished without reducing the welfare of at least one person, while holding the welfare of all other people constant. For example, if traffic volume were to increase beyond the efficient level, the cost of the additional volume would exceed the benefits of that output. Suppose that the value of the benefits of an additional one-mile trip are $2.00 to the traveller, but that the cost of the added trip is $2.10. Further suppose that the consumer of the added trip pays $2.00, so that this person stays at the same level of well-being (utility). However, the additional cost of the trip of $.10 must have been imposed on someone, who is worse off by that amount.

The argument in favor of the congestion toll is somewhat more complex if we take as given a level of traffic volume that is greater than the efficient output, the result that happens if there is no congestion toll. If travellers only face a price equal to AVC, then they produce and consume a level of traffic volume that exceeds the efficient level. The imposition of a toll means that toll revenue must be collected from those travellers who continue to travel on the highway at the time in question. These travellers have, of course, experienced an increase in the cost of travel in order to accomplish the reduction in traffic volume. These travellers have been made worse off unless some form of compensation can be arranged for them. We suppose that no form of compensation can be arranged because we think that no compensation could be paid that toll payers would not see as tied to tollway use. So what is the argument in favor of a policy that clearly makes some people worse off?

The argument is that the benefits of the policy outweigh the costs; there is a potential Pareto improvement. Hypothetically it is possible that those upon whom costs are imposed can be compensated out of the benefits of the program, with something left over. Those upon whom costs are imposed in fact are not compensated. The critical point is that the real average cost of traffic volume has declined at the same time toll revenue has been collected from tollway users. The toll revenue is money that has been redistributed from tollway users to the tollway authority (that is, the government, or society are large). In the process real resources in the form of person and vehicle time have been freed to engage in other activities with economic value. The toll revenue is not a cost to society as a whole, but only a redistribution of income. If society is neutral about the net effect of this redistribution of income, then the congestion toll policy generates a net benefit (benefits

exceed costs). However, the fact that travellers are made worse off by a congestion toll explains the political opposition to road pricing schemes.

Many empirical estimates of the relationship between traffic volume and traffic density exist in the literature. Recall that, for equation (5), the volume-density function can be considered to be a production function. McDonald, d'Ouville, and Liu (1999) included a study of hourly traffic volume on the Eisenhower Expressway in Chicago for a week in May 1994. The highway under study has three lanes in each direction. The outbound portion of the highway was studied, and the estimated function related traffic volume for an hour to the occupancy of the highway (that is, the percentage of the highway occupied by vehicles during that time). A plot of the data shows clearly that traffic volume rises with occupancy, reaches a maximum at an occupancy rate of about 22%, and then declines. Regression results are as follows:

(9) Vol = -18,008.5 – 522.36 Occ + 11,393.7 ln(Occ),
    (11.3) (15.2)   (14.8)

where Vol = traffic volume per hour, Occ = occupancy rate for that hour, unsigned t values are in parentheses, the R squared is .614, and the sample size is 150. The estimated equation implies that

(10) dVol/dOcc = -522.38 + 11,393.7/Occ.

Traffic volume is at a maximum of 5718 vehicles per hour (1906 per lane) when the occupancy rate is 21.8%. The mean occupancy rate for the morning and afternoon rush periods combined is 23.8%. Hypercongestion can be defined as occupancy (densitiy) in excess of the value at which traffic volume is a maximum – the addition of a vehicle to the traffic actually reduces traffic volume.

The model presented in this section is fine as far as it goes, but it does not represent important aspects of real road systems. The purpose of a congestion toll is to reduce the traffic at peak times, but what happens to those travellers who change their behavior in response to the toll and choose not to travel on that route at that time? Suppose they travel on a different route that is also subject to congestion, or suppose they travel at another time of day with traffic congestion. The message of the basic model is that an efficient congestion toll must be imposed on all routes at all times of day, which is likely a practical impossibility. The efficient toll is zero during the middle of the night, of course, but the design of efficient congestion tolls for a real urban area is a very complex business.

4   Extensions of the basic traffic congestion model

The first theoretical study of a complication to the basic traffic congestion model was conducted by McDonald (1995). I considered two routes that are perfect substitutes in the minds of the travellers for getting from an origin to a destination, but imposed the condition that only one route can have a congestion toll. The point of the article is to study the complication of traffic diversion to another congested highway. The basic theoretical result is that the "second-best" optimal toll in this case is

(11) efficient toll = $V_t(dAVC_t/dV_t) + (dV_f/dV_t)V_f(dAVC_f/dV_f)$,

where the subscripts t and f refer to the tollway and the freeway. The efficient toll now equals the standard toll from equation (8) above plus an adjustment term that equals the change in traffic volume on the freeway per unit change in volume on the tollway times the marginal congestion cost on the freeway. If the two routes are substitutes, traffic volume on the freeway goes up as volume on the tollway goes down, and the adjustment term is negative. Numerical examples studied by McDonald (1995) showed that this second-best toll could be only 25% to 50% of the efficient toll that would be imposed on both routes in accordance with equation (8), depending on the capacities and speed limits on the two routes. For example, if the two routes are identical in capacity and speed limit, the efficient second-best toll is only 26% of the efficient toll that would be imposed on both routes, and this toll yields only 8% of the potential welfare gain. In short, given that the administration of a toll has a real cost, there may be situations in which it is better not to impose a congestion toll. This situation can arise if there is a serious problem of traffic diversion onto congested routes upon which a toll cannot be imposed for practical or political reasons.

On the other hand, if the tollway and freeway are complements, traffic volumes on the two roads change in the same direction. This happens if, for instance, only a portion of the route from point A to point B can have a toll imposed. As an example, consider the bridges to New York City or San Francisco that lead directly to local arterial streets. In this case the adjustment term for the second-best toll is positive; the toll is higher than the standard first-best toll to account for the traffic congestion on the complementary streets. The notion of a toll on these bridges as a general congestion "entry fee" is closely related to the city center tolls in use in Singapore, London, and a few other cities that were discussed above. Imposition of efficient second-best tolls becomes a very complex matter if there are substitute and complementary routes that are subject to congestion and, for some reason, cannot have a toll imposed.

The articles in this special issue of the Review of Network Economics present different and more complex extensions of the basic traffic congestion-pricing model. As a group they represent a reasonably complete collection of current extensions of the model. Researchers are currently making rapid progress in formulating more realistic traffic congestion models that can serve as better guides for policy.

The article by David Boyce and Qian Xiong extends the basic model to the case of a complex road network that represents the actual system of highways and arterial streets in the Chicago metropolitan area. They consider a particular time of day, take the total number of trips between each pair of origins and destinations as given, and compare the "user-optimal" and "system-optimal" assignments of travellers to routes. By user optimal they mean that the cost to the user of travelling from zone A to zone B is the same on those routes that are actually in use. Users base their decisions on average variable cost rather than the marginal cost to society that includes congestion cost. The system optimal solution means that total travel time for all the trips is minimized; the routes to which drivers are assigned have equal <u>marginal</u> travel times (including marginal congestion cost), and no unused route has a lower marginal travel time. The system optimal solution implies a set of congestion tolls to achieve this equality of marginal travel times, but Boyce and Xiong do not focus their attention on these tolls. Rather, their purpose is to explore the

change in route choices that are implied by the change to the system optimal solution. Their basic finding is that major arterial streets carry more traffic under the system optimal solution compared to the user optimal assignment.

The article by Louie Nan Liu includes two times of day (peak and off-peak), and characterizes the highway system as consisting of two routes. Given two routes, he explores the possibility that only one route will be subject to a congestion toll – as compared to the cases in which a toll is imposed on both routes or on neither route. This example is studied because, as mentioned above, the imposition of tolls on all routes may be a practical or political impossibility. The basic result is that the welfare gain from a system of "partial" congestion tolls depends critically upon the extent to which the highway system is covered by tolls, but that this gain from partial tolling systems does not exceed 50% of the potential gain. Liu's study also includes the intriguing result that, with partial congestion tolls, it may be efficient to reward travellers (that is, a negative toll) who drive in the off-peak period when congestion is very low.

Edmond d'Ouville extends the model used by Liu to the long run in which the highway facilities themselves are variable. There are two roads and two periods, and the roads are imperfect substitutes for commuting. Note an immediate result; if the two roads are perfect substitutes but one cannot have a toll imposed, it would be efficient only to build the road that can carry a congestion toll – and to build it with efficient capacity for traffic. The cases considered by d'Ouville include no tolls, peak-period tolls only on both routes, off-peak tolls only on both routes, peak and off-peak tolls on one route, and peak and off-peak tolls on both routes (the first-best solution). The case of no tolls is used as a benchmark, and the model is constructed to build in mild hypercongestion in the peak period in the absence of tolls. Recall that hypercongestion is defined as the situation in which the addition of a vehicle on the highway (that is, an increase in density) reduces traffic volume. This starting point is consistent with the empirical results noted above for an expressway in Chicago. A general result is that the efficient amount of road capacity is smaller under any system of tolls compared to the case of no tolls. The capacity under the first-best solution is 83% of capacity without tolls, and hypercongestion is eliminated.

Akiyama, Mun, and Okushima extend the model to consider a network model developed for the Osaka metropolitan area. Their model does not include peak and off-peak time periods; their tolls are uniform with respect to time of day. Their paper compares alternative ring toll systems. They also consider the "first-best" solution of efficient congestion tolls on all routes, a no-toll solution, and alternative toll systems (including the current system) on existing toll roads. They find that a ring toll system is a reasonably efficient congestion toll system for a real urban area. A next step for their research is to build peak and off-peak periods into the model.

5   Conclusion

This special issue of the Review of Network Economics is devoted to road pricing in practice and to examples of the latest theoretical extensions of the basic model of traffic congestion and urban road pricing. Researchers such as Boyce and Xiong are developing sophisticated methods for modelling an urban road network and computing efficient tolls. Other scholars such as Liu and d'Ouville are working on models with peak and off-peak periods in which only some routes in the network can carry congestion tolls, while

Akiyama, Mun, and Okushima have reached the point of being able to model reasonably realistic policy alternatives such as a toll ring system. Progress in this field is rapid, so do not be surprised if some day you find that you are paying a congestion toll (or altering your behavior to avoid one).

6   References

McDonald, J. (1995) "Urban Highway Congestion: An Analysis of Second-Best Tolls," *Transportation*, 22: 353-369.

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

Small, K. and J. Gomez-Ibanez (1998) "Road Pricing for Congestion Management: The Transition from Theory to Policy," in K. Button and E. Verhoef (eds.) *Road Pricing, Traffic Congestion and the Environment*. Edward Elgar: Cheltenham, England.

Winston, C. and B. Bosworth (1992) "Public Infrastructure," in H. Aaron and C. Schultze, (eds.) *Setting Domestic Priorities*. The Brookings Institution: Washington, DC.