JOURNAL OF ADVANCED TRANSPORTATION
J. Adv. Transp. 2014; **48**:575–587
Published online 17 September 2012 in Wiley Online Library (wileyonlinelibrary.com). **DOI**: 10.1002/atr.1210

Taxi services with search frictions and congestion externalities

Teng Yang[1], Hai Yang[1]* and Sze Chun Wong[2]

[1]*Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China*
[2]*Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China*

SUMMARY

Built upon the previous studies, this paper incorporates both bilateral taxi–customer search frictions and congestion externalities into the economic analyses of the equilibrium properties of taxi markets. We take account of congestion externalities by adopting a realistic distance-based and delay-based taxi fare structure. We first investigate comparative static effects of regulatory variables of taxi fare and fleet size on the market and then examine the properties of the Pareto-efficient solutions for simultaneous maximization of social welfare and taxi profit in the congested market. Copyright © 2012 John Wiley & Sons, Ltd.

KEY WORDS: taxi market; search friction; congestion externality; Pareto efficiency; market equilibrium

1. INTRODUCTION

Traditionally, economic analyses of taxi markets are made by economists under various types of regulation such as entry restriction and price control. Most studies have used aggregate models with the basic assumptions that the demand for taxi rides is a decreasing function of the expected fare and the expected customer waiting time; expected customer waiting time is decreasing with the total vacant taxi-hours; and the cost of operating a taxi is constant per hour [1–6].

In most large cities, taxis contribute significantly to traffic congestion. Unfortunately, most previous economic analyses of taxi services are generally based on a constant average taxi ride time or distance and ignored the effects of traffic congestion. The only economic and network analysis of taxi services with congestion externalities was made by Wong *et al*. [7] and Yang *et al*. [8], where a realistic taxi fare structure including both a distance-based charge and a delay-based charge is adopted in investigating the nature of equilibrium and regulation in the taxi market.

Recently, Yang *et al*. [9] used a meeting function to spell out the search and meeting frictions that arise endogenously as a result of the spatial feature of the area and the customer-taxi moving decisions on networks. They proved the existence of the stationary competitive equilibrium achieved at fixed fare prices when the demand of the customer matches the supply of taxis or there is market clearing at the prevailing searching and waiting times in every meeting location. Yang and Yang [10] further offered some new interesting insights into the properties of the equilibrium of taxi services considered by economists by explicitly taking into account the bilateral customer-taxi searching and meeting in an aggregate taxi market. They considered a general meeting function, in which the meeting rate increases faster (slower) than linearly with proportionate increases in the number of unserved customers and vacant taxis, thereby dictating increasing (decreasing) returns to scale of production of customer-taxi meetings.

*Correspondence to: Hai Yang, Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China. E-mail: cehyang@ust.hk

Built upon the work by Yang *et al.* [8] and Yang and Yang [10], this paper investigates the equilibrium properties of taxi markets by incorporating both bilateral taxi–customer search frictions and congestion externalities, which is clearly a more realistic representation of actual taxi markets. The paper is organized as follows. In the next section, we introduce the essential elements and the basic analytical model required to characterize taxi services with search frictions and congestion effects, and examine comparative static effects of regulatory variables of taxi fare and fleet size on the market. In Section 3, we examine the properties of the Pareto-efficient solutions for simultaneous maximization of social welfare and taxi profit. The situation with constant returns to scale in the meeting function is particularly considered for illustration. A numerical example is given in Section 4 to elaborate the analytical results. General conclusions are given in Section 5.

2. THE MODEL

2.1. Basic assumptions

We consider a one-hour modeling period of an aggregate market in a stationary equilibrium state as set out in [10]. Suppose each taxi takes only one customer; the rate of meetings between customers and taxis depends on the size of two pools: the number of unserved customers and the number of vacant taxis at a given instant. Clearly, the number of unserved customers any time, denoted by N^{c}, is given by $N^{\mathrm{c}} = w^{\mathrm{c}}Q$, where w^{c} and Q are respectively average customer wait/search time, and customer arrival rate or customer demand into the market per unit time (one hour). The number of vacant taxis, denoted by N^{vt}, is given by $N^{\mathrm{vt}} = w^{\mathrm{t}}T^{\mathrm{vt}}$, where w^{t} is the average taxi wait/search time expected for picking up each customer and T^{vt} is the arrival rate or supply of vacant taxis into the market per unit time. Thus, the meeting rate $m^{\mathrm{c-t}}$ between customers and taxis is given as a function of N^{c} and N^{vt} as follows:

$$m^{\mathrm{c-t}} = M(N^{\mathrm{c}},, N^{\mathrm{vt}}) = M(w^{\mathrm{c}}Q, w^{\mathrm{t}}T^{\mathrm{vt}}) \tag{1}$$

where $\partial m^{\mathrm{c-t}}/\partial N^{\mathrm{c}} > 0$ and $\partial m^{\mathrm{c-t}}/\partial N^{\mathrm{vt}} > 0$ in their domain $N^{\mathrm{c}} \geq 0$, $N^{\mathrm{vt}} \geq 0$. Furthermore, $m^{\mathrm{c-t}} \to 0$ as either $N^{\mathrm{c}} \to 0$ or $N^{\mathrm{vt}} \to 0$. Note that in a stationary equilibrium state with market clearing, we always have

$$Q \equiv T^{\mathrm{vt}} \equiv m^{\mathrm{c-t}} = M(w^{\mathrm{c}}Q, w^{\mathrm{t}}T^{\mathrm{vt}}) \tag{2}$$

In an aggregate market, the meeting rate between customers and taxis may increase faster (slower) than linearly with proportionate increases in the number of unserved customer and the number of vacant taxis, which reflects increasing (decreasing) returns to scale of the production of customer-taxi meetings. Let α_1 and α_2 be the elasticity of the meeting rate with respect to the number of unserved customers and the number of vacant taxis, respectively, at a given instant in time:

$$\alpha_1 \equiv \frac{\partial M}{\partial N^{\mathrm{c}}}\frac{N^{\mathrm{c}}}{M} \quad \left(= \frac{\partial M}{\partial N^{\mathrm{c}}}\frac{w^{\mathrm{c}}Q}{M} = w^{\mathrm{c}}\frac{\partial M}{\partial N^{\mathrm{c}}} \right) \tag{3}$$

$$\alpha_2 \equiv \frac{\partial M}{\partial N^{\mathrm{vt}}}\frac{N^{\mathrm{vt}}}{M} \quad \left(= \frac{\partial M}{\partial N^{\mathrm{vt}}}\frac{w^{\mathrm{t}}T^{\mathrm{vt}}}{M} = w^{\mathrm{t}}\frac{\partial M}{\partial N^{\mathrm{vt}}} \right) \tag{4}$$

\where the equalities in the brackets in Equations (3) and (4) hold in a stationary equilibrium state, and the second equality in the brackets stems from identity relation (2). It is generally expected that $0 < \alpha_1, \alpha_2 \leq 1$. The meeting function that is homogeneous of degree $\alpha_1 + \alpha_2$ is said to exhibit increasing, constant, or

DOI: 10.1002/atr

decreasing returns to scale if $\alpha_1 + \alpha_2 > 1$, $\alpha_1 + \alpha_2 = 1$, and $\alpha_1 + \alpha_2 < 1$, respectively.[a] In addition, we say that the bilateral meeting function is symmetric if $\alpha_1 \equiv \alpha_2$. Otherwise, it is asymmetric with an asymmetric factor α_2/α_1. The technical properties of the meeting function such as the returns to scale play an essential role in determining the many different aspects of market equilibrium [10].

As congestion externalities are now taken into account, we use a realistic taxi fare structure consisting of three components: a flag fall or a constant initial flat charge, a distance-based charge, and a delay-based charge [8]. Let P^0 denote the initial flag-fall charge per ride (dollars), β^{d} the fare charge per occupied unit distance (dollar per kilometer), β^{t} the fare charge per delay hour (dollar per hour), L the average taxi ride length (kilometer; assumed to be a given constant), T the average taxi ride time (hour; congestion dependent), and T^0 the average taxi ride time without congestion (hour; assumed to be a given constant).

The total trip fare P (dollars) is given by

$$P = P^0 + \beta^{\mathrm{d}} L + \beta^{\mathrm{t}} \left(T - T^0 \right) \tag{5}$$

where $T - T^0 \geq 0$. Let κ and τ denote the value of customer waiting time and in-vehicle ride time. Then the full trip price, denoted by μ, for a taxi ride is given by $\mu = P + \kappa w^{\mathrm{c}} + \tau T$. The demand function is assumed to be a strictly decreasing and differentiable function of the full trip price:

$$Q = f(\mu) = f(P + \kappa w^{\mathrm{c}} + \tau T) \tag{6}$$

where $f' = \mathrm{d}f/\mathrm{d}\mu < 0$ over $\mu > 0$.

The taxi utilization or occupancy rate U is defined as the portion of occupied service time and is given by

$$U = \frac{TQ}{N} \quad (0 < U < 1.0) \tag{7}$$

Note that, for an average taxi ride time T, the taxi fleet size N, number of vacant taxis N^{vt}, number of occupied taxis N^{o}, and customer demand Q are related to each other by

$$N^{\mathrm{o}} = QT \tag{8}$$

$$N^{\mathrm{vt}} = N - QT \tag{9}$$

With $U = TQ/N$, we have $N^{\mathrm{vt}} = (U^{-1} - 1)TQ$ from Equation (9). Substituting N^{vt} into Equation (1) together with $Q = m^{\mathrm{c-t}}$, we have

$$Q = M\left(w^{\mathrm{c}} Q, \left(\frac{1}{U} - 1 \right) TQ \right) \tag{10}$$

For given T, Equation (10) can be solved for waiting time w^{c} as a function of the customer demand Q and the taxi utilization rate U [10]. Namely,

[a] It is worth mentioning that the taxi market does in fact often have increasing returns to scale (Manski and Wright, 1976; Shroeter, 1983) [2], which is common in many queuing processes and public transit systems (Mohring, 1972). That is, a simultaneous increase in the arrival rate and the number of servers will, if not disparate, decrease the customer waiting time and increase the server utilization rate [10].

DOI: 10.1002/atr

$$w^{c} = W^{c}(Q, U) \tag{11}$$

It is also shown that the $W_U^c = \partial w^c / \partial U > 0$, but the sign of $W_Q^c = \partial w^c / \partial Q$ depends on the returns to scale $(\alpha_1 + \alpha_2)$ in the meeting function.

Naturally, the in-vehicle travel time for a given average length of taxi ride is given as a function of the vehicle density on the road. Here, our focus is on the impact of change in taxi fleet size on traffic congestion and hence taxi services. For simplicity, we ignore normal traffic, and the taxi ride time is a function of the number of vacant and occupied taxis only: [b]

$$T = t(N^{vt}, N^{o}) \tag{12}$$

Let t_1 and t_2 denote the partial derivatives of T in function (12) in N^{vt} and N^{o}, respectively. We have $t_1 > 0$ and $t_2 > 0$.

We now revisit taxi fare structure (5) with congestion effects. In view of $P^0 + \beta^d L$ in Equation (5) being constant for a given average ride distance L, we now use P^f to denote this fixed component of taxi fare: $P^f = P^0 + \beta^d L$. Therefore, in the presence of congestion externalities, the total taxi fare per ride P is a function of three components: fixed fare charge P^f, hourly delay charge β^t, and average taxi ride time T. Namely,

$$P = p\left(P^f, \beta^t, T\right) \tag{13}$$

Let p_1, p_2, and p_3 denote the partial derivatives of P in P^f, β^t, and T in function (13), respectively. We have $p_1 > 0$, $p_2 > 0$, and $p_3 > 0$.

2.2. *Comparative static effects of regulatory variables*

In the presence of congestion externalities, with the distance and delay-based taxi fare structure, we now have three market regulatory variables of constant fare charge P^f, delay-based charge rate β^t, and taxi fleet size N. These regulatory variables are regarded as independent variables.

For simplicity, we assume that both vacant and occupied taxis have about the same marginal impact on traffic flow (or running speed) in the considered cruising taxi market, namely, $t_1 = t_2$. From Yang *et al.* [8], we have the following results:

$$\frac{\partial P}{\partial P^f} = p_1 > 0 \tag{14}$$

$$\frac{\partial P}{\partial \beta^t} = p_2 > 0 \tag{15}$$

$$\frac{\partial P}{\partial N} = t_1 p_3 > 0 \tag{16}$$

[b] Normal trafffic is taken into account by assuming a fixed demand and the same moving speed as taxis subjec to endogenous congestion [10]. Nonethelerss, such a consideratoion in the current context makes the analysis analytically intractable.

DOI: 10.1002/atr

$$\frac{\partial T}{\partial P^{\mathrm{f}}} = \frac{\partial T}{\partial \beta^{\mathrm{t}}} = 0 \tag{17}$$

$$\frac{\partial T}{\partial N} = t_1 = t_2 > 0 \tag{18}$$

With a general bilateral searching and meeting function, which characterizes the search frictions between vacant taxis and unserved customers, we now look at the effects of the regulatory variables on customer waiting time. Taking the partial derivatives of w^{c} in (11) with respect to P^{f}, β^{t}, and N, respectively, and utilizing Equations (17) and (7), we obtain

$$\frac{\partial w^{\mathrm{c}}}{\partial P^{\mathrm{f}}} = W_Q^{\mathrm{c}} \frac{\partial Q}{\partial P^{\mathrm{f}}} + W_U^{\mathrm{c}} \left(\frac{T}{N} \frac{\partial Q}{\partial P^{\mathrm{f}}} + \frac{Q}{N} \frac{\partial T}{\partial P^{\mathrm{f}}} \right) = W_Q^{\mathrm{c}} \frac{\partial Q}{\partial P^{\mathrm{f}}} + W_U^{\mathrm{c}} \frac{T}{N} \frac{\partial Q}{\partial P^{\mathrm{f}}} \tag{19}$$

$$\frac{\partial w^{\mathrm{c}}}{\partial \beta^{\mathrm{t}}} = W_Q^{\mathrm{c}} \frac{\partial Q}{\partial \beta^{\mathrm{t}}} + W_U^{\mathrm{c}} \left(\frac{T}{N} \frac{\partial Q}{\partial \beta^{\mathrm{t}}} + \frac{Q}{N} \frac{\partial T}{\partial \beta^{\mathrm{t}}} \right) = W_Q^{\mathrm{c}} \frac{\partial Q}{\partial \beta^{\mathrm{t}}} + W_U^{\mathrm{c}} \frac{T}{N} \frac{\partial Q}{\partial \beta^{\mathrm{t}}} \tag{20}$$

$$\frac{\partial w^{\mathrm{c}}}{\partial N} = W_Q^{\mathrm{c}} \frac{\partial Q}{\partial N} + W_U^{\mathrm{c}} \frac{T}{N} \frac{\partial Q}{\partial N} + W_U^{\mathrm{c}} \frac{Q}{N} \frac{\partial T}{\partial N} - W_U^{\mathrm{c}} \frac{TQ}{N^2} \tag{21}$$

where $\partial Q/\partial P^{\mathrm{f}}$, $\partial Q/\partial \beta^{\mathrm{t}}$, and $\partial Q/\partial N$ are given as follows,

$$\frac{\partial Q}{\partial P^{\mathrm{f}}} = f' \frac{\partial P}{\partial P^{\mathrm{f}}} + \kappa f' \frac{\partial w^{\mathrm{c}}}{\partial P^{\mathrm{f}}} + \tau f' \frac{\partial T}{\partial P^{\mathrm{f}}} \tag{22}$$

Substituting Equations (14), (17), and (19) into (22), we can obtain

$$\frac{\partial Q}{\partial P^{\mathrm{f}}} = \frac{N f' p_1}{N - \kappa N f' W_Q^{\mathrm{c}} - \kappa T f' W_U^{\mathrm{c}}} \tag{23}$$

Similarly, we have

$$\frac{\partial Q}{\partial \beta^{\mathrm{t}}} = \frac{N f' p_2}{N - \kappa N f' W_Q^{\mathrm{c}} - \kappa T f' W_U^{\mathrm{c}}} \tag{24}$$

$$\frac{\partial Q}{\partial N} = -\frac{\kappa U f' W_U^{\mathrm{c}}}{N - \kappa N f' W_Q^{\mathrm{c}} - \kappa T f' W_U^{\mathrm{c}}} + \frac{N f' p_3 + \tau N f' + \kappa Q f' W_U^{\mathrm{c}}}{N - \kappa N f' W_Q^{\mathrm{c}} - \kappa T f' W_U^{\mathrm{c}}} \cdot \frac{\partial T}{\partial N} \tag{25}$$

According to Yang *et al.* [8], we have these results: $\partial Q/\partial P^{\mathrm{f}} < 0$ and $\partial Q/\partial \beta^{\mathrm{t}} < 0$ in Equations (23) and (24), respectively. However, Equation (25) describes the aggregate impact of taxi fleet size on customer demand in two opposite manners, which is different from [9,10], in which the congestion externalities are ignored. From [10], we can conclude that the first term of the right-hand side in Equation (25) is always positive, which means that if everything else is equal, increase in taxi fleet size will reduce customer waiting time and thus increases customer demand [8]. As $\partial T/\partial N > 0$,

$W^c_U = \partial w^c/\partial U > 0$, the second term of the right-hand side is always negative, which represents the negative impact of congestion on customer demand arising from the entry of one additional taxi into the market. From this equation, we thus conclude that customer demand is not always increasing with taxi fleet size because of the congestion effects.

3. PARETO-EFFICIENT SOLUTION

In this section, we consider the choice of taxi fleet size and fare structure to maximize social welfare S and total taxi profit Π simultaneously, or we seek Pareto-efficient solution of taxi services, which naturally gives rise to a bi-objective maximization problem.

Assuming a constant hourly operation cost c (dollar per hour) for both vacant and occupied taxis and a predetermined constant taxi fare charge, $P^f = P^0 + \beta^d L$,[c] then we have the following bi-objective maximization problem:

$$\max_{(\beta^t,N)\in\Omega} \begin{pmatrix} S(\beta^t, N) \\ \prod(\beta^t, N) \end{pmatrix} \tag{26}$$

where $\Omega = \{(\beta^t, N) : \beta^t \geq 0, N \geq 0\}$ as defined before. Social welfare is the sum of the customer surplus and the producer surplus. Different from [8], here we ignored the additional congestion delay cost to normal traffic arising from taxi movements for simplicity. Regarding Q, w^c, and T as implicit functions of (β^t, N), we have the social welfare given by

$$S(\beta^t, N) = \int_0^Q f^{-1}(z)\mathrm{d}z - \kappa Q w^c - \tau TQ - cN \tag{27}$$

The total taxi profit is given by

$$\prod(\beta^t, N) = PQ - cN = \left(P^f + \beta^t\cdot(T - T^0)\right)Q - c\cdot(N^{vt} + N^o) \tag{28}$$

We can apply the so-called ε-constraint method [11] to derive the Pareto-optimality conditions. Suppose that (β^{t*}, N^*) is a Pareto-optimal solution and (Q^*, T^*) denotes the corresponding customer demand and average taxi ride time, then we have the following nonlinear programming problem:

$$\max_{(\beta^t,N)\in\Omega} S(\beta^t, N) = \int_0^Q f^{-1}(z)\mathrm{d}z - \kappa Q w^c - \tau TQ - cN \tag{29}$$

subject to

$$\prod = \left(P^f + \beta^t\cdot(T - T^0)\right)Q - cN \geq \left(P^f + \beta^{t*}\cdot(T^* - T^0)\right)Q^* - cN^* \tag{30}$$

[c] Note that P^f is the corresponding taxi fare in the absence of congestion effect and thus is assumed to be already determined when the market does not exhibit congestion. Only the delay-based hourly charge rate is determined when congestion is built into the market. This would make a sensible comparison of the market cases with and without traffic congestion.

DOI: 10.1002/atr

Now we form the following Lagrange function:

$$L(\beta^{\mathrm{t}}, N) = \int_0^Q f^{-1}(z)\mathrm{d}z - \kappa Q w^{\mathrm{c}} - \tau T Q - cN \\ + \lambda\left[\left(P^{\mathrm{f}} + \beta^{\mathrm{t}}\cdot(T - T^0)\right)Q - cN - \left(P^{\mathrm{f}} + \beta^{\mathrm{t}*}\cdot(T^* - T^0)\right)Q^* - cN^*\right] \tag{31}$$

where $\lambda \geq 0$ is the Lagrange multiplier. Applying the first-order optimality conditions with respect to β^{t}, $\partial L/\partial \beta^{\mathrm{t}} = 0$, and utilizing Equations (20) and (24), we have

$$f' P = \kappa f' Q W_Q^{\mathrm{c}} + \kappa f' U W_U^{\mathrm{c}} - \frac{\lambda}{1+\lambda} Q \tag{32}$$

From $\partial L/\partial N = 0$ and substituting Equation (32) in the result, we have

$$\kappa W_U^{\mathrm{c}} \frac{U^2}{T} = c + \left(\tau Q + \kappa W_U^{\mathrm{c}} \frac{Q^2}{N}\right) \frac{\partial T}{\partial N} \tag{33}$$

Equation (33) is the general result of the Pareto-efficient solution with congestion externalities. The equation implies that if $\partial T/\partial N = 0$, namely, if there is no congestion effect, we have $\kappa U^2 W_U^{\mathrm{c}} = cT$. We thus arrive at the same results as in Yang and Yang [10], who concluded that when there are constant returns to scale in the meeting function, the customer waiting time will reduce to a function of taxi utilization rate alone, and the taxi utilization rate and customer waiting time are constant along the Pareto-efficient frontier. If the congestion effect cannot be ignored, the taxi utilization rate and customer waiting time are clearly not constant, even with constant returns to scale in the meeting functions.

Here, we show that, with constant returns to scale in the meeting functions, taxi utilization rate is no longer a constant along the Pareto-efficient frontier from social optimum (SO) to monopoly optimum (MO) in the presence of congestion externalities. For both increasing and decreasing returns to scale, we are unable to establish the results analytically. However, as demonstrated later through numerical example, similar results can be obtained.

Considering constant returns to scale, from [10], we have

$$W_U^{\mathrm{c}} = \frac{\alpha_2}{\alpha_1} \frac{w^{\mathrm{c}}}{(1-U)U} > 0 \tag{34}$$

The meeting function is said to exhibit constant returns to scale if $\alpha_1 + \alpha_2 = 1$. In this case, we always have $W_Q^{\mathrm{c}} = 0$ [10], and customer waiting time reduces to a function of taxi utilization rate alone as

$$w^{\mathrm{c}} = \zeta \left(\frac{U}{1-U}\right)^{\frac{\alpha_2}{\alpha_1}} \tag{35}$$

where ζ, $\zeta > 0$, is a constant that can be determined once the meeting function is specified explicitly. Substituting Equations (34) and (35) into (33), we can solve for the Pareto-efficient taxi utilization rate as follows:

$$U=\left(1+\left(\frac{\kappa\zeta\frac{\alpha_2}{\alpha_1}\left(1-\frac{N}{T}\frac{\partial T}{\partial N}\right)}{cT+\tau QT\frac{\partial T}{\partial N}}\right)^{\alpha_1}\right)^{-1}=\left(1+\left(\frac{\kappa\zeta\frac{\alpha_2}{\alpha_1}\left(1-e_N^T\right)}{cT+\tau QT\frac{\partial T}{\partial N}}\right)^{\alpha_1}\right)^{-1} \tag{36}$$

where $e_N^T=(N/T)(\partial T/\partial N)>0$ denotes the elasticity of the average taxi ride time with respect to the taxi fleet size.

From Equation (36), we can see that if $\partial T/\partial N=0$ or $e_N^T=0$ (without congestion effect), we can obtain the following constant taxi utilization rate as given in [10]:

$$U=\left(1+\left(\frac{\alpha_2}{\alpha_1}\frac{\kappa\zeta}{cT^0}\right)^{\alpha_1}\right)^{-1} \tag{37}$$

However, with congestion effects or $\partial T/\partial N>0$, we can readily see that, for a given number of taxi fleet size N on the Pareto-efficient frontier, the Pareto-efficient taxi utilization rate with congestion effects will be greater than the counterpart in the absence of congestion effect. In other words, for Pareto-efficient taxi services, traffic congestion tends to increase taxi utilization rate.

To look at how taxi utilization rate varies from SO to MO along the Pareto-efficient frontier with congestion effects, we simply consider a linear traffic flow model $v=a-bk$, where speed v is a linear function of traffic density k with parameter $a,b>0$; and traffic density is assumed to be given by $k=N/\bar{L}$, where $\bar{L}$ is the total road length in the network. With this assumption of linear traffic flow model and $T=L/v$ where L is the average length of taxi rides defined before, we can easily obtain

$$e_N^T=\left(\frac{a}{b}\frac{\bar{L}}{N}-1\right)^{-1} \tag{38}$$

Thus, e_N^T increases with taxi fleet size N. If we further assume that traffic is operating within the normal flow regime, or equivalently, $0<a/2\le v<a$, we can easily prove that $0<e_N^T\le 1$.

Yang and Yang [10] proved that taxi fleet size N and customer demand Q decrease along the Pareto-efficient frontier from SO to MO in the absence of congestion effect. One would expect that this is true as well in the presence of (mild) congestion effects, and therefore, from Equation (36), the Pareto-efficient taxi utilization rate decreases along the Pareto-efficient frontier from SO to MO when the meeting function exhibits constant returns to scale. This can be contrasted with the corresponding observation of constant taxi utilization rate in the absence of congestion externalities in [10]. We will further verify this observation by numerical examples.

4. A NUMERICAL EXAMPLE

In this section, we present a numerical example to illustrate the findings obtained so far. The demand function is specified to be of the following negative exponential form:

$$\begin{aligned}Q&=f(\mu)=f(P+\kappa w^{\mathrm{c}}+\tau T)\\&=\tilde{Q}\exp\left\{-\alpha\left(P^0+\beta^{\mathrm{d}}L+\beta^{\mathrm{t}}(T-T^0)+\kappa w^{\mathrm{c}}+\tau T\right)\right\}\end{aligned} \tag{39}$$

where $\tilde{Q}$ is the potential customer demand per hour and assumed to be $\tilde{Q}=1.0\times10^5$ (trip per hour). With reference to the calibrated data of Hong Kong [4], we take $L=30$ (kilometer), $\tau=35$ (HKD per hour), $\kappa=60$ (HKD per hour), and $\alpha=0.03$ (1/HKD). The hourly operation cost per taxi is assumed to be $c=50$ (HKD per hour). To characterize the congestion effects, we adopt a linear speed-density function: $v=a-bk$, where $k=N/\bar{L}$ and $\bar{L}$ (kilometer) is the total length of the roads in the network. Here we note again that we have simply ignored normal vehicle flow because our focus is on the

DOI: 10.1002/atr

impact of congestion on taxi services.[d] The parameter values are assumed to be $a = 100$ (kilometer per hour), $b = 0.67$ (square kilometer per hour), and $\bar{L} = 500$ (kilometer). Additionally, we have $T = L/v$ and $T^0 = L/a$. Therefore, the average running speed of vehicles on the network is given by $v = a - bN/\bar{L}$, which is a simplified version of Yang *et al.* [8] by assuming zero normal traffic demand.[e]

We first verify some observations from analysis of comparative static effects. Assuming a fixed component of taxi fare $P^f = P^0 + \beta^d L = 50$ (HKD) and a fare charge per one-hour delay $\beta^t = 84$ (HKD per hour), Figures 1 and 2 show the change of customer demand and customer waiting time with taxi fleet size. As observed in Figure 1, customer demand first increases with taxi fleet size and then decreases with further increase in taxi fleet size. This is clearly due to the congestion effect. When taxi fleet size is small or congestion is insignificant, increase in taxi fleet size reduces customer waiting time and thus induces more demand, but further excessive increase will bring about severe congestion and thus increases taxi ride time and fare and thereby reduces customer demand. Furthermore, from Figure 2, customer waiting time always decreases with taxi fleet size even in the presence of congestion effects regardless of the returns of scale of the bilateral meeting functions.

Figures 3–5 portray the iso-social welfare and iso-profit contours within the two-dimensional space of the taxi utilization rate and taxi fleet size. The Pareto-optimal solution sets together with the SO and MO solution are clearly identified and depicted in the figures. The corresponding representative numerical solutions with or without congestion externalities for the MO and SO are presented in Tables 1–3, respectively. From Tables 1–3, we can observe that the taxi fare with congestion increases, taxi fleet size with congestion decreases, and customer waiting time with congestion increases under both SO and MO solutions. More important, taxi utilization rate with congestion increases under both SO and MO solutions because of the congestion effects. Customer demand and social welfare all decrease under SO and MO solution if we take congestion effects into account. The social welfare decreases can be attributed mainly to both the increased journey time and the reduced customer surplus (reduced customer demand). It is worth to note that the taxi profits associated with the MO and SO exhibit an opposite trend of change with congestion effects in the cases of increasing and constant returns to scale, although it is not intuitively obvious whether taxi profits would increase or decrease with congestion effects, because taxi fare increases but customer demand decreases and taxi fleet size decreases at SO. Nevertheless, the changes of taxi profits at SO with and without congestion effects in the cases of increasing and constant returns to scale are consistent with the theoretical observations made in [[8,10]] and in the current study.

As found in previous studies [10], taxi services at SO are associated with a negative, zero, and positive profit under increasing, constant, and decreasing returns to scale, respectively; and thus taxi services should be subsidized in the case. This is indeed true in the absence of congestion externality. Nevertheless, with the congestion effects, the SO price charged per taxi ride exceeds the corresponding marginal cost. Taxi profit becomes positive even in the case of increasing return to scale. However, in spite of higher markup pricing over marginal cost in the monopoly market, the total monopoly profit decreases with the congestion effects because of reduced customer demand. These observations are also consistent with the findings in [8].

From Figure 3, with increasing returns to scale, taxi utilization rate decreases monotonically as moving from the SO to the MO along the Pareto-efficient contract curve (Pareto-efficient set), similar to that observed in [10]. Nonetheless, the change in taxi utilization from SO to MO is $0.7922 - 0.7034 = 0.0888$ with congestion, larger than that $0.7581 - 0.6879 = 0.0702$ without congestion.

From Figure 4, with constant returns to scale, contract curve is no longer a straight line, which contrasts with the constant taxi utilization rate (or straight line) without congestion externalities [10]. Consistent with our theoretical observation in Section [7], because of congestion effects, taxi utilization rate actually decreases monotonically as moving from the SO to the MO along the Pareto-efficient contract curve.

[d] It is worth mentioning that normal vehicle movements are not considered in our execution of numerical analysis. In reality, normal vehicle journey time also increases with degree of congestion and thus taxi fleet size, even if its demand is fixed. This means that our analysis actually underestimated the congestion externality of taxi services.

[e] Values of parameter selected in the example do not necessarily represent the Hong Kong situation but are simply for illustrative purpose.

DOI: 10.1002/atr

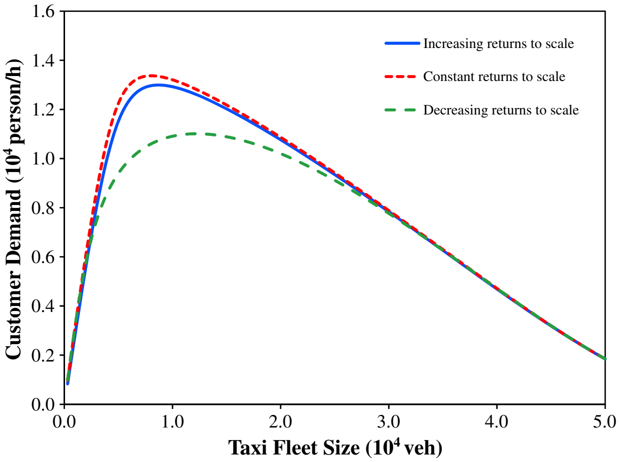

Figure 1. Customer demand versus taxi fleet size with constant and non-constant returns to scale.

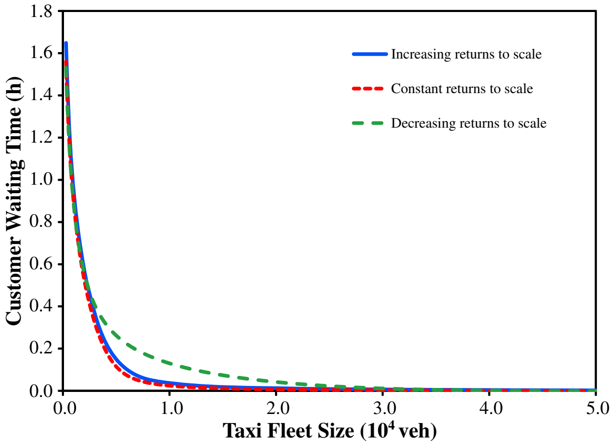

Figure 2. Customer waiting time versus taxi fleet size with constant and non-constant returns to scale.

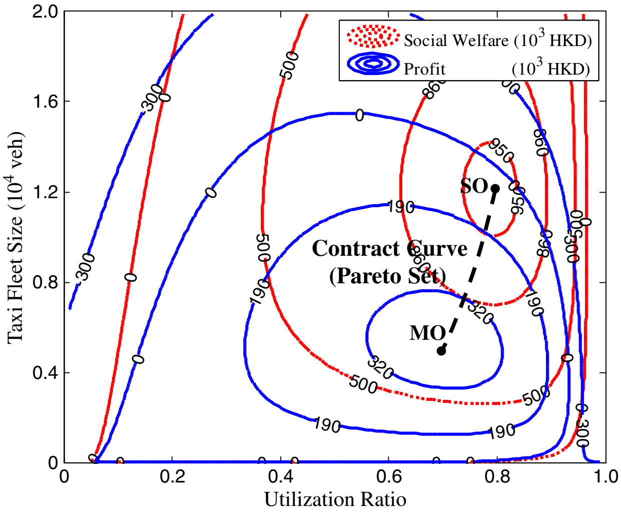

Figure 3. Pareto solution set under increasing returns to scale with congestion effects.

DOI: 10.1002/atr

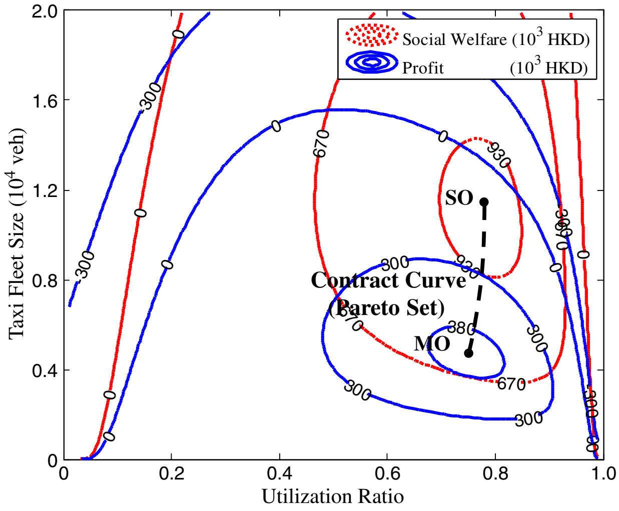

Figure 4. Pareto solution set under constant returns to scale with congestion effects.

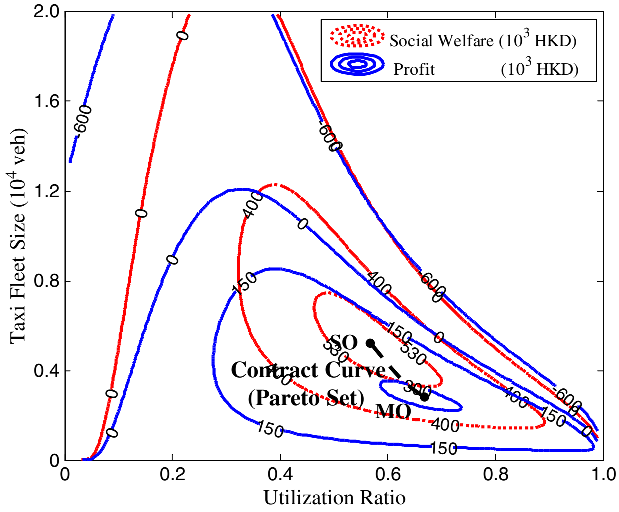

Figure 5. Pareto solution set under decreasing returns to scale with congestion effects.

Table I. The SO and MO solutions with increasing returns to scale.

Solution variable and performance measure	Monopoly optimum		Social optimum	
	Without congestion	With congestion	Without congestion	With congestion
Taxi fare (HKD)	50.6145	53.5017	16.1166	25.3955
Taxi fleet size (10^4 veh)	0.5685	0.5250	1.5490	1.2050
Customer waiting time (h)	0.1134	0.1240	0.0775	0.1020
Taxi utilization rate	0.6879	0.7034	0.7581	0.7922
Customer demand (10^4 person/h)	1.3035	1.1490	3.9143	2.6682
Taxi profit (10^5 HKD/h)	3.7550	3.5018	−1.4365	0.7510
Social welfare (10^6 HKD/h)	0.8100	0.7319	1.1611	0.9645

DOI: 10.1002/atr

Table II. The SO and MO solutions with constant returns to scale.

Solution variable and performance measure	Monopoly optimum		Social optimum	
	Without congestion	With congestion	Without congestion	With congestion
Taxi fare (HKD)	53.8245	56.4866	20.4776	28.4291
Taxi fleet size (10^4 veh)	0.5046	0.4700	1.3720	1.1000
Customer waiting time (h)	0.0912	0.0953	0.0913	0.1027
Taxi utilization rate	0.7324	0.7531	0.7325	0.7832
Customer demand (10^4 person/h)	1.2320	1.1055	3.3500	2.4484
Taxi profit (10^5 HKD/h)	4.1082	3.8948	0.0000	1.4607
Social welfare (10^6 HKD/h)	0.8214	0.7580	1.1167	0.9622

Table III. The SO and MO solutions with decreasing returns to scale.

Solution variable and performance measure	Monopoly optimum		Social optimum	
	Without congestion	With congestion	Without congestion	With congestion
Taxi fare (HKD)	73.4803	74.7191	54.4499	55.8908
Taxi fleet size (10^4 veh)	0.2924	0.2810	0.5407	0.5090
Customer waiting time (h)	0.1428	0.1393	0.2192	0.2144
Taxi utilization rate	0.6388	0.6615	0.5329	0.5734
Customer demand (10^4 person/h)	0.6226	0.5963	0.9604	0.9065
Taxi profit (10^5 HKD/h)	3.1130	3.0503	2.5260	2.5216
Social welfare (10^6 HKD/h)	0.5188	0.5038	0.5727	0.5543

For Figure 5, with decreasing returns to scale, taxi utilization rate increases monotonically as moving from the SO to the MO along the Pareto-efficient contract curve (Pareto-efficient set). Although this is still consistent with the case without congestion, the increase in taxi utilization from SO to MO is $0.6615 - 0.5734 = 0.0881$ with congestion in comparison with $0.6388 - 0.5329 = 0.1059$.

In summary, traffic congestion increases taxi utilization rate at Pareto-efficient solution inclusive of both SO and MO solutions. Meanwhile, traffic congestion also tends to make taxi utilization rate lower when moving from SO to the MO along the Pareto-efficient frontier, or it tends to rotate the Pareto-efficient contract curve in the clockwise direction.

5. CONCLUSIONS

In this paper, we investigated an aggregate taxi market by incorporating both bilateral taxi–customer search frictions and congestion externalities into the analyses. With a realistic distance-based and delay-based taxi fare structure, we first investigated comparative static effects of regulatory variables on the market. These market regulatory variables include constant fare charge, delay-based charge rate, and taxi fleet size. It is shown that customer demand is not always increasing with taxi fleet size if congestion effect is taken into account.

We examined the properties of the Pareto-efficient solutions for simultaneous maximization of social welfare and taxi profit in the congested market. In contrast with previous observations, we proved that taxi utilization rate decreases along the Pareto-efficient frontier from SO to MO with constant returns to scale, Pareto-efficient utilization rate with congestion effect is greater than its counterpart without congestion effect.

6. LIST OF SYMBOLS

N number of taxis in service or taxi fleet size
N^o number of occupied taxis anytime
N^{vt} number of vacant taxis anytime

DOI: 10.1002/atr

T^{vt}	arrival rate or supply of vacant taxis into the market per unit time (one hour).
Q	customer demand into the market per unit time (one hour).
N^c	number of unserved customers any time
w^c	average customer wait/search time
w^t	average taxi wait/search time expected for picking up each customer
m^{c-t}	meeting rate between customers and taxis per unit time (one hour)
P^0	initial flag-fall charge per ride (HKD)
β^{d}	fare charge per occupied unit distance (HKD/km)
β^{t}	fare charge per delay hour (HKD/h)
L	average taxi ride length (km) (assumed to be a constant)
T	average taxi ride time (h) (congestion-dependent)
T^0	average taxi ride time without congestion (h) (assumed to be a constant)
P^f	fixed component of taxi fare, $P^f = P^0 + \beta^d L$
c	constant hourly operation cost ($/h) for both occupied and vacant taxis
α_1	elasticity of meeting rate with respect to the number of unserved customers
α_2	elasticity of meeting rate with respect to the number of vacant taxis
κ	value of customer waiting time
τ	value of in-vehicle ride time.
μ	full trip price for a taxi ride, $\mu = P + \kappa w^c + \tau T$
U	taxi utilization or occupancy rate
S	total social welfare
Π	total taxi profit

ACKNOWLEDGEMENTS

The authors wish to express their thanks to three anonymous reviewers for their useful comments on an earlier version of the paper. The work described here was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (project no. HKUST 6212/07E).

REFERENCES

1. Douglas GW Price regulation and optimal service standards: the taxicab industry. *Journal of Transport Economics and Policy* 1972; **6**(2):116–127.
2. Arnott R. Taxi travel should be subsidized. *Journal of Urban Economics* 1996; **40**(3):316–333.
3. Cairns RD, Liston-Heyes C Competition and regulation in the taxi industry. *Journal of Public Economics* 1996; **59** (1):1–15.
4. Yang H, Wong SC, Wong KI. Modeling urban taxi services in road networks: progress, problem and prospect. *Journal of Advanced Transportation* 2001; **35**(3):237–258.
5. Yang H, Wong SC. A network model of urban taxi services. *Transportation Research* 1998; **32B**(4):235–246.
6. Yang H, Wong SC, Wong KI. Demand–supply equilibrium of taxi services in a network under competition and regulation. *Transportation Research* 2002; **36B**(9):799–819.
7. Wong KI, Wong SC, Yang H. Modeling urban taxi services in congested road networks with elastic demand. *Transportation Research* 2001; **35B**(9):819–842.
8. Yang H, Ye M, Tang WH, Wong SC. Regulating taxi services in the presence of congestion externalities. *Transportation Research* 2005; **39A**(1):17–40.
9. Yang H, Leung WY, Wong SC, Bell MGH. Equilibria of bilateral taxi–customer searching and meeting on networks. *Transportation Research* 2010; **44B**(8–9):819–842.
10. Yang H, Yang T. Equilibrium properties of taxi markets with search frictions. *Transportation Research* 2011; **45B**(4):696–713.
11. Geoffrion AM Solving bicriterion mathematical programs. *Operations Research* 1967; **15**(1):39–54.