OPTIMAL TRANSIT PRICING WITH CROWDING AND TRAFFIC CONGESTION: A DYNAMIC EQUILIBRIUM ANALYSIS[1]

André de Palma, Ecole Normale Supérieure de Cachan
Robin Lindsey, University of British Columbia
Guillaume Monchambert, KU Leuven

Introduction

Economists have long advocated congestion pricing as the best way to tackle traffic congestion. Yet congestion pricing is still fairly rare, and various second-best policies for congestion relief continue to gain attention. A leading candidate is to subsidize transit fares in order to attract people out of their cars. Subsidization is politically popular but it has several limitations. First, reducing fares below marginal social cost creates a deadweight loss from induced trips and it contributes to crowding which is a serious problem in many cities[2]. Second, if transit is a poor substitute for driving large fare reductions are needed to make a dent in traffic congestion. Third, if the own-price elasticity of car trips is large then any potential benefits from congestion relief will be largely offset by latent demand (Duranton and Turner, 2011). Finally, lowering fares exacerbates transit deficits.

Cities vary widely in their fare policies. Many levy fares that are constant throughout the day. Others have adopted some degree of time variation — either as peak-period surcharges (e.g., London and Washington, D.C.) or off-peak discounts (e.g., Singapore and Melbourne). The main goal of this paper is to analyze optimal fare policies when traffic congestion and transit crowding are both present. We use a dynamic model that accounts for trip-timing decisions and the evolution of transit crowding and traffic congestion over the course of a peak travel period. The focus is on how transit fares should be set to simultaneously address traffic congestion and transit crowding externalities, and how the level and time structure of fares affect overall efficiency of the two-mode system.

Literature Review

There are many studies of second-best transit pricing in the presence of traffic congestion.[3] One of the first is Glaister (1974) who used a model featuring cars and buses, peak and off-peak time periods, and parametric cross-price demand elasticities between each of the four mode-time period choices. Glaister showed that peak and off-peak fares should both be set below marginal social cost. The peak fare may be below the off-peak fare, and either fare can be zero or even negative. Glaister and Lewis (1978) extended Glaister's (1974) model to include a rail mode and congestion interaction between cars and buses. They explored the potential benefits from second-best transit pricing in the Greater London area. Proost and Van Dender (2008) conducted a similar analysis for London and Brussels using a more elaborate model. These and other studies reveal the role of own-price and cross-price demand elasticities in governing optimal fare policy. Nevertheless, their approach is limited by the use of discrete peak and off-peak time periods and parametric elasticities, and neglect of transit crowding.

Tabuchi (1993) advanced the treatment of time by using the bottleneck model to describe travelers' trip-timing decisions and the evolution of traffic congestion on the road. However, he assumed that transit service is provided by a rail system with sufficient capacity to deliver all passengers to the destination on time and without crowding. His model therefore features only a single fare, and cannot be used to study time-of-day fare variations. Huang (2000) built on Tabuchi (1993) by adding crowding costs, but retained the assumption that transit delivers users on time. Huang et al. (2007) relaxed this assumption by supposing that rail service is provided on multiple trains according to a timetable. However, they did not analyze optimal pricing for either mode. Kraus (2012) uses a similar model to examine how transit usage depends on the pricing of roads. He ignores crowding costs and assumes that train fares are set according to first-best pricing principles. de Palma, Kilani and Proost (2015) and de Palma, Lindsey and Monchambert (2015) do allow for crowding, but assume that transit is the only travel mode so that first-best transit pricing is de facto optimal.

The Model

The model incorporates components of the models in Huang (2000), Huang et al. (2007) and de Palma, Lindsey and Monchambert (2015). One origin is connected to one destination by a road and a train service with a separate right of way. Utility from travel is described by a quasi-linear utility function $U\left(N_R, N_A\right)+g$, where N_A is the number of car (automobile) trips, N_R is the number of rail trips, and g is a composite numeraire consumption good. Function $U\left(\cdot\right)$ is strictly quasiconcave so that car trips and rail trips are imperfect substitutes.

As in the Vickrey (1969) model, trip-timing preferences are described by a piecewise linear schedule delay cost function. A traveler departing at time *t* and arriving at time t_a incurs a combined travel time and schedule delay cost of

$$\alpha\left(t_a - t\right) + \beta\left(t^* - t_a\right)^+ + \gamma\left(t_a - t^*\right)^+ ,$$

where t^* is desired arrival time at the destination, α is the unit cost of time spent traveling, β is the unit cost of arriving early, and γ is the unit cost of arriving late.

Congestion on the road takes the form of queuing behind a bottleneck. The cost of a car trip departing at *t* and arriving at t_a is:

$$C_A\left(t\right) = \bar{C}_{A0} + \alpha q\left(t\right) + \beta\left(t^* - t - q\left(t\right)\right)^+ + \gamma\left(t + q\left(t\right) - t^*\right)^+ + \pi\left(t\right),$$

where C_{A0} is the free-flow cost of a car trip, $q\left(t\right)$ is queuing delay and $\pi\left(t\right)$ is the road toll (if any) at time *t*.

To simplify analysis, travel time by train is normalized to zero so that $t_a = t$. Train service is assumed to be provided continuously and at a constant capacity rate over a fixed time interval $\left[t_0, t_e\right]$ where $t_0 < t^* < t_e$. The cost of a train trip at *t* is:

$$C_R\left(t\right) = \bar{C}_{R0} + \lambda n\left(t\right) + \beta\left(t^* - t\right)^+ + \gamma\left(t - t^*\right)^+ + \tau\left(t\right),$$

where C_{R0} is the fixed cost of a train trip (e.g., the time cost of access and egress time), $n\left(t\right)$ is the number of users taking the train at time *t*, λ is a parameter measuring disutility from crowding, and $\tau\left(t\right)$ is the fare at time *t*.

Users have heterogeneous preferences.[4] There are two user groups[5], 1 and 2, that differ with respect to parameters α , β , γ and λ , but have the same values of γ / β and t^* . Parameter values satisfy four conditions. First, $\beta_1 \leq \beta_2$ so that group 2 has stronger on-time preferences than group 1. Second,

$\beta_1/\alpha_1 < \beta_2/\alpha_2$. This implies that group 2 tolerates queuing more than group 1. Group 2 arrives by car closer to t^* than group 1, and creates a higher marginal external congestion cost when queuing occurs. Third, $\beta_1/\lambda_1 < \beta_2/\lambda_2$. This implies that group 2 tolerates crowding more than group 1, and arrives by train closer to t^* than group 1. Finally, $\gamma_1/\beta_1 = \gamma_2/\beta_2$ so that the two groups value late arrival relative to early arrival by the same ratio. Train loads are determined by the numbers of users in each group.

Results

Preliminary results have been derived for the regimes shown in Table 1:

Table 1: Administrative and pricing regimes

Fare policy	**Road toll**	**Transit fare**
First-best optimum		
First-best	Optimal time-varying	Time-varying
Second-best	None	Time-varying
Third-best	None	Time-varying
Flat fare	None	Flat
Free transit	None	None

The first-best optimum and free transit regimes serve as benchmarks against which the efficiency of the other regimes is measured. In the first-best, second-best and third-best pricing regimes, the fare can be varied freely over time but it is anonymous in the sense that it cannot depend on whether a user belongs to group 1 or group 2. In the first-best pricing regime the road is priced optimally to eliminate queuing at the bottleneck and the fare schedule is chosen to optimize welfare. In the second-best pricing regime the road is not tolled and the transit operator takes traffic congestion into account when setting the fare schedule. By contrast, in the third-best pricing regime the operator behaves myopically and neglects traffic congestion. Finally, in the flat-are regime the fare is restricted to be the same for all trains but the level of the common fare can be optimized.

Several general properties of the regimes have been established.

First-best optimum: In the first-best optimum, for each group the numbers of trips taken by each mode are chosen to equalize their marginal social costs. Passenger loads are also distributed across trains to equalize the marginal social costs of each trip by members of the same group. Trains arriving closer to t^* carry higher loads.

First-best pricing: In this regime the road is optimally tolled and the fare can be varied freely over time. Nevertheless, the first-best optimum still cannot be achieved unless $\lambda_2 = \lambda_1$. To see why, consider an early arrival period and suppose group 1 travels during the interval $\left[t_0, \hat{t}\right]$ and group 2 during the interval $\left[\hat{t}, t^*\right]$. If $\lambda_2 > \lambda_1$, train loads must decrease at $\hat{t}$ in order to provide less crowded conditions for group 2, but users in group 1 can then reduce their trip costs by deviating from the optimum and taking a train just after $\hat{t}$. Conversely, if $\lambda_2 < \lambda_1$ train loads must increase at $\hat{t}$ but group 2 users can then reduce their costs by taking a train just before $\hat{t}$. Rescheduling trips this way can be deterred by introducing a suitable upward or downward jump in the fare at $\hat{t}$, but doing so upsets optimality conditions for numbers of trips and modal splits.[6]

Second-best pricing: When the road is not tolled, second-best pricing calls for a transit subsidy. The size of the subsidy for each group depends on the traffic congestion externality it creates, its own-price demand elasticity, and the cross-price elasticity between modes. Because group 2 creates a larger traffic congestion externality than group 1, the subsidy is higher — ceteris paribus — for group 2. Since group 2 travels on peak-period trains this requires lowering peak-period fares more than off-peak fares. However, this policy is constrained by trip rescheduling incentives as with the first-best pricing regime.

Third-best pricing: Third-best pricing entails setting fares as if first-best conditions apply. Fares are thus set as in the first-best pricing regime, and too many car trips are made.

Flat fare: In the flat-fare regime it is impossible to price discriminate either between trains or between groups. The level of the fare is chosen to balance the costs of traffic congestion (which calls for a low fare) and the costs of excessive transit trips (which calls for a high fare). More precisely, the fare level is set to balance overpricing off-peak trips and trips by group 1, and underpricing peak-period trips and trips by group 2.

Although the model is fairly simple, the presence of user heterogeneity complicates the analytics and precludes analytical solutions. Numerical analysis reveals that, when car trips and transit trips are good substitutes, third-best pricing can be much less efficient than second-best pricing and it can also perform less well than an optimal flat fare. Differences between the regimes narrow using more realistic assumptions about the degree of substitutability between modes.

A Numerical Example

In this section we present a numerical example that is calibrated to yield results broadly consistent with empirical evidence. Aggregate travel demand by each group is described by a representative individual with a linear-quadratic utility function: $U\left(N_{Ri}, N_{Ai}\right) = a_i\left(N_{Ri} + N_{Ai}\right) - b_i\left(N_{Ri} + N_{Ai}\right)^2 - d_i N_{Ri} N_{Ai}$, where a_i, b_i and d_i are positive parameters, and $d_i < b_i$, $i = 1, 2$.

The β_i and γ_i parameters affect the equilibrium only through the composite parameter $\delta_i \equiv \beta_i \gamma_i / \left(\beta_i + \gamma_i\right)$ and numerical values are assigned directly to this composite. Parameter values for the various components of the model are given in Table 2. Group 2 differs from group 1 in having a higher choke price on trips (i.e., $a_2 > a_1$), and stronger on-time preferences (i.e., $\delta_2 > \delta_1$). Other preference parameters are assumed to be the same for the two groups.

Table 2: Base-case parameter values

a_1	\$60/trip	$\bar{c}_{R1} = \bar{c}_{R2}$	\$8/trip
a_2	\$75/trip	$\bar{c}_{A1} = \bar{c}_{A2}$	\$5/trip
$b_1 = b_2$	\$0.02/trip^2	$\lambda_1 = \lambda_2$	\$0.0005/passenger
$d_1 = d_2$	\$0.0667/trip^2	t^*	immaterial
$\alpha_1 = \alpha_2$	\$20/h	g	immaterial
δ_1	\$8/h	s	5000 vehicles/h
δ_2	\$16/h	Service interval $\left[t_0, t_e\right]$	1 h

Results are shown in Table 3.

Table 3: Equilibria

	Regimes in order of increasing efficiency				
	Free transit (*n*)	Flat fare (*f*)	Third-best (*3*)	Second-best (*2*)	First-best (*o*)
N_{R1}	1,690	1,691	1,603	1,693	1,603
N_{A1}	1,858	1,858	1,882	1,857	1,882
N_{R2}	2,233	2,234	2,106	2,195	2,106
N_{A2}	2,247	2,247	2,282	2,259	2,282
$N_{R1} + N_{R2}$	3,923	3,925	3,709	3,887	3,709
$N_{A1} + N_{A2}$	4,106	4,105	4,165	4,115	4,165
Fixed toll for group 1	\$0	-\$0.014	\$0	-\$1.82	\$0
Fixed toll for group 2	\$0	-\$0.014	\$0	-\$1.73	\$0
$\hat{t}$	0.808	0.808	0.715	0.704	0.715
Full price elasticities (group 1, group 2)					
Auto, Auto	-0.350, -0.380	-0.350, -0.380	-0.349, -0.377	-0.351, -0.379	-0.349, -0.377
Rail, Rail	-0.460, -0.387	-0.459, -0.386	-0.540, -0.472	-0.458, -0.411	-0.540, -0.472
Auto, Rail	0.139, 0.128	0.139, 0.128	0.153, 0.145	0.139, 0.133	0.153, 0.145
Rail, Auto	0.128, 0.127	0.128, 0.127	0.136, 0.136	0.128, 0.130	0.136, 0.136
Welfare components					
Total costs	\$113,205	\$113,221	\$107,982	\$109,865	\$89,940
CS_1	\$84,023	\$84,047	\$81,261	\$84,071	\$81,261
CS_2	\$133,844	\$133,875	\$128,480	\$132,216	\$128,480
Fare revenue	\$0	-\$55	\$10,795	\$4,502	\$2,408
Toll revenue	\$0	\$0	\$0	\$0	\$18,042
Surplus	\$217,867	\$217,867	\$220,537	\$220,814	\$238,579
Rel. efficiency r^i	0	0	0.129	0.141	1

CS_i denotes the aggregate consumers' surplus of group *i*. The relative efficiencies of the regimes are compared using the index $r^i \equiv \left(W^i - W^n\right)/\left(W^O - W^n\right)$ where *W* is social surplus or welfare, *i* indexes the regime, *n* denotes the free-fare regime and *o* denotes the first-best optimum.

Free transit: In the free-transit regime there is no fare and no toll so that transit crowding and queuing congestion both impose external costs. About 4,000 trips are made by each mode. Group 2 takes more rail trips than group 1, but group 2 restricts its trips to the 20 percent of trains that arrive closest to on-time. The own-price elasticity of demand for automobile trips is a little over one third: in line with estimates for short-run elasticities. The own-price elasticity of demand for transit trips is a little higher. A rule of thumb is that the elasticity is about one third. However, long-run elasticities can be considerably larger (Schmiek, 2016). Similar elasticities obtain in the other four regimes. Thus, the elasticities can be interpreted as applicable over an intermediate time interval of perhaps 1-2 years. Cross-price elasticities are about one third the magnitude of the own-price elasticities — reflecting the fact that automobile and rail trips are rather imperfect substitutes.

First-best optimum: At the opposite extreme to free transit is the first-best optimum in which optimal numbers of trips are chosen or each group by each mode. Because $\lambda_2 = \lambda_1$, the first-best optimum can be supported by first-best anonymous (i.e., no discriminatory) pricing. Since queuing congestion is more severe than transit crowding in the free-transit regime, the first-best optimum entails somewhat fewer rail trips by each group and more automobile trips. Both groups end up worse off as apparent from the decline in their consumers' surplus, but the losses are outweighed by transit revenue and substantial toll revenue. The overall gain amounts to $12,325 or about $3 per trip.

Flat fare: The optimal flat fare turns out to almost zero ($0.014) and yields no perceptible welfare gain. The reason for this is that the benefits from setting a negative fare to alleviate queuing congestion almost exactly balance the benefits from setting a positive fare to reduce excessive rail trips. Both externalities are higher or group 2, but with a flat fare it is not possible to discriminate between the two groups using the pricing mechanism.

Third-best fare: In the third-best fare regime the fare is varied over time to fully internalize rail crowding costs without considering congestion on the road. The fare structure is therefore the same as for the first-best optimum and no flat-fare component is added to or subtracted from the schedule. Consequently, the numbers of trips by each group using each mode, price elasticities of demand and consumers' surplus are identical to those in the first-best optimum.

Second-best fare: In the second-best fare regime the fare is varied over time as in the third-best regime but the fare level is decreased to reduce automobile travel. Because group 2 creates a larger negative traffic congestion externality than group 1, the optimal downward shift is larger for group 2 than group 1. However, trip rescheduling incentives prevent the unrestricted second-best fare from being implemented and the overall welfare gain compared to the third-best pricing is very limited. moreover, the second-best fare only yields about one-seventh of the welfare gain achieved from first-best pricing. The reason for this is that (given the parameter values chosen for the example) traffic congestion is more costly than transit crowding and can only be alleviated directly by levying a time-varying toll.

Conclusions and Directions for Future Research

In this paper we have taken a simple, first-cut analysis at studying the optimal level and time structure of transit fares when transit crowding and traffic congestion are both significant externalities. Extensive sensitivity analysis will be required to determine the degree to which transit fare discounts can be used to reduce peak-period automobile trips that create the most congestion without overloading the transit system and exacerbating crowding.

REFERENCES

de Palma, A., Kilani, M., and Proost, S. (2015), Discomfort in mass transit and its implication for scheduling and pricing, Transportation Research Part B, 71(1), 1-18.

de Palma, A., Lindsey, R., and Monchambert, G. (2015), The economics of crowding in public transport. Working paper, November 17.

Duranton, G., and Turner, M. A. (2011), The fundamental law of road congestion: Evidence from US cities, The American Economic Review, 101, 2616-2652.

Glaister, S. (1974), Generalised consumer surplus and public transport pricing, Economic Journal, 84(336), 849-867.

Glaister, S., and Lewis, D. (1978), An integrated fares policy for transport in London, Journal of Public Economics, 9(3), 341-355.

Huang, H.J. (2000), Fares and tolls in a competitive system of transit and highway: The case with two groups of commuters, Transportation Research Part E, 36(4), 267-284.

Huang, H.J., Tian, Q., Yang, H., and Gao, Z.Y. (2007), Modal split and commuting pattern on a bottleneck-constrained highway, Transportation Research Part E, 43(5), 578-590.

Kraus, M. (2012), Road pricing with optimal mass transit, Journal of Urban Economics, 72(2-3), 81-86.

OECD (2014), Valuing convenience in public transport. Technical report, ITF Round Tables.

Parry, I. W.H., and Small, K.A. (2009), Should urban transit subsidies be reduced?, American Economic Review, 99(3), 700-724.

Proost, S., and Van Dender, K. (2008), Optimal urban transport pricing in the presence of congestion, economies of density and costly public funds, Transportation Research Part A, 42(9), 1220-1230.

Prud'homme, R., Koning, M., Lenormand, L., and Fehr, A. (2012), Public transport congestion costs: the case of the Paris subway, Transport Policy, 21, 101-109.

Schmiek, P. (2016), Dynamic estimates of fare elasticity for U.S. public transit, Transportation Research Record, 2538, 96-101.

Small, K.A., and Verhoef, E.T. (2007), The Economics of Urban Transportation, Routledge.

Tabuchi, T. (1993), Bottleneck congestion and modal split, Journal of Urban Economics, 34, 414-431.

Veitch, T., Partridge, J., and Walker, J. (2013), Estimating the costs of over-crowding on Melbourne's rail system. 36th Australasian Transport Research Forum. Brisbane, Queensland, Australia.

Vickrey, W. S. (1969), Congestion theory and transport investment, American Economic Review, 59(2), 251-260.

[1] Regular paper.

[2] See OECD (2014), Prud'homme et al. (2012) and Veitch et al. (2013).

[3] For reviews see Small and Verhoef (2007, Section 4.5) and Parry and Small (2009).

[4] Heterogeneous preferences are a crucial element of the model. Without heterogeneity it is optimal, as in Kraus (2012), to internalize transit crowding costs by varying fares over time. Deviation from first-best pricing is limited to applying a uniform subsidy for all trains so that the time profile of the fare is the same as when car travel is efficiently priced (or not an option).

[5] Limiting heterogeneity to two types not only simplifies the analysis but also facilitates understanding the implications of heterogeneity in the various preference parameters.

[6] Arbitrage-like behaviour of this sort would also occur in a model with discrete train service if the headway between trains is sufficiently short.