Optimal transit pricing with crowding and traffic congestion: A dynamic equilibrium analysis

André de Palma, Ecole Normale Supérieure de Cachan, Laplace 302, 61 avenue du Président Wilson, 94235 Cachan, France. andre.depalma@ens-cachan.fr; 33 1 47405575

* Robin Lindsey, Sauder School of Business, University of British Columbia robin.lindsey@sauder.ubc.ca; 1 (604) 822-3323

Guillaume Monchambert, Ecole Normale Supérieure de Cachan, Laplace 302, 61 avenue du Président Wilson, 94235 Cachan, France. guillaume.monchambert@ens-cachan.fr

* Corresponding author.

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 disadvantages. 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. 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 vary by time of day in order to simultaneously address traffic congestion and transit crowding externalities.

---

1 See OECD (2014), Prud'homme et al. (2012) and Veitch et al. (2013).
LITERATURE REVIEW

There are many studies of second-best transit pricing in the presence of traffic congestion. 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 et al. (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 et al. (2015). One origin is connected to one destination by a road and a train service operating on a separate right of way. Utility from travel is described by a quasi-linear utility function $U(N_A, N_R) + g$, 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 $U(\cdot)$ 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

\[ \alpha(t_a-t) + \beta(t^*-t_a)^+ + \gamma(t_a-t^*)^+, \]

---

2 For reviews see Small and Verhoef (2007, Section 4.5) and Parry and Small (2009).
where $t^*$ is desired arrival time at the destination, $\alpha$ is the unit cost of time spent traveling, $\beta$ is the unit cost of arriving early, and $\gamma$ 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:

$$C_A(t) = C_{A0} + \alpha q(t) + \beta (t^*-t-q(t))^+ + \gamma (t+q(t)-t^*)^+ + \pi(t),$$

where $C_{A0}$ is the free-flow cost of a car trip, $q(t)$ is queuing delay and $\pi(t)$ 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 $[t_0,t_e]$ where $t_0 < t^* < t_e$. The cost of a train trip at $t$ is:

$$C_R(t) = C_{R0} + \lambda n(t) + \beta (t^*-t)^+ + \gamma (t-t^*)^+ + \tau(t),$$

where $C_{R0}$ is the access cost of a train trip, $n(t)$ is the number of users taking the train at time $t$, $\lambda$ is a parameter measuring disutility from crowding, and $\tau(t)$ is the fare at time $t$.

Users have heterogeneous preferences.\(^3\) There are two user groups, 1 and 2, that differ with respect to parameters $\alpha$, $\beta$, $\gamma$ and $\lambda$, but have the same values of $\gamma/\beta$ and $t^*$.\(^4\) Parameter values satisfy three 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. Train loads are determined by the numbers of users in each group.

---

\(^3\) 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).

\(^4\) Limiting heterogeneity to two types not only simplifies the analysis but also facilitates understanding the implications of heterogeneity in the various preference parameters.
RESULTS

Preliminary results have been derived for the following regimes:

<table>
<thead>
<tr>
<th>Regime</th>
<th>Fare policy</th>
<th>Road toll</th>
<th>Transit fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First-best optimum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>First-best</td>
<td>Optimal TOD</td>
<td>TOD</td>
</tr>
<tr>
<td>3</td>
<td>Second-best</td>
<td>None</td>
<td>TOD</td>
</tr>
<tr>
<td>4</td>
<td>Third-best</td>
<td>None</td>
<td>TOD</td>
</tr>
<tr>
<td>5</td>
<td>Flat fare</td>
<td>None</td>
<td>Flat</td>
</tr>
<tr>
<td>6</td>
<td>Free transit</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

Regimes 1 and 6 serve as benchmarks against which the efficiency of the other regimes is measured. In regimes 2-4, the fare can be varied freely over time but it is anonymous in the sense that it cannot depend on whether a user is in group 1 or group 2. In regime 2 the road is priced optimally to eliminate queuing at the bottleneck and the fare schedule is chosen to optimize welfare. In regime 3 the road is not tolled and the transit operator takes traffic congestion into account when setting the fare schedule. By contrast, in regime 4 the operator behaves myopically and neglects traffic congestion. Finally, in regime 5 the fare is restricted to be the same for all trains but the level of the common fare can be optimized.

Several properties of the regimes have been established.

Regime 1: In the first-best optimum, the numbers of trips taken by each mode are chosen to equalize their marginal social costs for each group. 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.

Regime 2: In regime 2 the road is optimally tolled and the fare can be varied freely over time, but 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 $[t_0, \hat{t}]$ and group 2 during the interval $[\hat{t}, t^*]$. 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.\(^5\)

\(^5\) Arbitrage would also occur in a model with discrete train service if the headway between trains is sufficiently short.
Regime 3: 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 in Regime 2.

Regime 4: Third-best pricing entails setting fares as if first-best conditions apply. Fares are thus set as in regime 2 and too many car trips are made.

Regime 5: 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 overpricing off-peak trips and trips by group 1, and underpricing peak-period trips and trips by group 2.

Preliminary numerical analysis indicates that third-best pricing results is much less efficient than second-best pricing. Third-best pricing can also perform less well than an optimal flat fare when car trips and transit trips are perfect substitutes. Differences between the regimes narrow using more realistic assumptions about the degree of substitutability between modes.

REFERENCES


