Mathematical Population Studies, 22:37–52, 2015
Copyright © 2015 KLD Associates, Inc.
ISSN: 0889-8480 print/1547-724X online
DOI: 10.1080/08898480.2013.836423

STOCHASTIC CONGESTION AND PRICING MODEL WITH ENDOGENOUS DEPARTURE TIME SELECTION AND HETEROGENEOUS TRAVELERS

Wuping Xin
KLD Engineering, Islandia, New York, USA

David Levinson
Department of Civil, Environmental, and Geo-Engineering,
University of Minnesota, Minneapolis, USA

In a stochastic roadway congestion and pricing model, one scheme (omniscient pricing) relies on the full knowledge of each individual journey cost and of early and late penalties of the traveler. A second scheme (observable pricing) is based on observed queuing delays only. Travelers are characterized by late-acceptance levels. The effects of various late-acceptance levels on congestion patterns with and without pricing are compared through simulations. The omniscient pricing scheme is most effective in suppressing the congestion at peak hours and in distributing travel demands over a longer time horizon. Heterogeneity of travelers reduces congestion when pricing is imposed, and congestion pricing becomes more effective when cost structures are diversified rather than identical. Omniscient pricing better reduces the expected total social cost; however, more travelers improve welfare individually with observable pricing. The benefits of a pricing scheme depend on travelers' cost structures and on the proportion of late-tolerant, late-averse, and late-neutral travelers in the population.

Keywords: commute; congestion pricing; omniscient and observable pricing schemes; road pricing; second-best pricing; social cost and individual welfare; stochastic network equilibrium; traveler departure patterns

1. INTRODUCTION

Congestion pricing can be modeled using econometrics, queuing theory, game theory, or with a bottleneck model. Econometric models employ time-dependent demand and delay functions with crude specifications of queue variation. In queuing models, queue values are stochastic but arrival rates are treated as exogenous and insensitive to equilibrium fees (Daniel, 1992). Game theory explains behaviors in conflicting situations (Littlechild and Thompson, 1977; Hildebrand et al., 1990; Kita, 1999; Hansen and Wenbei, 2001; Marcucci and Marini, 2003). Levinson (2005) and Zou and Levinson (2006) explored the "micro-foundations" of congestion and pricing, considering the interactions of departure strategies of two or more travelers or more.

Address correspondence to David Levinson, Department of Civil, Environmental, and Geo-Engineering, University of Minnesota, 500 Pillsbury Drive S.E., Minneapolis, MN, 55455 USA. E-mail: dlevinson@umn.edu

Vickrey (1969) developed the first bottleneck model to study the efficiency gains and the temporal distribution of departure times resulting from congestion pricing. In this model, individual departure times are endogenized and the variation of congestion over the rush hour is determined within the model (Arnott et al., 1998). Smith (1984) proved the existence of no-fee single bottleneck equilibrium, while Daganzo (1985) showed that Smith's equilibrium is unique. Ben-Akiva et al. (1984) modeled dynamic adjustments of commuters. Arnott et al. (1990a) developed a network model of parallel routes and analyzed the efficiency gains of step tolls instead of continuously varying tolls. Daniel (1992, 1995, 2001) combined Vickrey's bottleneck model with queuing theory for airport operations, with random arrivals and adjustments of scheduled arrival time.

We combine a bottleneck model with stochastic queuing for roadway congestion and pricing. Similar to the user equilibrium in route choice models, there is an analogous equilibrium in terms of selection of departure time. Each traveler selects the departure time that minimizes personal travel cost. At equilibrium, travelers cannot improve their individual travel costs by changing departure time unilaterally. Travelers vary in their late-acceptance levels. We focus on the effects of late-averse, late-tolerant, and late-neutral travelers on congestion patterns with and without pricing. A bottleneck exists immediately before arriving. The travel time from origin to bottleneck is not constant, but has a probability distribution. We model the queue as a Markov-Poisson process and take two pricing schemes. Omniscient pricing relies on the knowledge of each individual's journey cost and of early and late penalties; observable pricing involves only queuing delay.

2. MODEL

N travelers plan their trips between their homes and a single destination. Their desired arrival time is distributed uniformly over a period of time. There exists a bottleneck immediately before arriving. The bottleneck's queuing capacity is M and its deterministic service time is t_s per traveler. At this bottleneck, travelers are served one at a time. If two travelers or more arrive simultaneously, only one is served while the others queue. The travel time between the downstream end of the bottleneck and the destination is negligible as in the bottleneck model. Unlike most conventional bottleneck models, the travel time from home to the bottleneck is an independent identically distributed random variable following a certain probability distribution characterized by mean μ, variance σ^2, maximum $t_{\max}$, and minimum $t_{\min}$. The travel time has a random congestion-free component upstream of the bottleneck and a queuing delay component (conditional on the total number of vehicles) at the bottleneck. Travel time is independent from one traveler to the other for the congestion-free component, because travelers take independent routes (Figure 1). The total journey time from home to destination is the sum of travel time from home to the bottleneck, the queuing delay, and a fixed service time at the bottleneck. In Figure 1, the queuing delay time at the bottleneck is "random," in as much as the queue is unpredictable. The queuing delay is deterministic when the service time per vehicle is deterministic.

The time horizon is divided by an interval of t_s. D_n denotes the selected departure time of traveler n, and $p_n(t|D_n)$ denotes the probability that traveler n departing at D_n arrives at the bottleneck at time t. The expected bottleneck arrival rate at time t is:

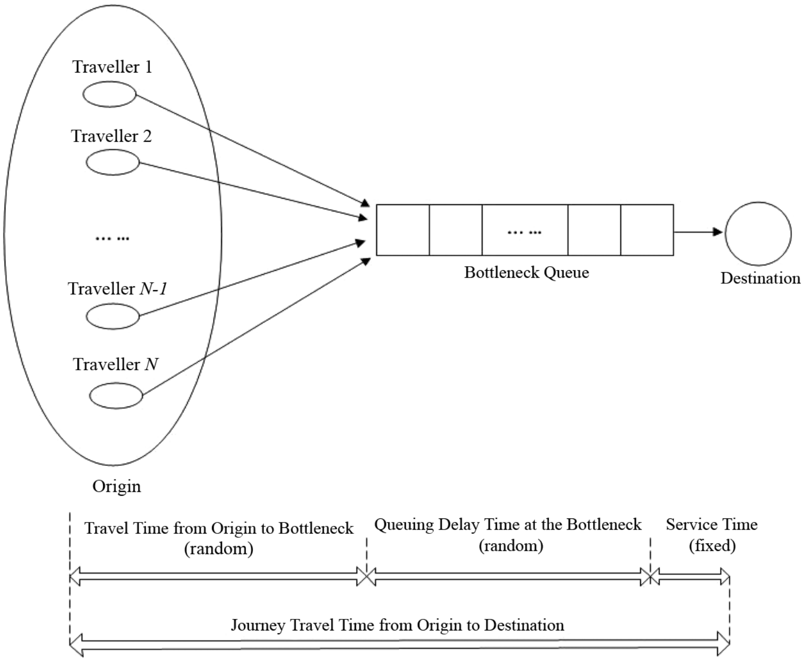
Traveller 1
Traveller 2
... ...
Traveller *N-1*
Traveller *N*
Origin
... ...
Bottleneck Queue
Destination
Travel Time from Origin to Bottleneck (random)
Queuing Delay Time at the Bottleneck (random)
Service Time (fixed)
Journey Travel Time from Origin to Destination

Figure 1. *N* travelers model diagram. There exists a bottleneck immediately before the destination. Total journey time is composed of travel time from the origins to the bottleneck and queuing delay at the bottleneck.

$$\lambda(t) = \sum_{n=1}^{N} p_n(t|D_n). \tag{1}$$

$q(t)$ denotes the $M+1$ dimensional probability vector describing the distribution of the total number of queuing vehicles at the bottleneck:

$$q(t) = (q_0(t), ...q_M(t))^T \tag{2}$$

with "T" denoting transposition and $q_i(t)$ is the probability of i queuing vehicles at the bottleneck at time t. The dynamics of $q(t)$ are described using a Markov chain as:

$$q(t+1) = Q(t)q(t), \tag{3}$$

where $Q(t)$ is the transition matrix:

$$\begin{pmatrix} \frac{\lambda^0(t)e^{-\lambda(t)}}{0!} & \frac{\lambda^0(t)e^{-\lambda(t)}}{0!} & 0 & 0 & \ldots & 0 \\ \frac{\lambda^1(t)e^{-\lambda(t)}}{1!} & \frac{\lambda^1(t)e^{-\lambda(t)}}{1!} & \frac{\lambda^0(t)e^{-\lambda(t)}}{0!} & 0 & \ldots & 0 \\ \frac{\lambda^2(t)e^{-\lambda(t)}}{2!} & \frac{\lambda^2(t)e^{-\lambda(t)}}{2!} & \frac{\lambda^1(t)e^{-\lambda(t)}}{1!} & \frac{\lambda^0(t)e^{-\lambda(t)}}{0!} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \frac{\lambda^0(t)e^{-\lambda(t)}}{0!} \\ 1-\sum_{k=0}^{M-1}\frac{\lambda^k(t)e^{-\lambda(t)}}{k!} & 1-\sum_{k=0}^{M-1}\frac{\lambda^k(t)e^{-\lambda(t)}}{k!} & 1-\sum_{k=0}^{M-2}\frac{\lambda^k(t)e^{-\lambda(t)}}{k!} & \ldots & \ldots & \frac{\lambda^0(t)e^{-\lambda(t)}}{0!} \end{pmatrix}. \tag{4}$$

The element $Q_{ij}(t)$ of the transition matrix $Q(t)$ at row i and column j describes the probability of j queuing vehicles at time t and i queuing vehicles at time $t+1$:

$$Q(t) = \left(Q_{ij}(t)\right)_{i,j} \text{ with } Q_{ij}(t) = P(\pi(t+1) = i|\pi(t) = j),\ i,j = 1,2,...M, \tag{5}$$

where $\pi(t)$ represents the total number of vehicles queuing at time t at the bottleneck. J_n is the unit journey cost, E_n the unit early penalty, L_n the unit late penalty, and A_n the desired arrival time of traveler n. The total travel cost of traveler n departing home at D_n and arriving at the bottleneck at t is:

$$\begin{aligned} C_n(t|D_n) = J_n\left((t - D_n) + \sum_{k=0}^{M} q_k(t)k\right) + E_n\left(\sum_{k=0}^{A_n-1} q_k(t)(1A_n - (t+k))\right) \\ + L_n\left(\sum_{k=A_n-t}^{M} q_k(t)(t+k-A_n)\right). \end{aligned} \tag{6}$$

The first term on the right hand side of Eq. (6) is the expected journey cost of traveler n, including the cost of traveling to the bottleneck $J_n(t - D_n)$ and the expected queuing cost $J_n\left(\sum_{k=0}^{M} q_k(t)k\right)$. The unit cost of travel time is treated as the same for free-flow time and queuing time. The expected queuing cost is computed by averaging the waiting time over the probability distribution of the total number of queuing vehicles at time t. The waiting time of traveler n equals the total number of queuing vehicles because the time scale is discretized by the fixed bottleneck service time and vehicles are served one at a time. The variable k refers to both the total number of queuing vehicles at the bottleneck at time t, and the queuing delay the vehicle is about to experience. The second term on the right hand side of Eq. (6) gives the expected cost of being early in relation to the desired arrival time A_n, the third term gives the expected cost of being late. From Eq. (6), the expected total personal cost (ETPC) of traveler n over all possible arrival times t is:

$$\sum_{t=0}^{\infty} p_n(t|D_n)C_n(t|D_n). \tag{7}$$

Traveler n will select a departure time D_n to minimize the traveler's expected total personal cost,

$$D_N^* = \arg\min_{D_n}\left(\sum_t p_n(t|D_n)C_n(t|D_n)\right). \tag{8}$$

The equilibrium is achieved when no traveler can lower the expected total personal cost by changing the departure time unilaterally. This gives the user-optimal departure time schedule:

$$D_{uo} = (D_1^*, D_2^*, \ldots, D_i^*, \ldots, D_N^*)^T, \tag{9}$$

where

$$D_i^* = \arg\min_{D_i}\left(\sum_{t=0}^{\infty} p_i\big(t|D_i\big).C_i\big(t|D_1^*, D_2^*, \ldots, D_i, \ldots, D_N^*\big)\right), \forall i = 1, 2, \ldots, N. \quad (10)$$

The expected total social cost (ETSC) of N travelers is:

$$\sum_{n=1}^{N}\left(\sum_{t=0}^{\infty} p_n(t|D_n)C_n(t|D_n)\right). \quad (11)$$

The system-optimal departure pattern is the one that optimizes the expected total social cost:

$$D_{so} = (D_1^{'}, D_2^{'}, \ldots, D_N^{'})^T = \arg \min_{(D_1, D_2, \ldots, D_N)^T}\left(\sum_{n=1}^{N}\sum_{t=0}^{\infty} p_n(t|D_n)C_n(t|D_n)\right). \quad (12)$$

2.1. Pricing Schemes

Omniscient pricing. If the transportation administrative agency is omniscient about each individual traveler's cost structure, defined as the valuation of J_n, E_n, L_n, and the desired arrival time A_n, for $n = 1, 2, 3, \ldots, N$, then a time-dependent "omniscient" pricing scheme imposing congestion fee, contingent upon the actual arrival time t, is:

$$F(t) = \sum_{n=1}^{N}\sum_{u=0}^{t} p_n(u|D_n)\frac{\partial C_n(u|D_n)}{\partial \lambda(t)}. \quad (13)$$

This fee is equivalent to the expected increase in the costs of all travelers caused by the increased expected arrival rate at time t. A time-dependent congestion fee as defined in Eq. (13) results in a departure pattern that minimizes the expected total social cost in Eq. (11). The first order condition to minimize the expected total social cost defined in Eq. (11) is:

$$\frac{\partial \text{ETSC}}{\partial D_n} = \sum_{t=0}^{\infty}\frac{\partial p_n(t|D_n)}{\partial D_n}C_n(t|D_n) + \sum_{n=1}^{N}\sum_{t=0}^{\infty} p_n(t|D_n)\frac{\partial C_n(t|D_n)}{\partial D_n} = 0, \forall n = 1, 2, \ldots, N. \quad (14)$$

The second term in Eq. (14) is the marginal increase in *all travelers'* costs induced by the nth traveler's departure time selection.

When the congestion fee $F(t)$ is imposed, each traveler n selects a departure time D_n to minimize the traveler's expected total personal cost. The traveler n's expected total personal cost (ETPC) including the congestion fee is:

$$\sum_{t=0}^{\infty} p_n(t|D_n)(C_n(t|D_n) + F(t)). \quad (15)$$

If travelers treat their own cost C_n and the congestion fee $F(t)$ parametrically and fix them as constant when selecting their departure times, then the first order condition for ETPC to be minimized is:

$$\frac{\partial \mathrm{ETPC}_n}{\partial D_n} = \sum_{t=0}^{\infty} \frac{\partial p_n(t|D_n)}{\partial D_n}(C_n(t|D_n) + F(t)) = 0, \forall n = 1, 2, \ldots, N. \tag{16}$$

Arrange Eq. (16) to obtain:

$$\frac{\partial \mathrm{ETPC}_n}{\partial D_n} = \sum_{t=0}^{\infty}\left(\frac{\partial p_n(t|D_n)}{\partial D_n}C_n(t|D_n)\right) + \sum_{t=0}^{\infty}\left(\frac{\partial p_n(t|D_n)}{\partial D_n}F(t)\right) = 0, \forall n = 1, 2, \ldots, N \tag{17}$$

$$\Rightarrow \frac{\partial \mathrm{ETPC}_n}{\partial D_n} = \sum_{t=0}^{\infty}\left(\frac{\partial p_n(t|D_n)}{\partial D_n}C_n(t|D_n)\right) + \sum_{t=0}^{\infty}\left(\frac{\partial p_n(t|D_n)}{\partial D_n}\sum_{n=1}^{N}\sum_{u=0}^{t} p_n(u|D_n)\frac{\partial C_n(u|D_n)}{\partial \lambda(t)}\right)$$

$$\Rightarrow \frac{\partial \mathrm{ETPC}_n}{\partial D_n} = \sum_{t=0}^{\infty}\left(\frac{\partial p_n(t|D_n)}{\partial D_n}C_n(t|D_n)\right) + \sum_{n=1}^{N}\sum_{u=0}^{\infty} p_n(u|D_n)\sum_{t=0}^{u}\left(\frac{\partial p_n(t|D_n)}{\partial D_n}\frac{\partial C_n(u|D_n)}{\partial \lambda(t)}\right).$$

From Eq. (1), $\frac{\partial p_n(t|D_n)}{\partial D_n} = \frac{\partial \lambda(t)}{\partial D_n}$, thus

$$\begin{aligned}\frac{\partial \mathrm{ETPC}_n}{\partial D_n} &= \sum_{t=0}^{\infty}\left(\frac{\partial p_n(t|D_n)}{\partial D_n}C_n(t|D_n)\right) + \sum_{n=1}^{N}\sum_{u=0}^{\infty} p_n(u|D_n)\sum_{t=0}^{u}\left(\frac{\partial \lambda(t)}{\partial D_n}\frac{\partial C_n(u|D_n)}{\partial \lambda(t)}\right)\\ &= \sum_{t=0}^{\infty}\left(\frac{\partial p_n(t|D_n)}{\partial D_n}.C_n(t|D_n)\right) + \sum_{n=1}^{N}\sum_{u=0}^{\infty} p_n(u|D_n)\frac{\partial C_n(u|D_n)}{\partial D_n}, \forall n = 1, 2, \ldots, N.\end{aligned} \tag{18}$$

Eq. (18) is the same first order condition as Eq. (14) for minimizing the expected total social cost (ETSC) when no fee is imposed.

Observable pricing. Usually an individual traveler's cost structure is unobservable to the transportation agency. A more realistic pricing scheme would only impose a congestion fee upon queuing delay. This gives observable price $\tilde{F}(t)$ as:

$$\tilde{F}(t) = \sum_{n=1}^{N}\sum_{t=0}^{\infty} p_n(u|D_n)\frac{\partial \tilde{C}_n(u|D_n)}{\partial \lambda(t)}, \tag{19}$$

where $\tilde{C}_n(t|D_n) = J_n\left(\sum_{k=0}^{M} q_k(t)k\right)$ is the queuing delay cost for traveler n. The congestion fee defined in Eq. (19) equals the marginal increase of the queuing cost for all travelers in relation to the marginal increase of expected arrival rate at time t.

Numerical computation of congestion fees. To compute the congestion fees $F(t)$ and $\tilde{F}(t)$, numerical evaluations of $\frac{\partial C_n(u|D_n)}{\partial \lambda(t)}$ and $\frac{\partial \tilde{C}_n(u|D_n)}{\partial \lambda(t)}$ in Eq. (13) and (19) are needed. These are equivalent to computing $\frac{\partial q_k(u)}{\partial \lambda(t)}, \forall k = 1, 2, \ldots, M; u = t+1, t+2, \ldots$ The matrix $Q'(t)$ is:

$$e^{-\lambda(t)} \begin{pmatrix} \frac{0\lambda^{-1}(t)-\lambda^0(t)}{0!} & \frac{0\lambda^{-1}(t)-\lambda^0(t)}{0!} & 0 & \ldots & 0 \\ \frac{1\lambda^{-1}(t)-\lambda^1(t)}{1!} & \frac{1\lambda^{-1}(t)-\lambda^1(t)}{1!} & 0 & \ldots & 0 \\ \frac{2\lambda^{-1}(t)-\lambda^2(t)}{2!} & \frac{2\lambda^{-1}(t)-\lambda^2(t)}{2!} & \frac{0\lambda^{-1}(t)-\lambda^0(t)}{0!} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \frac{0\lambda^{-1}(t)-\lambda^0(t)}{0!} \\ \sum_{k=0}^{M-1} \frac{\lambda^k(t)-k\lambda^{k-1}(t)}{k!} & \sum_{k=0}^{M-1} \frac{\lambda^k(t)-k\lambda^{k-1}(t)}{k!} & \ldots & \ldots & -1 \end{pmatrix}. \tag{20}$$

The chain rule leads to

$$\left(\frac{\partial q_1(u)}{\partial \lambda(t)}, \frac{\partial q_2(u)}{\partial \lambda(t)} \ldots \frac{\partial q_M(u)}{\partial \lambda(t)} \right)^T = Q(u) \ldots Q(t+2) Q(t+1) Q'(t) \begin{pmatrix} q_1(t) \\ q_2(t) \\ \ldots \\ q_M(t) \end{pmatrix}, \tag{21}$$

$\forall u = t+1, t+2, \ldots\ldots$

The congestion fees $F(t)$ and $\tilde{F}(t)$ are evaluated from Eq. (21). The derivation is similar to that of Daniel's (1995).

3. NUMERICAL EXAMPLE

The proposed congestion and pricing model was implemented as a standalone C++ program to facilitate the study of different congestion patterns in relation to various late-acceptance levels of travelers with and without pricing. Here, "late" is specific to a desired arrival time. For the simulation experiment, the travel time distribution from the origin to the bottleneck location takes the *double truncated normal* distribution DTN (μ, σ^2; $t_{\min}$, $t_{\max}$), where μ is the average, σ^2 the variance, $t_{\min}$ the minimal travel duration and $t_{\max}$ its maximal value. The program also allows other types of distribution such as Weibull or lognormal as well as user defined distributions. Specifically, in the simulation experiment, we assume

$\mu = 15$ min
$\sigma = 10$ min (this will be varied later)
$t_{\min} = 5$ min
$t_{\max} = 25$ min
Total number of travelers $N = 30$
Bottleneck queuing capacity $M = 15$ vehicles
Bottleneck service time 1 min/vehicle.

The bottleneck queuing capacity M is the maximum total number of vehicles which can be stored upstream of the bottleneck. The queuing capacity constraint was found never binding in the simulations. The feasible time horizon was taken between 14:00 and 16:00 pm, and the desired arrival times were distributed uniformly between 15:00 and 15:15 pm.

Travelers are late-averse, late-tolerant, or late-neutral. Late-averse travelers have higher penalties for being late than for being early so that they should depart early for fear of being late. Late-tolerant travelers have lower penalties for being late than for being early. Late-neutral travelers have identical early and late penalties. For simplicity, the settings for journey cost, early penalty, and late penalty are:

- Late-averse travelers: $J = 10$, $L = 20$, and $E = 10$;
- Late-tolerant (early-averse) travelers: $J = 10$, $L = 20$, and $E = 40$;
- Late-neutral travelers: $J = 10$, $L = 20$, and $E = 20$.

Results with other parameter values are similar. In the basic deterministic bottleneck model, an equilibrium with a finite departure rate exists only if $E < J$ (Arnott et al., 1990b). With the total number of drivers (rather than a continuum) and random arrival times, this constraint does not apply and $\mathrm{E} > J$ is allowed.

3.1. Congestion Pattern Without Pricing

When no pricing is imposed, four scenarios are considered:

1. All the travelers are late-averse (Figure 2);
2. All the travelers are late-tolerant (early-averse) (Figure 2);
3. All the travelers are late-neutral (Figure 3); and

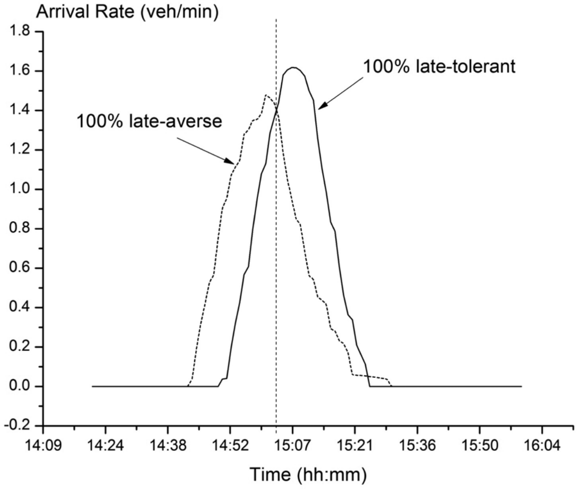

Figure 2. Expected bottleneck arrival rate: a traveler population with 100% of late-averse travelers against a traveler population of 100% late-tolerant travelers, no pricing.

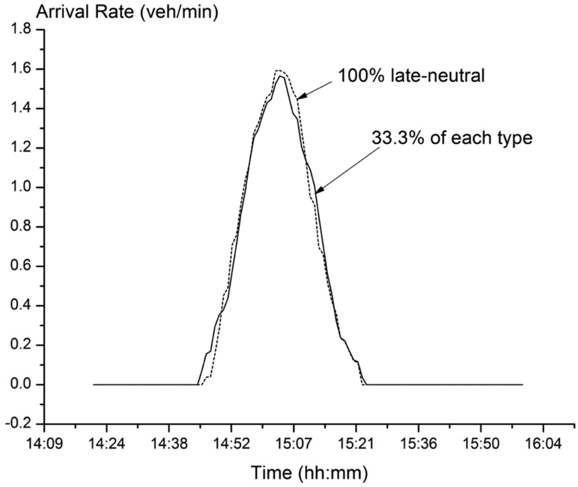

Figure 3. Expected bottleneck arrival rate: a traveler population with 100% of late-neutral travelers against a traveler population with 1/3 of late-averse, 1/3 of late-tolerant, and 1/3 of late-neutral travelers, no pricing.

4. One third of the travelers is late-averse, one third is late-tolerant, and one third is late-neutral (Figure 3).

Figures 2 and 3 show the congestion patterns under these scenarios. Figure 2 compares the congestion patterns under scenario 1 and scenario 2, which represent two extreme situations. When no pricing is imposed, under scenario 1, the congestion pattern is skewed toward the left, meaning arriving early; while under scenario 2 the congestion pattern is skewed toward the right, meaning that travelers do not favor early departure. These congestion patterns are expected because, under scenario 1, all travelers are late-averse and avoid arriving late by selecting early departures. Under scenario 2 all travelers are late-tolerant and do not care so much about being late because they have higher penalties for being early. In contrast to the “skewed” pattern in Figure 2, Figure 3 shows that when all travelers are late-neutral or when the three types of travelers are distributed equally, the congestion patterns are approximately symmetric around the desired arrival time. This is because, when late-acceptance is distributed equally, the influence of late-averse travelers and the influence of late-tolerant travelers on congestion counterbalance each other, resulting in a pattern similar to the one obtained when all travelers are late-neutral.

3.2. Congestion Pattern with Pricing

When pricing (either omniscient or observable) is imposed, three scenarios are considered:

1. One third of the travelers is late-averse, one third is late-tolerant, and one third is late-neutral (Figure 4);

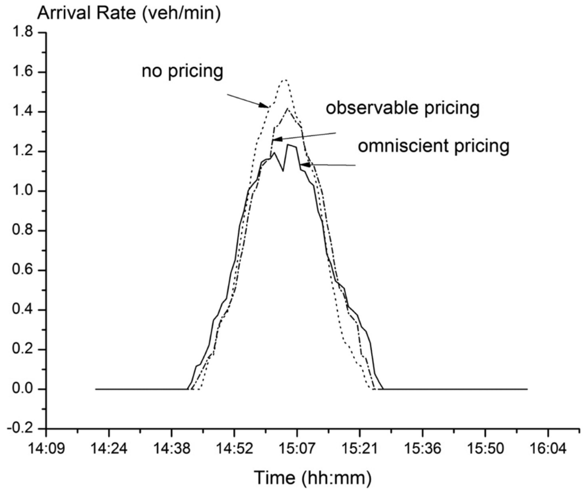

Figure 4. Congestion pattern: the traveler population is composed of 1/3 late-averse, 1/3 late-tolerant, and 1/3 late-neutral travelers, with pricing.

2. All travelers are late-averse (Figure 5); and
3. All travelers are late-tolerant (early-averse) (Figure 6).

Figures 4, 5, and 6 show the effects of different pricing schemes on congestion patterns. Omniscient pricing is most effective in suppressing the congestion at peak hours and in redistributing demands over a longer time horizon. The observable pricing scheme is also effective, but because it is based only on queuing cost and does not include early and late penalties, congestion is suppressed to a lesser extent.

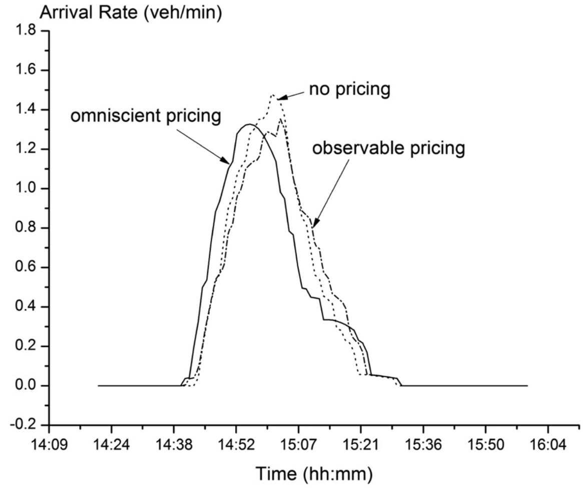

Figure 5. Congestion pattern: the traveler population is composed of 100% late-averse travelers, with pricing.

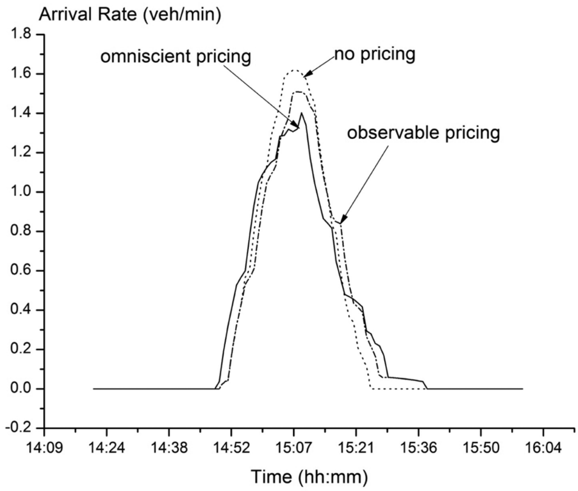

Figure 6. Congestion pattern: the traveler population is composed of 100% late-tolerant travelers, with pricing.

Moreover, one interesting observation is that congestion is more often suppressed with heterogeneous travelers (Figure 4) than with a single traveler type (Figures 5 and 6). This implies that congestion pricing is more effective when travelers' cost structures are diversified than when cost structures are identical. Figures 5 and 6 also suggest that the arrival pattern for the omniscient pricing scheme lies generally on the left of the arrival pattern for the no pricing scheme, whereas the arrival pattern for the observable pricing scheme lies on the right of the no pricing scheme. This implies that omniscient pricing prompts early departures while observable pricing postpones them.

3.3. Influence on Social Cost and Revenue

Table 1 presents the influence of different pricing schemes on the expected total social cost (ETSC), excluding congestion fees. The expected total social cost when the congestion fee is excluded is minimized under the omniscient pricing scheme. This is consistent with the theory that omniscient pricing minimizes the expected total social cost excluding the congestion fee, as shown in Eq. (14) and (18). However, the reduction of social cost from observable to omniscient pricing varies

Table 1. Percentage reduction in expected total social cost excluding congestion fee with pricing. ETSC excluding congestion fee = expected journey cost + expected early penalty + expected late penalty

Pricing scheme	1/3 each type	100% late averse	100% late tolerant	100% late neutral
Baseline ETSC excluding congestion fee	9298	8355	11351	9387
Observable Pricing	2.17%	1.19%	2.27%	1.76%
Omniscient Pricing	3.95%	1.37%	3.42%	3.59%

Table 2. Percentage increase in expected total social cost including congestion fee with pricing (revenue not returned to travelers). ETSC = expected journey cost + expected early penalty + expected late penalty

Pricing scheme	1/3 each type	100% late averse	100% late tolerant	100% late neutral
Baseline ETSC including congestion fee	9298	8355	11351	9387
Observable Pricing	15.8%	15.4%	14.4%	16.7%
Omniscient Pricing	14.9%	10.3%	22%	15.8%

with traveler compositions. When all travelers are late-averse, then the absolute reduction in social cost provided by omniscient pricing is less than in other traveler populations. The structure of the traveler population modifies the benefits, which a pricing scheme could bring, particularly when travelers are all late-averse.

Table 2 gives the expected total social cost with the addition of the congestion fee under different pricing schemes. When the congestion fee is taken into account, omniscient pricing results in lower expected total social cost than does observable pricing. Omniscient pricing reduces the ETSC because savings by omniscient pricing counterbalance the addition of a congestion fee. As the expected total social cost (with the addition of a congestion fee) is closely correlated with social welfare, the results in Table 2 imply that social welfare is improved with omniscient pricing compared with observable pricing. However, when all travelers are late-tolerant, the ETSC (with the addition of a congestion fee) resulting from observable pricing is less than with omniscient pricing, because the addition of a congestion fee exceeds the initial cost savings. If revenue is returned to travelers, the welfare outcome becomes different. This suggests again that all late-averse or all late-tolerant travelers' compositions deserve extra attention when analyzing pricing benefits.

Table 3 presents the revenues generated by tolling. Omniscient pricing generates more revenues (except when all are late averse). This can be because omniscient pricing includes the journey cost and early/late penalties for individual travelers, while observable pricing is based on queuing delay only. When all travelers are late-averse, the revenue generated with omniscient pricing becomes less than with observable pricing. However, Table 2 indicates that the expected total social cost under omniscient pricing is still smaller than with observable pricing. The amount of revenue generated by a certain pricing scheme cannot be employed solely to assess the effectiveness of that pricing scheme. This is even clearer when revenues are returned to travelers as equal shares (Table 4).

Table 4 presents the total number of travelers who become better off if revenues are returned as equal shares to each individual traveler. While overall social welfare improves with omniscient pricing (Table 2), with returned toll revenue, more travelers are better off with observable pricing than with omniscient pricing in each

Table 3. Expected revenue generated with different pricing schemes

Pricing scheme	1/3 each type	100% late averse	100% late tolerant	100% late neutral
Observable Pricing	1672	1386	1890	1729
Omniscient Pricing	1751	977	2890	1820

Table 4. Total number of travelers with improved personal welfare. An equal share of revenue is returned to each traveler (30 travelers in total)

Pricing scheme	1/3 each type	100% late averse	100% late tolerant	100% late neutral
Observable Pricing	28	24	30	30
Omniscient Pricing	27	20	24	26

case. This may come from the fact that omniscient pricing redistributes travel demand more uniformly. Subsequently, the congestion pattern is more spread out such that some travelers depart very early while others depart very late, resulting in a broader distribution of travel costs compared to observable pricing.

Cohen (1987) and Arnott et al. (1994) for a deterministic bottleneck have shown that influences of congestion pricing on welfare depend on the heterogeneity captured in the cost parameters. Daniel (2001) found that heterogeneity of costs tends to improve the welfare-distributional influences of congestion pricing. These earlier findings are consistent with the results presented here.

3.4. Influence of Travel Time Variability

Arnott and Kraus (1994) showed that if tolls can be varied freely over time and travelers cannot overtake each other, anonymous tolls (tolls which are independent of traveler type) suffice to enable an optimum. With pure bottleneck queuing congestion, the optimal toll should be able to eliminate queues while maintaining the capacity flow through the bottleneck. We set the observable pricing scheme in a similar logic. However, simulations indicate that the observable toll performs worse than the omniscient toll. This can be due to the fact that the arrival rate of vehicles at the bottleneck is random, contrary to the basic bottleneck model. Additional simulations are conducted to test smaller values of the standard deviation σ of travel time. Base case values of $J=20$, $L=20$, $E=20$, $\sigma=10$, $\mu=15$ are used, which corresponds to a large travel time variability (coefficient of variation CV = 0.67). Keeping other parameters unchanged, the influence of σ is tested by varying CV from 0.67 to 0.26. The expected total social cost (without tolls) under each pricing scheme is presented in Table 5. Under every scenario, omniscient pricing performs better than observable pricing. However, as the coefficient of variation decreases (the variability of travel time diminishes), the expected total social costs implied by observable pricing become closer to those under omniscient pricing. This suggests that when arrival rates at the bottleneck have a lower variance, a second-best observable pricing scheme performs almost as well as the first-best pricing,

Table 5. Percentage reduction in expected total social cost with different CV. ETSC excluding congestion fee = expected journey cost + expected early penalty + expected late penalty, with J = 20, L = 20, and E = 20, CV ranges from 0.67 to 0.26

Pricing scheme	CV = 0.67	CV = 0.53	CV = 0.4	CV = 0.33	CV = 0.26
Baseline ETSC excluding congestion fee	14241	14243	14266	14229	14713
Observable Pricing	3.8%	4.4%	4.6%	3.6%	3.1%
Omniscient Pricing	4.9%	4.9%	5.4%	3.9%	3.4%

which is the omniscient one. This is consistent with Arnott and Kraus (1994). From Table 5, the base ETSC without pricing has a tendency to increase when the CV is decreased, because when travel time has a lower variance, the system becomes less heterogeneous and the travel cost of each individual traveler weights more.

4. CONCLUSION

In this stochastic congestion and pricing model, travelers have a uniformly distributed desired arrival time for a given origin-destination pair, and each traveler selects the ‘‘best’’ departure time to minimize his or her expected total personal cost, which is the sum of journey cost, early penalty, late penalty, and a congestion fee. The tolling agency is either omniscient and knows every traveler’s cost structure (their detailed valuation of journey cost as well as early and late penalties), or observes only queuing delays. The simulations of different pricing schemes with varied compositions of late-averse, late-tolerant, and late-neutral travelers show that omniscient pricing is the most effective in suppressing peak hour congestion and distributing demands over a longer time horizon. Heterogeneity of travelers reduces congestion when pricing is imposed. Congestion pricing is then more effective when cost structures are diversified rather than identical.

When compared to observable pricing, omniscient pricing reduces the expected total social cost without congestion fee; this implies that omniscient pricing is better for social welfare than observable pricing. However more travelers improve welfare individually with observable rather than omniscient pricing. The ultimate benefits of any pricing scheme depend on the travelers’ cost structure and on the frequency of late-tolerant, late-averse, and late-neutral travelers in the entire population.

The effect of the heterogeneity of traveler compositions is not straightforward without a sensitivity analysis because heterogeneity affects costs with and without tolling, and the welfare benefits of tolling depends on the difference in costs. Moreover, neither the observable nor the omniscient pricing scheme eliminates queuing at the bottleneck during travel because arrival rates are random and there is a tradeoff between having high arrival rates which lead to queuing, and insufficient arrival rates which cause throughput to fall below the capacity of the bottleneck.

REFERENCES

Arnott, R., Palma (de), A., and Lindsey, R. (1994). The welfare effects of congestion tolls with heterogeneous commuters. *Journal of Transport Economics and Policy*, *28*(2): 139–161.

Arnott, R. J. and Kraus, M. (1994). When are anonymous congestion charges consistent with marginal cost pricing. *Journal of Public Economic*, *67*(1): 45–64.

Arnott, R. J., Palma (de), A., and Lindsey, C. R. (1990a). Departure time and route choice for the morning commute. *Transportation Research Part B:Methodological*, *24*(3): 209–228.

Arnott, R. J., Palma (de), A., and Lindsey, C. R. (1990b). Economics of a bottleneck. *Journal of Urban Economics*, *27*(1): 111–130.

Arnott, R. J., Palma (de), A., and Lindsey, C. R. (1998). Recent developments in the bottleneck model. In K.J. Button and E. T. Verhoef (Eds.), *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*. Cheltenham, UK: Edward Elgar Publishing, 79–110.

Ben-Akiva, M. E., Cyna, M., and Palma (de), A. (1984). Dynamic model of peak period congestion. *Transportation Research Part B:Methodological*, *18*(4): 339–355.

Cohen, Y. (1987). Commuter welfare under peak-period congestion tolls: Who gains and who loses? *International Journal of Transport Economics*, *14*(3): 238–266.

Daganzo, C. (1985). The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation Science*, *19*(1): 29–37.

Daniel, J. I. (1992). *Peak-load-congestion pricing and optimal capacity of large hub airports: With application to the Minneapolis-St Paul airport* (Doctoral dissertation). Department of Economics, University of Minnesota, Twin Cities, MI.

Daniel, J. I. (1995). Congestion pricing and capacity of large hub airport: A bottleneck model with stochastic queues. *Econometrica*, *63*(2): 327–370.

Daniel, J. I. (2001). Distributional consequences of airport congestion pricing. *Journal of Urban Economics*, *50*(2): 230–258.

Hansen, M. and Wenbei, W. (2001, November). *An airline's choice of aircraft size in a competitive environment*. Paper presented at INFORMS Conference, Miami, FL.

Hildebrand, M., Prentice, B., and Lipnowski, I. (1990). Enforcement of highway weight regulations: A game theoretic model. *Journal of the Transportation Research Forum*, *30*(2): 442–452.

Kita, H. (1999). A merging-giveaway interaction model of cars in a merging section: A game theoretic analysis. *Transportation Research Part A: Policy and Practice*, *33*(3/4): 305–312.

Levinson, M. D. (2005). Micro-foundations of congestion pricing: A game theory perspective. *Transportation Research Part A: Policy and Practice*, *39*(7–9): 691–704.

Littlechild, S. C. and Thompson, G. F. (1977). Aircraft landing fees: A game theory approach. *The Bell Journal of Economics*, *8*: 186–203.

Marcucci, E. and Marini, M. (2003). Political acceptability of road pricing policies under individual specific uncertainty. In J. Schade and B. Schflag (Eds.), *Acceptability of Transport Pricing Strategies*. Oxford: Elsevier Science, 279–297.

Smith, M. J. (1984). The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck. *Transportation Science*, *18*(4): 385–394.

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

Zou, X. and Levinson, D. (2006). A multi-agent congestion and pricing model. *Transportmetrica*, *2*(3): 237–249.

APPENDIX: ALGORITHM TO FIND THE EQUILIBRIUM

Under the proposed congestion and pricing model, a traveler's expected total personal cost (ETPC) is $\sum_{t=0}^{\infty} p_n(t|D_n)C_n(t|D_n)$ when no congestion fee is imposed, or $\sum_{t=0}^{\infty} p_n(t|D_n)(C_n(t|D_n) + F(t))$ with omniscient pricing, or $\sum_{t=0}^{\infty} p_n(t|D_n)(C_n(t|D_n) + \tilde{F}(t))$ with observable pricing. The equilibrium is obtained when no traveler can reduce his or her expected total personal cost by changing departure time unilaterally. This means, at equilibrium $(D_1^*, D_2^*..., D_i^*, \ldots, D_N^*)$,

1) Without pricing:

$$\begin{aligned}&\sum_{t=0}^{\infty} p_i(t|D_i^*)C_i(t|D_1^*, \ldots, D_i^*, \ldots, D_N^*) \leq \\ &\sum_{t=0}^{\infty} p_i(t|D'_i)C_i(t|D_1^*, \ldots, D_i', \ldots, D_N^*), \forall D_i{}', i = 1, 2, \ldots, N.\end{aligned} \tag{22}$$

2) With omniscient pricing:

$$\begin{aligned}&\sum_{t=0}^{\infty} p_i(t|D_i^*)(C_i(t|D_1^*,\ldots,D_i^*,\ldots,D_N^*)+F(t)) \leq \\ &\sum_{t=0}^{\infty} p_i(t|D_i')(C_i(t|D_1^*,\ldots,D_i',\ldots,D_N^*)+F(t)), \forall D_i', i=1,2,\ldots,N.\end{aligned} \tag{23}$$

3) With observable pricing:

$$\begin{aligned}&\sum_{t=0}^{\infty} p_i(t|D_i^*)(C_i(t|D_1^*,\ldots,D_i^*,\ldots,D_N^*)+\tilde{F}(t)) \leq \\ &\sum_{t=0}^{\infty} p_i(t|D'_i)(C_i(t|D_1^*,\ldots,D_i',\ldots,D_N^*)+\tilde{F}(t)), \forall D_i', i=1,2,\ldots,N.\end{aligned} \tag{24}$$

Under these equilibrium conditions, the heuristic to search the equilibrium is:
Step 0:

- Initialize departure time vector $D=(D_1, D_2 ..., D_i, \ldots, D_N)$;
- Compute the expected total personal cost vector ETPC$^{[0]}$ = (ETPC$_1$,ETPC$_2$, ETPC$_3$,...,ETPC$_N$) based on the initialized departure time vector; and
- Set $n=1$.

Step 1:

- For traveler *n,* keep D_i constant for$i=1, 2, \ldots, N$, $i \neq n$, search for D_N^* such that his expected total personal cost is minimized; and
- Set $D_n = D_N^*$.

Step 2:

- Set $n=n+1$; and
- If $n \leq N$ go to step 1; else go to step 3.

Step 3:

- Update the expected total personal cost vector based on the new departure time vector *D* resulting from Step 1 and Step 2; denote the updated new expected total personal cost vector as ETPC$^{[1]}$; and
- If ||ETPC$^{[1]}$ – ETPC$^{[0]}$ ||$_2 < \varepsilon$, stop; else set ETPC$^{[0]}$ = ETPC$^{[1]}$ and go to step 1. Here ε is a predefined tolerance.