Discussion Papers
Department of Economics
University of Copenhagen

No. 17-25

Vickrey Meets Alonso:
Commute Scheduling and Congestion in a Monocentric City

Mogens Fosgerau, Jinwon Kim and Abhishek Ranjan

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (E)

Vickrey Meets Alonso: Commute Scheduling and Congestion in a Monocentric City

Mogens Fosgerau[*] Jinwon Kim[†] Abhishek Ranjan[‡]

November 8, 2017

Abstract

This paper studies the interaction between dynamic traffic congestion and urban spatial equilibrium, using a model that is a straight unification of the Vickrey (1969) bottleneck congestion model and the Alonso (1964) monocentric city model. In a monocentric city with a bottleneck at the entrance to the CBD, residents choose their commute departure time jointly with residential location and housing consumption. Commuters arrive at the bottleneck in sequence sorted by residential location, so that more distant residents arrive later. The socially optimal toll makes central residents commute earlier in the morning than they would without the toll, which in turn induces a city that is less dense in the center and more dense further out. This is the opposite effect of what is found in models with static congestion.

Keywords: congestion; toll; land use; bottleneck model; monocentric model

JEL Classification Numbers: D62, R14, R41

[*] University of Copenhagen. Email: *mogens.fosgerau@econ.ku.dk*.

[†] California State University, Long Beach. Email: *jinwon.kim@csulb.edu*.

[‡] Technical University of Denmark. Email: *abhra@dtu.dk*.

We thank Gilles Duranton, Jonathan Hall, Ken Small, Erik Verhoef, and two anonymous referees for insightful comments and suggestions.

1 Introduction

This paper studies the interaction between dynamic traffic congestion and urban spatial equilibrium, formulating a model that is a straight unification of the Vickrey (1969) bottleneck model and the Alonso (1964) monocentric city model. It leads to conclusions that contradict the conclusions based on monocentric city models with static congestion, showing that it is indeed important to take congestion dynamics into account.

Our model is an extension of both the Vickrey and the Alonso models. The Vickrey (1969) bottleneck model, later for alized by Arnott, de Palma, Lindsey (1990, 1993), has become a workhorse for the economic analysis of traffic congestion. We extend this standard bottleneck model by incorporating spatial heterogeneity of commuters and endogenous urban equilibrium. At the same time, our model is an extension of the monocentric city model, developed by Alonso (1964), Muth (1967), and Mills (1972). We extend it by incorporating the consumer's commute scheduling problem and dynamic traffic congestion.

The timing of traffic plays no role in static congestion models, but takes centre stage in dynamic congestion models. They take into account that real traffic congestion is dynamic in the sense that traffic flows at some point in time affects later flows through the persistence of queues. To capture the dynamic nature of traffic congestion, Vickrey (1969) considered a bottleneck with a fixed capacity that commuters must pass to arrive at the destination. Commuters with identical scheduling preferences choose their optimal departure time from home. Then congestion arises in equilibrium since it is impossible for all commuters to pass the bottleneck at the same time. The queue can, however, be eliminated by the imposition of an appropriate time-varying toll.

The Vickrey bottleneck, however, model misses an important aspect of urban traffic congestion, namely space. In particular, the Vickrey formulation of scheduling preferences is additively separable in trip duration and arrival time and linear in trip duration. So, distance to the bottleneck does not matter for the Vickrey analysis of how travelers choose their arrival times at the bottleneck. In reality, however, trip distance may matter for a commuter's timing of trips. For example, a commuter located farther from the center may have to depart earlier from home to arrive at the destination at his preferred time than commuters at nearby locations, who travel shorter distances. Congestion dynamics at the bottleneck will then be influenced by the spatial distribution of commuters in the city.

It is only recently that research has recognized the importance of spatial heterogeneity of commuters in the congestion dynamics. Fosgerau and de Palma (2012) incorporate spatial heterogeneity of commuters in the bottleneck model and show how congestion dynamics at the bottleneck are influenced by the spatial distribution of population in the city. Arnott and DePalma (2011) consider a traffic corridor with dynamic flow congestion, that connects a continuum of residential locations to the central business district. Tsekeris and Geroliminis (2013) incorporate an empirical relationship between traffic flow and traffic density, which includes hypercongestion, to analyze how a city's road network and spatial structure influence congestion dynamics. Nevertheless, these models are incomplete because they take the population density to be exogenous.

We use the monocentric city framework developed by Alonso (1964), Muth (1969), and Mills (1972) to incorporate spatial heterogeneity of commuters into the bottleneck model. Specifically, we consider a monocentric city, where the entrance to the central business district is a bottleneck, which is the same spatial structure as in Fosgerau and de Palma (2012). Unlike in Fosgerau and de Palma (2012), however, the spatial distribution of commuters is endogenous in our model. To endogenize the urban spatial equilibrium, the consumer is assumed to not only choose the timing of commute trips but also to choose housing consumption and residential location. With the resulting endogeneity of population density, we obtain a firmer microeconomic foundation for commute scheduling equilibrium. From the monocentric model framework, we obtain the link between transport costs and commuters' location and housing consumptions choices, which is the key component generating the regularities of the urban spatial structure in the standard urban models. This framework also allows us to investigate how the city at laissez-faire equilibrium diverges from the socially optimal city, where congestion externalities are corrected by imposition of an appropriate congestion toll.

The urban economics literature has long been concerned with how the urban land market interacts with traffic congestion, with a particular aim of comparing laissez-faire and optimal land-use patterns (Solow and Vickrey, 1971; Riley, 1974; Arnott, 1979; Arnott, Pines and Sadka, 1986; Wheaton, 1998) or evaluating the efficacy of the first-best and alternative second-best anti-sprawl policies (Anas and Rhee, 2006; Brueckner, 2007). These classical congested-city models mostly adopt a static congestion framework, where congestion at a location depends just on the number of commuters passing, regardless of the timing. The interactions between traffic congestion and the spatial distribution of population in the city are incorporated, but these models nevertheless do not capture the dynamic nature of urban traffic congestion. Moreover, the commuter's scheduling problem is ignored in these models, which implies that the optimal congestion tolls are effectively imposed on residential location, not on the timing of commute trips. In contrast, the toll is based on the timing of trips in our model, which allows us to investigate the effect of the optimal time-varying congestion toll on the urban spatial structure.

There have been a few attempts similar to ours. Ross and Yinger (2000) were probably the first to incorporate the consumer's scheduling problem in the monocentric model framework. Unlike in our model, however, they consider flow congestion that depends both on population density at each location and departure times of residents. Unfortunately, their model is not very tractable and fails to generate a realistic equilibrium solution.[1] Gubins and Verhoef (2014) obtain a more realistic equilibrium solution by adopting the bottleneck model framework in the same spatial structure as ours. Like us, Gubins and Verhoef (2014) specify consumer preferences requiring that scheduling preferences are not separable from housing consumption preferences. But, unlike us, their specification of scheduling preferences is separable in trip duration and the time of arrival at work. To achieve non-separability between scheduling and housing preferences, they add a new term in utility that lets the marginal utility from housing consumption increase with time spent at home. In our model, we avoid such new elements, employing simply scheduling preferences that are not separable in trip duration and arrival time.

[1] The only solvable equilibrium in their model is a never-ending rush hour.

The present analysis first characterizes urban spatial structure and commute scheduling outcomes in laissez-faire equilibrium. We find that the spatial variation in housing prices and population density is qualitatively the same as in the standard monocentric model. We then find that travelers arrive at the bottleneck in sequence sorted according to trip distance, so that residents located farther away arrive at the bottleneck later than those located closer to the center. This sorting property links the congestion dynamics at the bottleneck to the spatial distribution of population in the city.

After analyzing laissez-faire, we investigate the social optimum, achieved by the imposition of an appropriate congestion toll on commuters. Toll revenues are not returned to commuters. The analysis shows that residents tend to arrive at the destination earlier in social optimum than under laissez-faire. As a result of the shift in arrival schedules, residents located at some distant locations attain a lower commuting cost in social optimum than under laissez-faire while the central residents incur a higher commuting cost. By changing the commuting cost at different locations in this way, the socially optimal toll pushes the central residents to the suburbs, generating a city that is less dense in the center and more dense further out compared to a city in laissez-faire equilibrium.

The effect of optimal tolling on urban spatial structure in our model differs from the effect found in two groups of previous studies. First, our result differs from that in the standard bottleneck setup considered in Arnott (1998), where tolling has no effect on location incentives due to the fact that it does not alter anyone's commuting cost. Second, our result is in stark contrast to that in the traditional congested-city models adopting the static congestion assumption (e.g., Wheaton (1998), Brueckner (2007)). Since the external congestion cost increases with commute distance in these models, the optimal toll tends to make the city more dense and compact.

The paper is organized as follows. In Section 2, we set up the model and investigate the urban spatial equilibrium. In Section 3, we investigate the trip scheduling equilibrium under the laissez-faire policy of no tolls. In Section 4, we consider the socially optimal toll and investigate the equilibrium under this policy. In Section 5, we illustrate the equilibrium numerically, with and without optimal tolling, and carry out some comparative static analyses using numerical illustration. Finally, Section 6 concludes. Proofs of the analytical results are given in the appendix.

2 The model

We begin by presenting the setup of the bottleneck model with spatial heterogeneity of commuters, which is the framework suggested by Fosgerau and de Palma (2012). We then incorporate the commuter's scheduling problem into the monocetric model and derive some basic results regarding the urban spatial equilibrium.

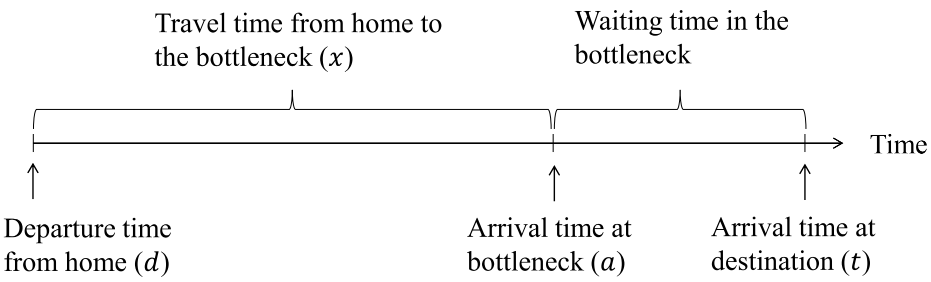
2.1 A bottleneck model with spatially heterogeneous commuters

The city is linear with a width of unity and contains a central business district (CBD) at its left end where all employment takes place. The city is *open* and its length is infinite, so there is no spatial boundary. The CBD has no physical extension. Commuters are city residents and live outside the CBD. The distance from a residence to the CBD is denoted x and is measured in time units. Travel speed is constant outside the CBD and hence x is also a measure of the spatial distance between the residence and the CBD.

Unlike in Fosgerau and de Palma (2012), the distribution of residents on distance is endogenous in our model. The number of residents within distance x is denoted $F(x)$ and the derivative $f(x) = F'(x)$ is the density of residents at distance x. The minimum distance of residents is x_0 and we let $x_0 = 0$ as the minimum distance does not impact our conclusions.

The entrance to the CBD is a bottleneck, which also has no physical extension. Every commuter must pass the bottleneck in order to enter the CBD. Commuters pass the bottleneck in the order of arrival. A commute schedule is described as follows. The commuter departs from home at time d and arrives first at the bottleneck at time a. There, he may experience some bottleneck delay before he arrives at his destination in the CBD at some later time t. A commuter located at distance x and departing at time

Figure 1: Commute schedule


Travel time from home to the bottleneck (x)
Waiting time in the bottleneck
Time
Departure time from home (d)
Arrival time at bottleneck (a)
Arrival time at destination (t)

d will arrive at the bottleneck at time $a = d + x$. The total commute time from home to the destination is the sum of travel time from home to the bottleneck, x, and the delay in the bottleneck, $t - a$. Figure 1 depicts the commute schedule.[2]

The bottleneck has a capacity of ψ persons per time unit. Congestion arises at the bottleneck when the rate at which travelers arrive at the bottleneck exceeds its capacity. To describe the congestion technology, denote by a_0 the time of the first arrival at the bottleneck and assume that arrivals take place at the time-varying rate $\rho(\cdot)$. The number of travelers who have arrived at the bottleneck location by time a is then given by $R(a) \equiv \int_{a_0}^{a} \rho(s)ds$. Denote by a_{q0} the most recent time before time a when there was no queue. Then $R(a) - R(a_{q0})$ commuters have joined the queue since time a_{q0}. They require $\left[R(a) - R(a_{q0})\right]/\psi$ time units to pass the bottleneck. The commuters exit the bottleneck at the rate ψ in the order that they arrived, which means that the individual who arrives at the bottleneck at time a arrives at the destination at time $t(a) = a_{q0} + \left[R(a) - R(a_{q0})\right]/\psi$. We may equivalently express this as $t(a) = a + \left[R(a) - R(a_{q0}) - \psi(a - a_{q0})\right]/\psi$, where the expression in brackets is the number of commuters left in the queue at time a.

Commuters have identical preferences regarding the timing of their commute, expressed by a cost function $c\,(d, t)$; we refer to it as "scheduling cost" or "commuting cost". The scheduling cost depends separately on both the clock time of departure, d, and the clock time of arrival, t. Therefore, it depends also on the travel time $t - d$, but

[2] The commute schedule in this model is different from that in the standard bottleneck model, where x is set to zero for all commuters, so that $d = a$, implying that the commuter arrives at the bottleneck immediately after the departure.

we will make assumptions that ensure that it does *not* reduce to a cost that depends only on the travel time $t - d$.

Following Fosgerau and de Palma (2012), we only impose weak assumptions on $c(d,t)$. By making the formulation of $c(d,t)$ as general as we can, we avoid the objection that our results are driven by very specific functional form assumptions. This is especially relevant in this paper, since our objection to the Arnott (1998) result that tolling has no effect on location incentives is that it derives specifically from his assumption regarding the cost function.

We assume that $c(d,t)$ is twice continuously differentiable with $c_1 < 0$ and $c_2 > 0$, meaning that commuters prefer to depart later and arrive earlier, ceteris paribus. We also assume that $c(d,t)$ is strictly convex and that $c(t-x,t)$ attains minimum for any (travel time) x. These minima are unique by convexity. Conditions 1 and 2 together ensure that $c(a-x,a)$ is convex as a function of a, so their combination is stronger than convexity.

Condition 1 $\forall\, d \leq t, c_{11}(d,t) + c_{12}(d,t) > 0$

Condition 2 $\forall\, d \leq t, c_{12}(d,t) + c_{22}(d,t) > 0$

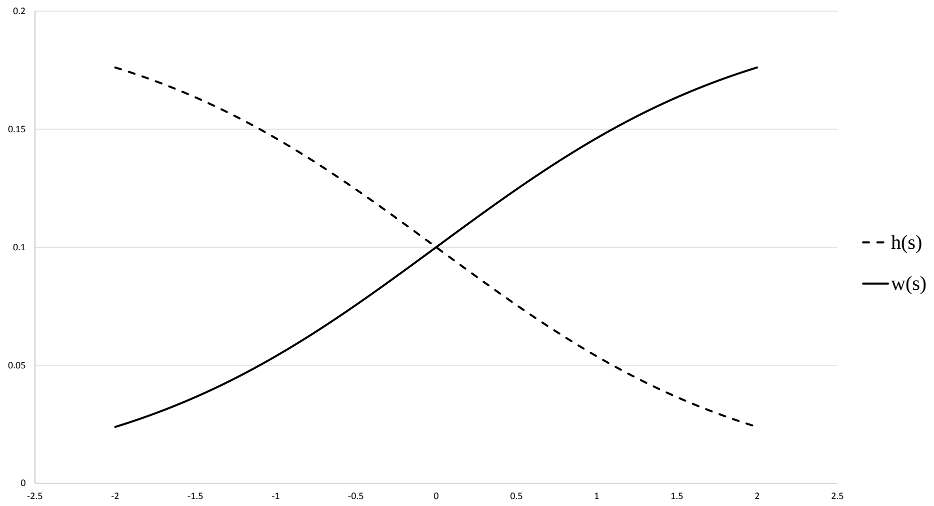
The final Condition 3 ensures that the substitution rate $-c_1(d,t)/c_2(d,t)$ decreases with d.[3]

Condition 3 $\forall\, d \leq t, c_{11}(d,t) - \frac{c_1(d,t)}{c_2(d,t)} c_{12}(d,t) > 0$

A special case is a scheduling cost defined in terms of utility rates at home and at work (Vickrey, 1973; Tseng and Verhoef, 2008; Fosgerau and Engelson, 2011). Denote by $h(s)$ the rate of utility achieved at time s at home prior to departure, and denote by $w(s)$ the utility rate achieved at work at time s after the completion of the commute. The utility rate during the trip is normalised to zero. Assume further that h, w are positive and differentiable with $h'(s) < 0 < w'(s)$ for all s, that $h(0) = w(0)$ and that w' is

[3] Fosgerau and de Palma (2012) assume also that $\forall\, d \leq t, c_{12}(d,t) - \frac{c_1(d,t)}{c_2(d,t)} c_{22}(d,t) > 0$. Under this condition, $-c_1(d,t)/c_2(d,t)$ decreases also with arrival time (t). Fosgerau and de Palma used this condition to prove the existence of the equilibrium.

Figure 2: Utility rates



continuous. These assumptions guarantee that the commuter prefers to be home prior to time 0 and prefers to be at work after time 0. Then we may define, as an example,

$$c\left(d,t\right) = -\int_{T_1}^{d} h\left(s\right) ds - \int_{t}^{T_2} w\left(s\right) ds, \tag{1}$$

where T_1, T_2 are arbitrary constants. Eq. (1) is then an example of a scheduling cost that satisfies all the above conditions and that, furthermore, is separable in departure and arrival time. In the general case, we do not assume separability and separability is not required for our results.

Our numerical example in Section 5.2 uses the cost formulation (1) as a special case of our general form, with $h\left(s\right) = ke^{-s}/\left(1+e^{-s}\right)$ and $w\left(s\right) = ke^{s}/\left(1+e^{s}\right)$, $k > 0$, leading to

$$c(d,t) = k\left(\ln\left[1+\exp(-d)\right] + \ln\left[1+\exp(t)\right]\right).$$

The utility rates h and w are illustrated in Figure 2. A commuter with fixed travel time would optimally depart at the (unique) time when the utility rates at home and at work are equal. A commuter with zero travel time would then depart and arrive at time 0. Commuters with positive travel time would depart prior to time 0 and arrive after time 0.

The conventional $\alpha - \beta - \gamma$ scheduling cost used in Vickrey (1969) and Small (1982) has the form $c\,(t, d) = \alpha(t - d) + \beta \max(0, t^* - t) + \gamma \max(0, t - t^*)$, where t^* is the preferred arrival time at the destination, α is the unit cost of travel time, β is the unit cost of arriving earlier than t^*, and γ is the unit cost of arriving later than t^*. It may be written in terms of utility rates with $h\,(s) = \alpha$ and $w\,(s) = (\alpha - \beta)$ if $s < t^*$ and $w\,(s) = \alpha + \gamma$ otherwise. The utility rate at work has a discontinuous jump and both utility rates are constant almost everywhere. The $\alpha - \beta - \gamma$ scheduling cost is then not a special case but a limiting case of the utility rate formulation (1) above.

The $\alpha - \beta - \gamma$ scheduling cost includes a single preferred arrival time. This is not the case for the scheduling cost (1) specified in terms of utility rates and even less for a general scheduling cost. Instead, the preferred departure and arrival times depend on the travel time.

Returning to a general scheduling cost, it is now convenient to write it in terms of the arrival time at the bottleneck, denoted by a. First, write the arrival time at the destination t as a function of the arrival time at the bottleneck a, so that $t = t(a)$. We have $t'(a) \geq 0$, since the queuing system obeys the first-in-first-out (FIFO) rule. In particular, when there is queue from time a_0 to a, $t(a) = a_0 + \frac{R(a)}{\psi}$ and $t'(a) = \frac{\rho(a)}{\psi} \geq 0$, as explained above. When there is no queue at time a, then $t\,(a) = a$. Since $d = a - x$, the departure time of the consumer living at distance x is also completely determined by the choice of arrival time at the bottleneck a. It is then possible to rewrite the scheduling cost in terms of a and x as $c\,(a - x, t\,(a))$.

2.2 Incorporating commute scheduling into the monocentric model

We now incorporate the consumer's commute scheduling problem into the monocentric model. Each resident in the city commutes to the CBD to earn income y. Consumer utility depends on housing consumption (equivalently land consumption) denoted by q, and on the consumption of a composite non-housing good denoted by e. Housing (land) rents are paid to absentee landlords, and the rental price per unit of land is p. We treat the scheduling cost, $c = c\,(a - x, t\,(a))$, as money metric and let it enter the budget constraint. With the price of the non-housing composite good normalized to unity, the budget constraint is then given by $e + c + pq = y$.

Consumers have strictly concave and three times continuously differentiable utility with negative definite hessian matrix. Elimination of e using the budget constraint allows utility to be written as $U\,(y - c - pq, q)$. The formulation that we use embodies an assumption of separability between e and $c(d, t)$ when q is given. This is a convenient simplification, but it is not strictly required: we could alternatively specify utility as depending on e, q, and c separately, where c would then not enter the budget constraint. We would then need to impose some appropriate restrictions on the second-order derivatives of utility and would expect to obtain qualitatively similar results compared to what we find with the present formulation.[4]

A consumer at location x chooses housing consumption q and arrival time a to maximize utility. The first-order condition for the choice of q is

$$-pU_1\left[y - c\,(a - x, t(a)) - pq, q\right] + U_2\left[y - c\,(a - x, t(a)) - pq, q\right] = 0. \tag{2}$$

The first-order condition for the choice of a is

$$c_1\,(a - x, t(a)) + c_2\,(a - x, t(a))\,t'(a) = 0, \tag{3}$$

which requires that the benefit from staying at home an extra unit of time equals the cost from arriving later at the destination. As (3) shows, the separability between non-housing consumption (e) and scheduling cost that we built into utility implies that, conditional on location x, the consumer's choice of departure time does not depend on his consumption. In contrast, we see from (2) that the consumer choice of housing consumption does depend on the scheduling cost.

Consumers are identical, so equilibrium also requires that consumers attain the same

[4] We just want to avoid the additional complexity that would be involved with this utility formulation. The standard bottleneck models also treat scheduling cost as money metric (e.g., Vickrey, 1969; Arnott, de Palma, and Lindsey, 1990, 1993; Fosgerau and de Palma, 2012).

utility level

$$U\left[y-c\left(a-x,t(a)\right)-pq,q\right]=\bar{U}, \tag{4}$$

such that no individual has incentive to change his residential location. The equilibrium utility, $\bar{U}$, is taken as an exogenous parameter since we are assuming an open city.[5]

We discuss the existence of equilibrium below, when we have established more properties of equilibrium. Granted existence, equations (2), (3), and (4) determine the equilibrium values for the key endogenous variables p, q, and a as functions of location (x). In particular, $p(x)$ is the bid-rent function for housing and $q(x)$ characterizes the spatial distribution of population (population densities) in the city. We compute the derivatives of these functions in this section. Spatial variations in a are a focus of the next section.

To derive the slope of the bid-rent function, differentiate (4) with respect to x using (2) and (3) to find that

$$p'(x)=\frac{c_1\left(a\left(x\right)-x,t(a\left(x\right))\right)}{q\left(x\right)}<0. \tag{5}$$

Thus, an increase in x leads to a utility-equalizing decline in p. Letting $c(x)$ denote the minimized scheduling cost at location x, $c'(x)=-c_1>0$ holds by the envelope theorem. This implies that the increasing scheduling cost associated with living further away from the CBD is compensated in equilibrium by a lower price of housing, the same principle as in the standard model. Note that (5) is the standard Alonso-Mills-Muth condition, except the marginal commuting cost here is not constant as in the standard model (see Brueckner (1987)).

Since utility is fixed, the increase in q with respect to x should exactly be the substitution effect of the decrease in p. This point is shown analytically by differentiating (2) with respect to x using (3) and (5), which yields

$$q'(x)=\eta p'(x)>0, \tag{6}$$

[5] We rely on the open-city assumption and the assumption that the city has no spatial boundary for analytical convenience. Our concern is on locations where residents experience congestion while we do not care much about what happens at the city boundary.

where $\eta \equiv U_1/\left(p^2U_{11} - 2pU_{12} + U_{22}\right)$. Substituting $p = U_2/U_1$ from (2) into η shows that $\eta = \partial MRS/\partial q|_{U=\bar{U}}^{-1}$, where $MRS \equiv U_2/U_1$. The convexity of indifference curves implies $\eta < 0$.

Housing production is suppressed in our model, and therefore the population density at x is simply the city width divided by housing consumption per resident $f(x) = 1/q(x)$. Note that $q'(x) > 0$ implies $f'(x) < 0$, i.e., that density decreases with increasing distance from the CBD.

Convexity of the indifference curves also implies that an increase in scheduling cost c for a resident at some location x increases his housing consumption in equilibrium.

These observations are summarized in the following proposition.

Proposition 1 *In equilibrium, the bid-rent curve slopes downward, i.e., $p'(x) < 0$, while individual housing consumption is increasing in x, i.e., $q'(x) > 0$, and population density falls with x, i.e., $f'(x) < 0$.*

3 Commute scheduling and congestion in laissez-faire equilibrium

In this section, we investigate the commuters' scheduling behavior and congestion dynamics under the laissez-faire policy of no tolls. In equilibrium, the commuter's scheduling cost is minimized given her other choices (see (3)), and the equal-utility condition (4) is also satisfied.

To begin with, it is useful to discuss the case where the capacity constraint at the bottleneck is not binding at any time during the day. In this case, there is no queue and every commuter arrives at the destination as soon as she arrives at the bottleneck, so that $t(a) = a$. The commuter located at distance x would then minimize $c(a - x, a)$ by choice of a. Let $a_*(x) \equiv argmin_a\ c(a - x, a)$ be the arrival time that would be chosen by a resident at x if there were no queue, expressed as a function of x. It is straightforward to show that Conditions 1 and 2 imply[6] that

$$0 < a_*'(x) < 1. \tag{7}$$

This means that commuters sort by distance to the CBD such that more distant commuters arrive later at the bottleneck. They also depart earlier since the derivative of the departure time $d(x)$ $(= a(x) - x)$ with respect to x is $a_*'(x) - 1 < 0$.

We now consider the general case where congestion may arise. The consumer located at distance x chooses arrival time at the bottleneck with the first-order condition (3) and the corresponding second-order condition written $SOC_a \geq 0$. Differentiating the first-order condition with respect to x shows (omitting some function arguments) that

$$SOC_a \cdot a'\left(x\right) = c_{11} - \frac{c_1}{c_2} c_{12}, \tag{8}$$

where $a\left(x\right)$ is the arrival time at the bottleneck of commuters located at x. But then

$$a'\left(x\right) > 0 \tag{9}$$

by Condition 3.

This argument relies only on the shape of scheduling cost and applies regardless of whether there is queue or not. So, no matter what the equilibrium looks like, more distant commuters arrive later at the bottleneck. We call this property as sorting in the sense that commuters sort in arrival time at the bottleneck. They do not similarly sort in departure time from home, since the sign of $d'(x)$ $(= a'(x) - 1)$ is ambiguous in this general case where congestion may arise.

It is generally possible that equilibrium population densities are so low that queuing does not emerge at all. As analysis of such a situation misses the point of this paper, we restrict attention to cases where maximum density in the city exceeds the bottleneck capacity. We have already established that the density decreases with distance from the

[6] The first-order condition for choice of a is $c_1(a - x, a) + c_2(a - x, a) = 0$. Differentiation of this equation with respect to x shows that $a_*' = (c_{11} + c_{12})/(c_{11} + 2c_{12} + c_{22})$.

bottleneck, so the condition amounts to $f(0) > \psi$. Proposition 2 below shows that this is sufficient to guarantee that there will be queuing at the bottleneck and that queuing begins at the time the commuter located at $x = 0$ arrives at the bottleneck.

Proposition 2 *In equilibrium, the first commuter arrives at the bottleneck earlier than she would prefer in the absence of queue, i.e.,* $a_0 \leq a_*(0)$*. In addition, when* $f(0) > \psi$*, queuing begins immediately after the first commuter arrives at the bottleneck.*

Sorting and the observation that $f' < 0$ implies that there can be just a single queuing interval. We denote the queuing interval by $[a_0, a_1]$. Due to sorting, the commuter located at 0 is the one to arrive at the bottleneck at time a_0 and we denote by x_1 the location of the commuter who arrives at the bottleneck at the time a_1 when the queue ends.

We look for results regarding the spatial variation in a during the queuing interval $[a_0, a_1]$. Fosgerau and de Palma (2012) (Theorem 3) impose conditions on their *exogenous* spatial distribution of the city population to ensure that all commuters queue in equilibrium. Their results regarding the spatial variation in a, conditioned on the spatial distribution of population, carry over to the present setting during the interval where the queue persists. These results are presented in the following proposition. Proposition 3 also comprises a statement regarding the more distant residents, who do not queue.

Proposition 3 *For* $x \in [0, x_1]$*,* $a(x)$ *satisfies the differential equation*

$$a'(x) = -\frac{c_2\left(a(x) - x, t(x)\right)}{c_1\left(a(x) - x, t(x)\right)} \frac{f(x)}{\psi} > 0, \tag{10}$$

where

$$t(x) = a_0 + \frac{F(x)}{\psi} \tag{11}$$

is the arrival time at the destination for residents located at x*. The equilibrium scheduling cost is thus given by*

$$c(x) = c\left(a(x) - x, a_0 + \frac{F(x)}{\psi}\right). \tag{12}$$

For $x > x_1$*, we have* $a\left(x\right) = a_*\left(x\right)$ *and* $t\left(x\right) = a\left(x\right)$*, since these residents do not queue. The arrival time function* $a(x)$ *is continuous for all* x*.*

Note again that arrivals at the bottleneck are sorted by distance, i.e., $a'(x) > 0$. Due to this sorting property, cumulative arrivals at time $a(x)$ equals cumulative residents at x, so that $R(a(x)) = F(x)$. This allows us to write the arrival time at the destination as a function of cumulative residents at x as seen in (11).[7]

By definition, the queue is exactly gone at time a_1 and this is the time when the resident at x_1 arrives at the bottleneck and at the destination. It follows that

$$a_1 = a_0 + \frac{F(x_1)}{\psi} = a_* (x_1) . \tag{13}$$

The following proposition states some additional features of this point.

Proposition 4 *At the location x_1 where queuing ends,*

$$a'(x_1) = t'(x_1)$$

and, for $x < x_1$ near x_1,

$$t (x) > a_* (x) .$$

Thus, the curves describing the arrival time at the bottleneck and the arrival time at the destination are tangent for the last resident to queue and residents closer to the CBD arrive later than they would have in the absence of queuing. A final piece of information regarding the shape of the queue comes from Fosgerau and de Palma (2012), whose result that the bottleneck delay, $a_0 + \frac{F(x)}{\psi} - a(x)$, is uni-modal as a function of distance x carries directly over to the current setting.

We are now ready to discuss the existence of equilibrium in general. We have used the condition that $f(0) > \psi$ to ensure that there is queuing in equilibrium. The existence of equilibrium is trivial without queuing, so we proceed with the case where there is queuing.

Consider a candidate a_0 for the arrival time at the bottleneck of the resident at location

[7] Since the capacity is fully utilized from time a_0 to $a(x)$, it takes $\frac{F(x)}{\psi} = \frac{R(a(x))}{\psi}$ time units for the precedent commuters to pass the bottleneck. So, the commuter at x arrives at the destination at time $a_0 + \frac{F(x)}{\psi}$.

0. By Proposition 2, we must have $a_0 \leq a_*(0)$. We also require the existence of initial values p_0, q_0 that solve (2) and (4) at $x = 0$ with $t(a_0) = a_0$.

With these initial values, the evolution of (q, p, a, t) as functions of x is uniquely determined by (6), (5), (10), and (11) with substitution of $f = \frac{1}{q}$. This is an application of the Picard-Lindelöf Theorem.[8] The differentiability requirements we made for U and c ensure that the conditions for using this theorem are fulfilled.

Let x_1 be the first $x > 0$ that solves $a(x) = t(x)$ and note that x_1 depends on a_0. It is now sufficient for an a_0 to lead to equilibrium that $a(x_1) = a_*(x_1)$. This is the case since all residents choose housing consumption and arrival time at the bottleneck according to their first-order conditions. The conditions that $a_0 \leq a_*(0)$ and $a(x_1) = t(x_1) = a_*(x_1)$ ensure that no resident wants to arrive at the bottleneck outside the interval $[a_0, a(x_1)]$.

Then equilibrium exists if such an a_0 exists and equilibrium is unique if there is just one such a_0. We do not have a proof that this is the case in general, and we therefore need to assume that it is the case. The assumption is true in our example below in Section 5.2.

4 Tolling and social optimum

In this section, we investigate the equilibrium under a time-varying toll charged at the entrance to the bottleneck. The toll revenue is not returned to commuters.

Commuters take the toll into account when choosing their commute schedule. We are treating scheduling cost as money metric and it enters utility together with non-housing consumption. The toll enters the budget constraint and then it becomes simply additive to the scheduling cost. We denote the toll at time a by $\tau(a)$ and we use subscripts τ to indicate functions and variables specific to the equilibrium with tolling.

Since the city is open, the utility of residents is fixed at some level, which implies that tolling makes no difference for the utility of residents. The welfare effect of tolling is then reflected in the toll revenue and the change in land rents.[9] We therefore use the following

[8] See e.g., Taylor (2011). The Picard-Lindelöf theorem is the standard result for guaranteeing existence and uniqueness of solutions to first order nonlinear differential equations.

[9] Another paper might use the current setup to analyse the provision of public goods or transport

measure of city welfare excluding the utility:

$$W \equiv \int_0^\infty p_\tau(x)dx + \psi \int_{a_{\tau 0}}^{a_{\tau 1}} \tau(a)da. \tag{14}$$

The city has width unity so the first term here is the total land rent, which is paid to absentee landlords. The toll is paid by ψ commuters per time unit during the interval $[a_{\tau 0}, a_{\tau 1}]$ so the second term in (14) is the total toll revenue. We assume that the total expenditure on the non-housing good equals its production cost, such that we may ignore consumption of the non-housing composite good in the welfare consideration. Then W is the total revenue that is extracted from the city and we seek a toll that maximizes this.

Note that under the open-city assumption, while the utility is exogenous, the entire population size is endogenous and thus changes in response to imposition of tolling. Then, setting a large toll will induce people to leave the city and reduce W. So, it makes intuitive sense that there exists a maximum of W in a reasonable set of tolls, as we will continue to establish and discuss.

Now, let Θ be the set of tolls that removes queue on the finite interval $[a_{\tau 0}, a_{\tau 1}]$ and that lead to equilibrium with the entry rate at the bottleneck at capacity when tolling is active and below capacity otherwise. We conjecture that the socially optimal toll belongs to this set, since there cannot be queueing in social optimum under bottleneck congestion. As in the laissez-faire case, we cannot prove in general that Θ is non-empty. It is however non-empty in the example that we provide in Section 5.2.

Proposition 5 lists a number of properties of the equilibrium under any toll in the set Θ:

Proposition 5 *Any toll $\tau \in \Theta$ has the following properties:*

1. *The evolution of the toll is given by*

$$\tau'\left(a_\tau(x)\right) = -c_1\left(a_\tau(x) - x, a_\tau(x)\right) - c_2\left(a_\tau(x) - x, a_\tau(x)\right). \tag{15}$$

infrastructure. That would connect to the self-financing literature on the Henry George Theorem (Arnott and Stiglitz, 1979) and self-financing roads (Mohring and Harwitz, 1962).

2. *Residents sort with $a'_\tau(x) > 0$.*

3. *The residential density is decreasing with distance: $f'_\tau(x) < 0$.*

4. *Tolling begins with the first traveler's arrival, i.e., $a_{\tau 0} = a_\tau(0)$, and there exists $x_{\tau 1}$ with $a_{\tau 1} = a_\tau(x_{\tau 1})$.*

5. *For $x \leq x_{\tau 1}$, we have $a_\tau(x) = a_{\tau 0} + \frac{F_\tau(x)}{\psi}$.*

6. *The last tolled commuter arrives at his preferred time: $a_{\tau 1} = a_*(x_{\tau 1})$.*

7. *The last commuter in the tolling interval pays zero toll: $\tau(a_{\tau 1}) = 0$.*

8. *The first commuter in the tolling interval pays zero toll: $\tau(a_{\tau 0}) = 0$.*

9. *The toll is increasing (decreasing) at arrival times for commuters who arrive before (after) their preferred time:*

$$\tau'(a_\tau(x)) \gtreqless 0 \Leftrightarrow a_\tau(x) \lesseqgtr a_*(x). \tag{16}$$

Now consider some toll $\tau \in \Theta$. We ask whether the toll is uniquely determined from $a_{\tau 1}$. Say that we know $a_{\tau 1}$. Then $x_{\tau 1} = a_*^{-1}(a_{\tau 1})$ is known, and given $\tau(a_{\tau 1}) = 0$, $f(x_{\tau 1})$ and $q(x_{\tau 1})$ are known. We know τ' from (15) and also $a'_\tau(x) = f_\tau(x)/\psi$ for $x \leq x_{\tau 1}$. From (5), we know $p'_\tau(x)$. Hence, we can back out the evolution of all quantities in the model and work backwards until we have determined $a_{\tau 0}$. Thus, from every $a_{\tau 1}$ corresponding to tolls in Θ, there is a unique toll in Θ and hence a unique value of $a_{\tau 0}$. Thus, we may consider $a_{\tau 0}$ as a function of $a_{\tau 1}$.

This function may however not be injective, in which case it will not have an inverse. We shall therefore require the following regularity condition to ensure that $a_{\tau 1}$ is increasing as a function of $a_{\tau 0}$ and vice versa. The condition is satisfied in our simulation example below.

Condition 4 *$a_{\tau 1}$, considered as a function of $a_{\tau 0}$, satisfies $\partial a_{\tau 1}/\partial a_{\tau 0} > 0$.*

The following proposition establishes then that welfare W is constant on Θ. This is trivially true if Θ consists of just one toll and we conjecture that this is the case.

Proposition 6 *Under condition 4, the welfare function W is constant on Θ.*

We are now ready to provide some results that compare social optimum to laissez-faire.

Proposition 7 *Assume Condition 4 and that $c_{12} = 0$.*

1. *The first traveler's arrival time is earlier in social optimum than under laissez-faire, i.e., $a_{\tau 0} \leq a_0$.*
2. *The scheduling cost is higher in social optimum than under laissez-faire for those located at $x = 0$, i.e., $c(a_{\tau 0} - 0, a_{\tau 0}) \geq c(a_0 - 0, a_0)$.*

It is important to discuss the first result that $a_{\tau 0} \leq a_0$. A central planner would ensure that departures are dispersed in time such that the arrival rate at the bottleneck equals the bottleneck capacity. This is exactly the same as in the basic Vickrey bottleneck analysis. But in contrast to the Vickrey analysis, the commuters here do not have identical scheduling preferences as these depend on the residential location. There is therefore also a scope for a central planner to increase welfare by changing the interval during which arrivals take place. Note that since the first arrival is earlier in social optimum than under laissez-faire, any commuter following the first commuter will also arrive earlier under social optimum than under laissez-faire (although this result is ultimately ambiguous due to our endogenous population density). This suggests that arrival times are overall too (inefficiently) late under laissez-faire.

To have intuition on this result, recall first that the central residents (either in optimum or in laissez-faire) arrive earlier than they would prefer in the absence of queue (see Proposition 2). But, since the central residents occupy the most advantageous location in terms of choosing the commute timing, they would want to delay departures as much as they could to get a time closer to $a_*(0)$. In contrast, the suburban residents who already arrive later than $a_*(x)$ tend to arrive *too* late under laissez-faire due to queuing. The optimal toll coordinates this asymmetry and induces earlier arrivals overall for the city residents.

In the standard bottleneck analysis, the optimal toll has no effect on the utility of commuters. The optimal toll simply replaces the cost of queueing with the cost of the toll. Then commuters are as well off as before, but the cost of congestion has been converted into toll revenue. As the next proposition shows, this is no longer the case when commuters are heterogeneous with respect to commute distance. We find that central residents lose from optimal tolling, which causes the residential density to decrease at central locations. Conversely, there are locations further out where the population density increases, reflecting that the residents there are made better off by the optimal toll.

Proposition 8 *The population density is lower under optimal tolling than under laissez-faire at locations near the bottleneck, i.e., $f_\tau(x) \leq f(x)$ for x near 0. Conversely, there are locations between 0 and x_1 where $f_\tau(x) \geq f(x)$.*

Since the scheduling cost (including the toll) is higher under tolling than under laissez-faire at locations near location 0 and since consumer utility is fixed at $\bar{U}$, the housing price as a compensating differential must be lower under tolling than under laissez-faire for residents at these locations. Since the price effect is comprised solely of a substitution effect (see (6)), the lower housing price in the center corresponds to a larger individual dwelling size and hence a lower population density, so that $f_\tau(x) \leq f(x)$ for x near 0.

This density effect of tolling is reversed at some non-central locations because residents there attain a lower scheduling cost (including the toll) in social optimum. As an example, it is useful to consider some suburban location between $x_{\tau 1}$ and x_1. Granted that $x_{\tau 1} < x_1$, residents at these locations under tolling arrive according to the schedule $a_*(x)$ without queuing or charge of toll (see Proposition 5). Tolling therefore clearly gives a benefit to these suburban residents. This benefit in terms of scheduling cost is compensated by a higher housing price, which results in a higher population density.

The density effect of tolling in our model is in stark contrasted to that in the congested-city models with the static congestion, where imposition of congestion tolling raises densities at all locations in the city, making the city more compact (e.g., Wheaton (1998), Brueckner (2007)). Their result is a consequence of the external congestion cost that is monotonically increasing with the commute distance. In the model of endogenous scheduling choices as ours, however, the suburban commuters traveling a longer distance do not necessarily generate a higher congestion externality.

Meanwhile, our result is similar to that in Gubins and Verhoef (2014), where dynamic congestion is considered in a spatial framework like ours. In the simulation in Gubins and Verhoef, however, tolling induces choices of larger housing and thus reduces densities from the CBD and out to some distance because tolling allows residents to spend more time at home by eliminating the queue, which gives a stronger incentive for residents to have a larger house. This result therefore crucially relies on their assumption that the marginal utility of spending time at home depends on the size of the house.

In our model, the city center is too dense because people are trying to locate near the bottleneck in order to improve their place in the (unpriced) queue. We can say that residents' preferences over location are distorted such that central locations are too favored over suburban locations. Importantly, this distortion is not simply assumed but is an implication of the general scheduling preferences used in our model.

5 Numerical examples

In this section, we illustrate the theoretical model by a numerical example. We also carry out a numerical comparative static analysis to see how the equilibrium configuration changes in response to a change in parameters. After investigating the laissez-faire equilibrium, we then find the social optimum, i.e., the equilibrium under the policy of optimal tolling, and investigate how the laissez-faire equilibrium diverges from the social optimum.

5.1 Outline of simulation

We run a simulation based on our model with Cobb-Douglas utility function of the form $U(e-c,q)=\ln[(e-c)q^{\gamma}]$. The functional form of scheduling cost is

$$c(d,t)=k\left(\ln\left[1+\exp(-d)\right]+\ln\left[1+\exp(t)\right]\right), \tag{17}$$

where k is a parameter that scales the commuting cost relative to expenditure. The consumer minimizes this expression by choice of a, which gives a as a function of x. Using the analytical results (10) and (11), the differential equation for $a(x)$ is written

$$a'(x)=\frac{1+\exp\left[a(x)-x\right]}{1+\exp\left[-\left(a_0+\frac{F(x)}{\psi}\right)\right]}\frac{f(x)}{\psi}. \tag{18}$$

The equilibrium scheduling cost is given by

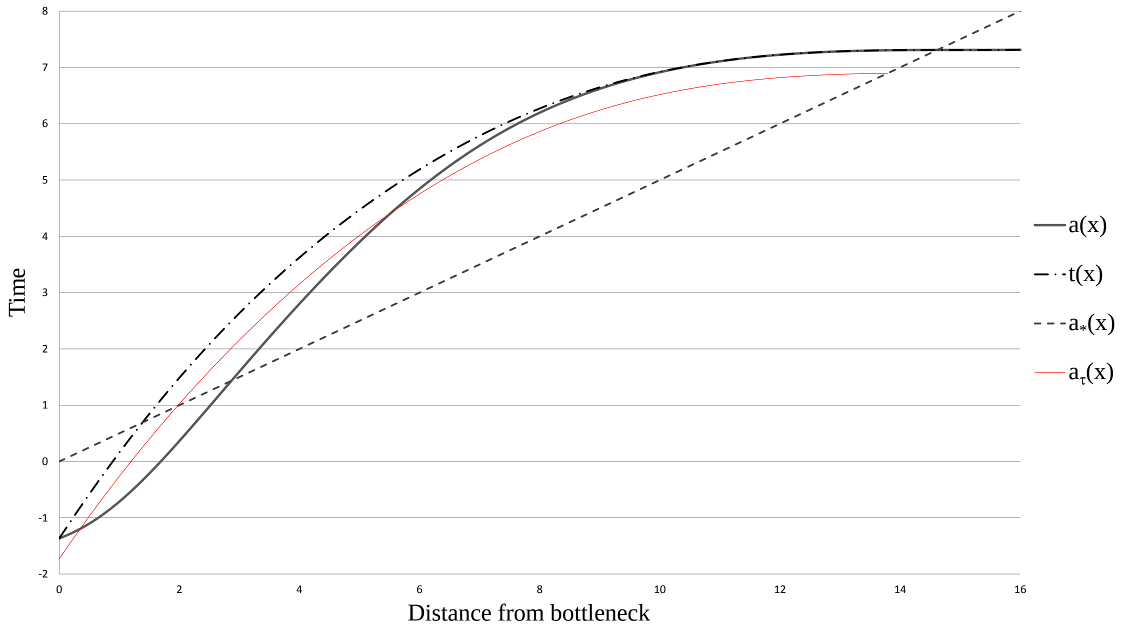
$$c(x)=\ln\left[1+\exp\left(-a(x)+x\right)\right]+\ln\left[1+\exp\left(a_0+\frac{F(x)}{\psi}\right)\right]. \tag{19}$$

This equation characterizes the interdependency between the equilibrium scheduling cost and the distribution of population in the city.

To find an equilibrium, we first arbitrarily set a value for a_0. Once a_0 is set, since $F(0)=0$, the value for $c(0)$ is determined from (19). $q(0)$ is also determined from the $q(x)$ solution for the utility maximization problem (see (2) and the budget constraint). Once these initial values are determined, the functions $a(x)$, $F(x)$, $c(x)$, and $q(x)$ can be computed sequentially using the differential equation in (18) for $a(x)$ and $F'(x)=f(x)=1/q(x)$. Among the set of arbitrary a_0 values, we search for the equilibrium value for a_0, such that the corresponding x_1 satisfies the condition (13). With the functional form assumed in (17), we find that $a_*(x)=x/2$.

The simulation relies on several exogenous parameter values, given as follows. The parametric utility ($\bar{U}$) is set at 0.78. The housing exponent in the Cobb-Douglas utility (γ) is set at 0.5. Income per household (y) are set at 3, and the unit cost of scheduling cost (k) is set at 0.2. The bottleneck capacity (ψ) is set at 0.4. These values are determined

Figure 3: Commute schedule variables as a function of location



to make the features of the model clearly visible on the figures.

5.2 Numerical results

5.2.1 Illustration of equilibrium

Figure 3 shows the equilibrium profiles for $a(x)$, $t(x)$ $(= a_0 + \frac{F(x)}{\psi})$, and $a_*(x)$ (recall that $a_*(x)$ is the optimal arrival time at the bottleneck in the absence of queueing for a commuter located at x). The point x_1 is the residential location where the three profiles meet. The numerical result indicates that the equilibrium value for x_1 is about 14.6. The $a(x)$ and $t(x)$ profiles are tangent at the point x_1 while the $a_*(x)$ profile crosses the other two curves from below, which is consistent with the theoretical result. The figure also shows that $a(0) < a_*(0)$, meaning that the central resident arrives earlier than the time he would choose if there were no queue. Residents in the middle part of the city arrive later than $a_*(x)$. The residents located beyond x_1 arrive according to the schedule $a_*(x)$. In the figure, the vertical gap between the $t(x)$ and the $a(x)$ curves at x indicates the waiting time for the resident at x. The queue begins immediately after the first arrival, and it persists until the person located at x_1 arrives at the bottleneck. The queue length initially increases with x, but it decreases and eventually vanishes at higher x values, which is also consistent with the theoretical prediction.

5.2.2 Numerical comparative static analyses

This section uses the simulation model to carry out a series of comparative static analyses. We compute the waiting-time profile, the waiting time at the bottleneck for residents at each location, and see how this profile shifts in response to a change in a parameter value. As another measure of congestion, we use the total bottleneck delay for all residents in the city, which is given by

$$\int_{a_0}^{a_1} \left(a_0 + \frac{R(s)}{\psi} - s \right) \rho(s) ds = \int_0^{x_1} \left(a_0 + \frac{F(x)}{\psi} - a(x) \right) f(x) dx. \tag{20}$$

The comparative static analysis also concerns the urban spatial structure.[10] In particular, we investigate how the population-density profile, which plots $f(x)$ $(= 1/q(x))$ against x, shifts in response to the changes in parameter values. Recall that the population density decreases with x. Since utility is fixed, q and p move in the opposite directions in response to a parameter change. So the downward-sloping bid-rent curve ($p(x)$ curve) shifts in the same direction as the density profile ($f(x)$ curve).

The exogenous parameters include income (y) and the bottleneck capacity (ψ). We also analyze the effect of changing the speed outside the bottleneck. The speed is incorporated by modifying the parameter x (time from residence to the bottleneck) into x/ω, where ω is a speed parameter. Specifically, travel takes x/ω time units from residence to the bottleneck, yielding the relationship $a = d + x/\omega$.[11] The ω parameter has been set at unity in the analytical section, but we now vary ω values to see its effect.

[10] The analytic comparative static signs on p and q and density are in many cases ambiguous. Unlike in the standard monocentric model, where the marginal commuting cost is exogenous, the scheduling cost as a function of location in our model depends on the distribution of population across other parts of the city (see (12)). So we cannot in general determine the sign of the impact of a parameter on the scheduling cost (c), and the impacts of a parameter on p and q are therefore ambiguous.

[11] With this modification, (18) and (19) become instead

$$a'(x) = \frac{1 + \exp\left[a(x) - \frac{x}{\omega}\right]}{1 + \exp\left[-\left(a_0 + \frac{F(x)}{\psi}\right)\right]} \frac{f(x)}{\psi}.$$

The comparative static exercises begin by using the initial values mentioned above as well as ω set at 1. In each case, we vary only the variable of interest, holding the other parameter values, which allows us to investigate the variable's partial effect.

Effects of an increase in incomes Figure 4 illustrates the effects of an increase in incomes in the city. Since the analytic comparative static derivations are complicated in our model, we mainly apply the interpretations of the standard open city analyses of Brueckner (1987) to our modified model. Since our model is an extension of the standard urban model, this approach would provide a good starting point.

Panel (a) in the figure shows that population densities increase throughout the city as income increases. As in Brueckner (1987), although utility ($\bar{U}$) in the open city is ultimately fixed, it is useful to decompose the impact of y into an initial rise in $\bar{U}$ and the adjustment in the other variables that would follow to offset the initial rise in $\bar{U}$. In particular, p would rise in response to the increase in $\bar{U}$ due to the rising demand for housing, which would be spurred by the inflow of migration. With the higher p, individual dwelling sizes will decrease, leading to a higher population density, with an increase in the overall population. The adjustment would continue until the original level of utility is restored.

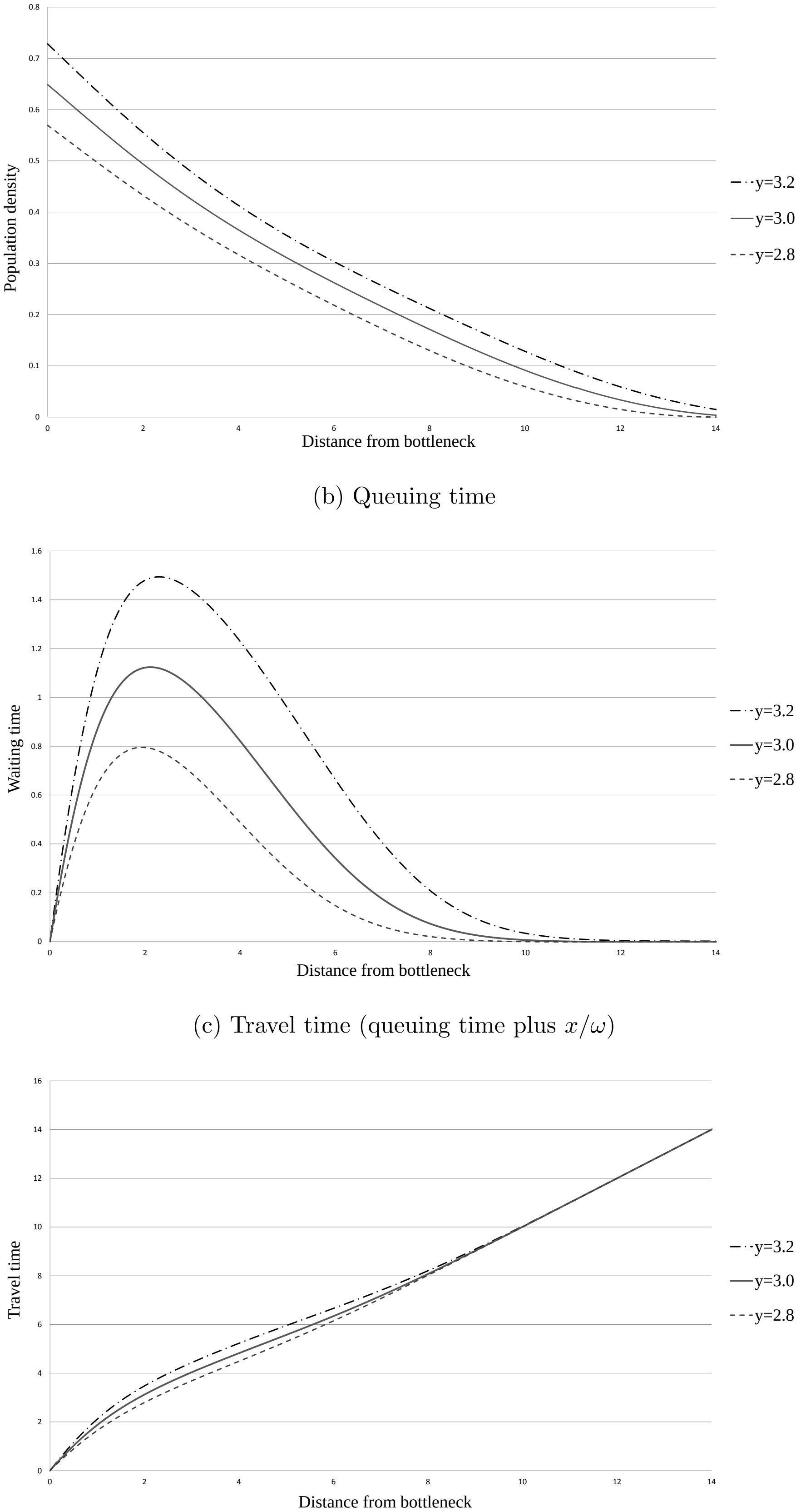
Since congestion is endogenous in the current model, not only p but also the congestion level would increase to offset the initial impact of y on $\bar{U}$. The mechanism would be through the higher population density leading to higher traffic flows and more congestion delays. Panel (b) in Figure 4 does indicate that the waiting time for all residents increases as incomes increase. The numerical example indicates that the aggregate bottleneck delays (for all individuals) increase by 62% when y increases from 3 to 3.2, which corresponds approximately to a point elasticity of 9.3. Finally, Panel (c) shows that the individual commuter's total travel time (i.e., x/ω plus the queuing time) is non-linear

and

$$c(x) = ln\left[1 + \exp\left(-a(x) + \frac{x}{\omega}\right)\right] + ln\left[1 + \exp\left(a_0 + \frac{F(x)}{\psi}\right)\right].$$

The arrival rate in the absence of queue is modified into $a'_*(x) = x/2\omega$.

Figure 4: Effects of income change



(a) Population density

(b) Queuing time

(c) Travel time (queuing time plus x/ω)

due to the non-linear queuing time, but it is still monotonic, i.e., commuters living farther out travel a longer time. We conclude that the higher-income city will be denser, more expensive to live in, and more congested than a city with lower incomes, which is consistent with real-world observations.

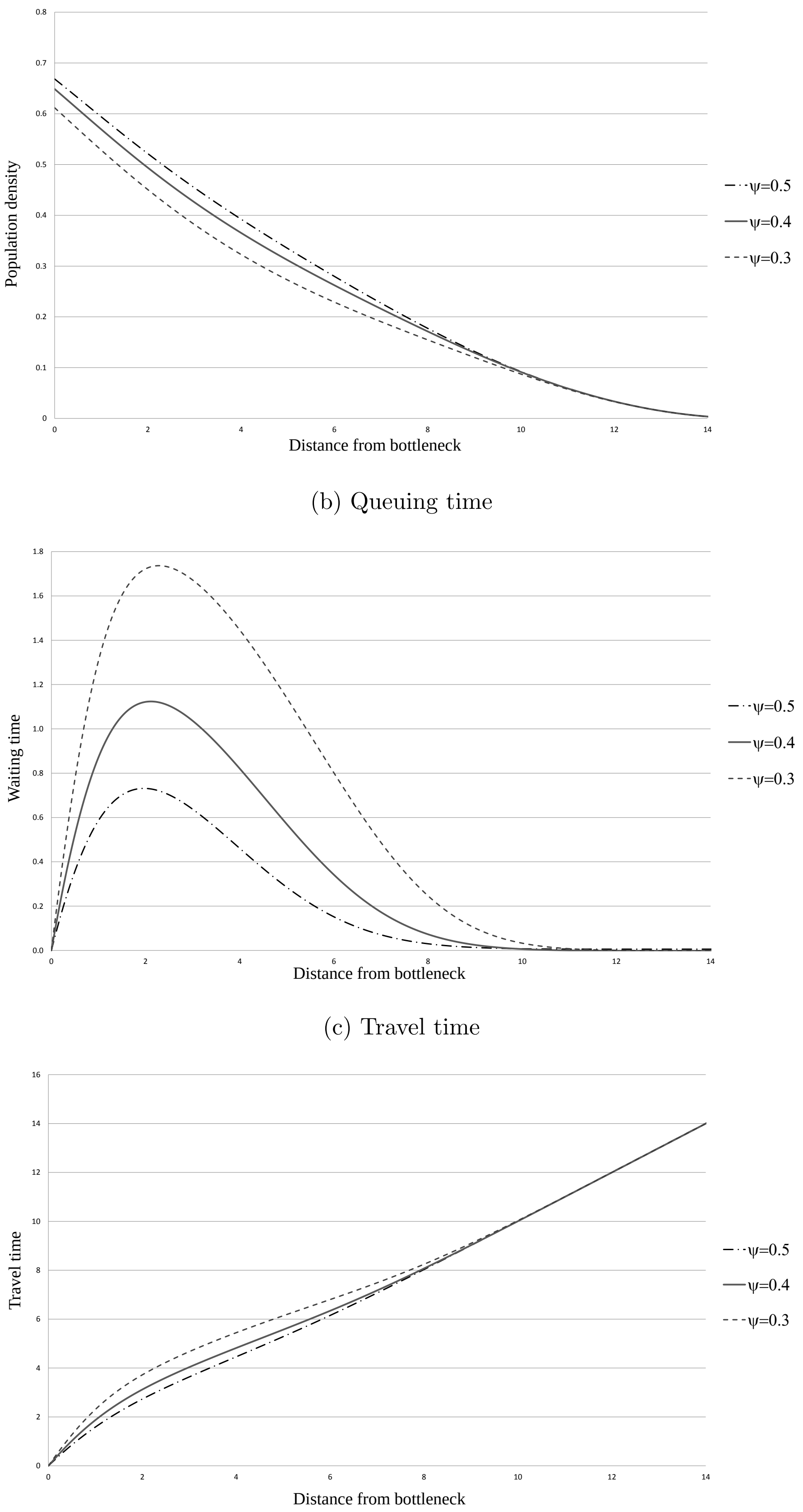
Effects of an increase in the bottleneck capacity Figure 4 numerically illustrates the effects of an increase in the bottleneck capacity. It first shows that population densities increase as the bottleneck capacity increases. The impact of ψ on densities is parallel with that of y because it would also cause an initial rise in utility ($\bar{U}$) by reducing the individual's commuting cost. In response to the initial increase in utility, p will increase, leading to smaller dwelling sizes and higher population densities, and the adjustment will continue until the original utility level is restored. The only difference from the effect of y is that the bottleneck capacity has no influence on densities at distant locations (beyond x_1), since the waiting time is unaffected by ψ for the residents at these locations.

For a given level of population and densities, a higher capacity would directly reduce the waiting time at the bottleneck. However, since the population would also increase, the effect of ψ on congestion is analytically ambiguous. According to our numerical simulation, despite the increased population, higher capacity leads to reductions in the queuing times (see Panel (b)). The aggregate bottleneck delays fall by 35% when ψ increases by 25% (from 0.4 to 0.5), which is equivalent to a point elasticity of -1.4. Finally, Panel (c) illustrates the individual commuter's travel time (x/ω plus queuing time) as a function of location, which again exhibits a non-linear but monotonic feature. We conclude that the city with a higher bottleneck capacity is more expensive to live in and more dense but less congested than in a city with a lower bottleneck capacity.

Effects of an increase in the speed outside of the bottleneck (ω) An increase in the speed parameter ω reduces the congestion-free travel time from home to the bottleneck (x/ω). Similarly to the effect of ψ, the increase in ω is associated with a lower commuting cost for a given distribution of population. To offset the resulting positive effect on utility ($\bar{U}$), p and population densities would increase along with the overall population size.

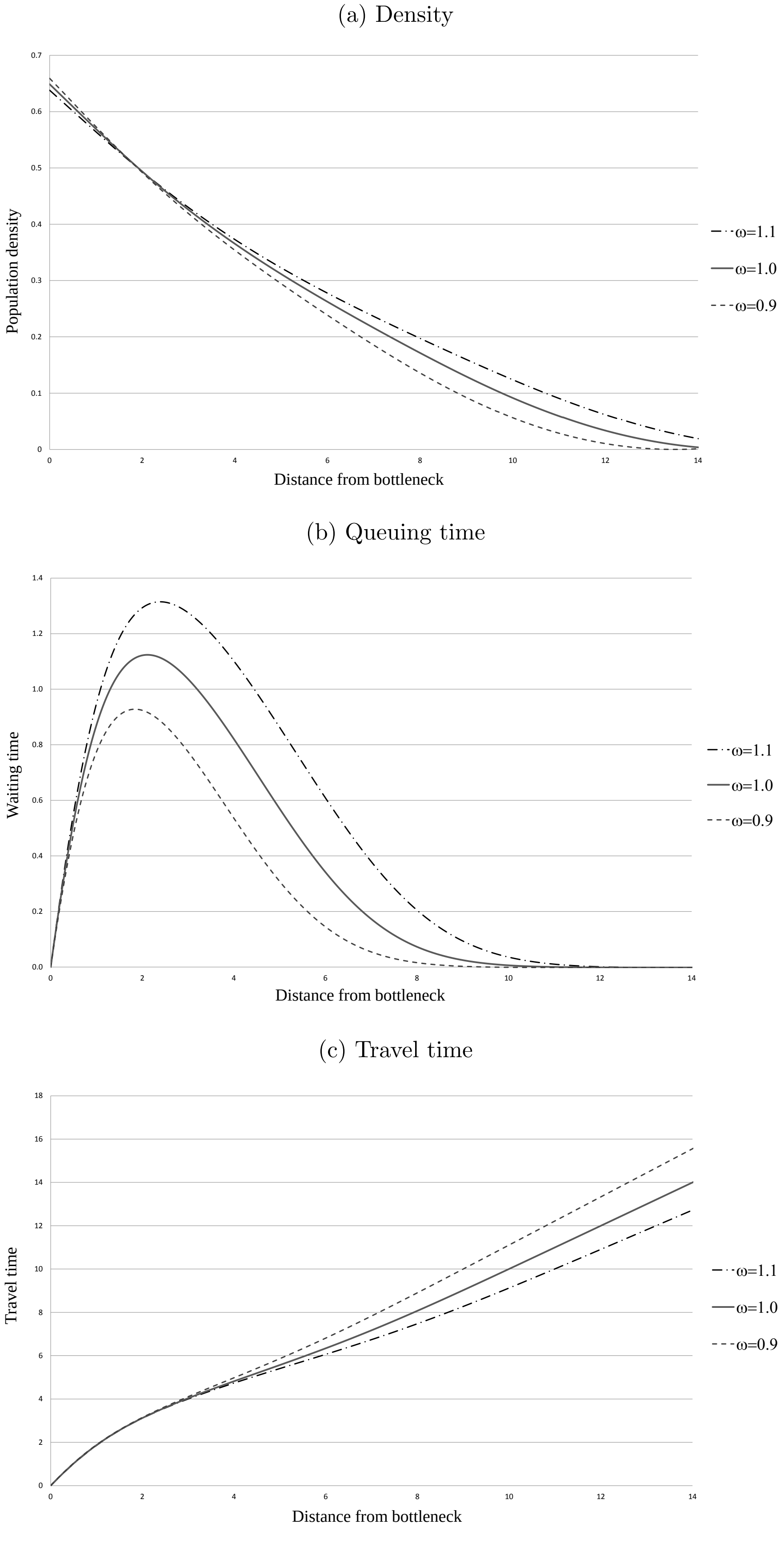
Figure 5: Effects of bottleneck capacity change

(a) Population density


(b) Queuing time

(c) Travel time

Figure 6: Effects of speed change


(a) Density

(b) Queuing time

(c) Travel time

However, the effect of ω is not necessarily constant over locations within city. Since the reduction in commuting cost (in response to the increase in ω) would be larger as x is higher (since the decrease in x/ω is larger as x is larger), the suburban locations now become more attractive relative to the central locations. As a result, the commuters who were located in the central locations would shift toward the suburb, lowering the densities at the center and raising them outward. Summing up the two effects (i.e., the effect of the increased population and the effect of migrations within the city), the effect on densities at the center would analytically be ambiguous. Our numerical example, illustrated in Panel (a) in Figure 6, shows that the increase in ω lowers densities in the center and increase them outside.

With the (overall) higher population densities and the resulting increased traffic, each individual's waiting time increase as ω increases. Figure 6 confirms that this prediction is right. Because the population size also increases, the aggregate bottleneck delays must unambiguously increase. The numerical result indicates that the aggregate bottleneck delays increase by 28% when ψ increases by 10% (from 1 to 1.1). Finally, Panel (c) show that although the queuing time increase as the speed increases (due to the effect of the higher population), the individual's travel time (x/ω plus queuing time) nevertheless tend to fall because the increased ω directly lowers the congestion-free travel time (see x/ω), especially at distant locations with a greater x. We conclude that as the speed outside of the bottleneck increases, the city becomes more expensive, more dense, and more congested.

5.2.3 Tolling and social optimum

Finding the social optimum The procedure for finding the social optimum is as follows. The first step is to arbitrarily select $x_{\tau 1}$, such that tolls are imposed for all residents located inside $x_{\tau 1}$ but not for those located beyond $x_{\tau 1}$.

Using the boundary condition, $a_{\tau 1} = a_*(x_{\tau 1})$, and given $a_*(x) = x/2$, we can then determine the value of $a_{\tau 1}$ and also $c(x_{\tau 1})$ using (19). Given $\tau(a_{\tau 1})$ set at zero, $q_\tau(x_{\tau 1})$ (and also $f_\tau(x_{\tau 1})$) is determined from the $q_\tau(x)$ solution for the utility-maximization problem (see (2) and the budget constraint). Once these boundary values are determined, we then use the differential equation $a'_\tau(x) = f_\tau(x)/\psi$ to determine the values for $a_\tau(x)$ for $x < x_{\tau 1}$ until the point $x = 0$. At the same time, we determine $\tau(a_\tau(x))$ for $x < x_{\tau 1}$ by the differential equation (15), that works backward in time from the point $x_{\tau 1}$.[12] The corresponding values for $q_\tau(x)$ and $f_\tau(x)$ are also computed using the utility-maximization conditions for $q_\tau(x)$. Note that the arbitrary chosen $x_{\tau 1}$ is not necessarily the value corresponding to the social optimum. So, to find the social optimum, we find sets of equilibrium values with varying $x_{\tau 1}$ values and compute the corresponding values for W (see (14)). We then pick the $x_{\tau 1}$ value that gives the maximum of W.[13] As shown in Propositions 5 and 6, the welfare-maximizing $x_{\tau 1}$ is the value that leads to $\tau(a_{\tau 0}) = 0$ at location $x = 0$.

Comparison between social optimum and laissez-faire The thin solid (red) line in Figure 3 plots the arrival time at the bottleneck (and at the destination) in social optimum at each x. The numerical result confirms that the central resident arrives earlier in social optimum than under laissez-faire, i.e., $a_\tau(0) < a(0)$. We can also see that all residents located inside x_1 arrive at the destination earlier in social optimum than under laissez-faire.

Figure 7 plots the optimal toll that is imposed on each resident at x. The toll at each end point is zero, i.e., $\tau(a_\tau(0)) = \tau(a_{\tau 1}) = 0$, which confirms the results of Proposition 5. We can also see that residents meeting a longer queue are charged a higher congestion toll, which is consistent with the standard externality-based logic for congestion tolling.

Figure 8 depicts scheduling costs as a function of location both at the laissez-faire equilibrium and in social optimum. The solid line plots the scheduling cost of an individual located at each x under laissez-faire. The dotted line plots $c_\tau(x) + \tau(a_\tau(x))$,

[12] Under the maintained functional forms, the first-order condition for a_τ corresponding to (15) is given by

$$\tau'(a_\tau) \quad = \quad k\left(\frac{\exp\left(-a_\tau + x\right)}{1 + \exp\left(-a_\tau + x\right)} - \frac{\exp\left(a_\tau\right)}{1 + \exp\left(a_\tau\right)}\right).$$

[13] We compute land rents until the point $x = 5.4$ since land rents beyond x_q are not affected by tolling. The boundary value of 5.4 would correspond to an exogenous agricultural rent of 0.0004. Equivalently (and more simply), we could search for $x_{\tau 1}$ until the equilibrium satisfies $\tau(a_\tau(x_0)) = 0$.

i.e., scheduling costs in social optimum, which includes the scheduling cost itself and the toll. The numerical result confirms that the scheduling cost at $x = 0$ is higher in social optimum than under laissez-faire, i.e., $c(a_{\tau 0} - 0, a_{\tau 0}) > c(a_0 - 0, a_0)$. We can also see that the inequality, $c_\tau(x) + \tau(a_\tau(x)) > c(x)$, continues to hold near $x = 0$. However, this pattern is reversed at some non-central locations, so that $c_\tau(x) + \tau(a_\tau(x)) < c(x)$ holds for x in a certain interval. For larger x, there is no change since residents there do not queue under laissez-faire.

Tolling influences population densities through its effect on the scheduling cost. Figure 9 depicts density profiles under laissez-faire and in social optimum. The figure confirms that the population densities are lower in social optimum than under laissez-faire at the central locations. As explained in the analytic section, this is a consequence of a higher scheduling cost for these central residents. On the contrary, since the scheduling costs (including the toll) are lower under tolling than under laissez-faire for residents located farther away, residents now move there, resulting in a higher density at these locations, as illustrated in Figure 9. Note that housing prices $p(x)$ are lower in the central area and higher in the suburban area under tolling than under laissez-faire, while housing consumption $q(x)$ exhibits the opposite pattern.

Finally, Figure 10 shows the non-housing consumption profiles before and after tolling, i.e., $e(x) = y - c(x) - p(x)q(x)$ and $e_\tau(x) = y - c_\tau(x) - \tau(a_\tau(x)) - p_\tau(x)q_\tau(x)$, respectively. The simulation indicates that the central residents reduce their non-housing consumption under tolling due to their higher commuting costs (including the toll). In contrast, the suburban residents increase their non-housing consumption under tolling. This analysis suggests that the political acceptability of congestion tolling may vary by the location of the residents, if the residents are only conscious of these pecuniary effects.

6 Conclusion

This paper has presented a unification of the bottleneck model of congestion dynamics and the monocentric city model. The model generates a number of new insights regarding

Figure 7: Optimal toll as a function of location

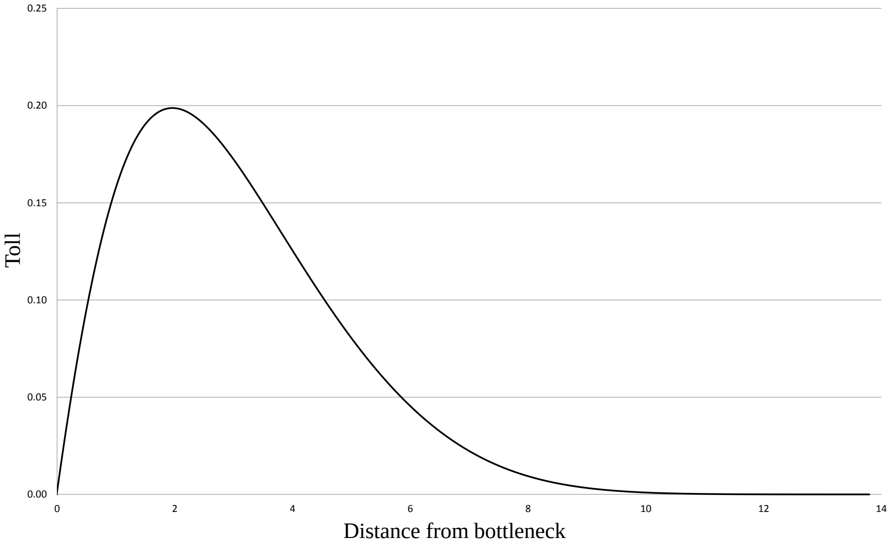


Figure 8: Scheduling-cost profiles with and without tolling

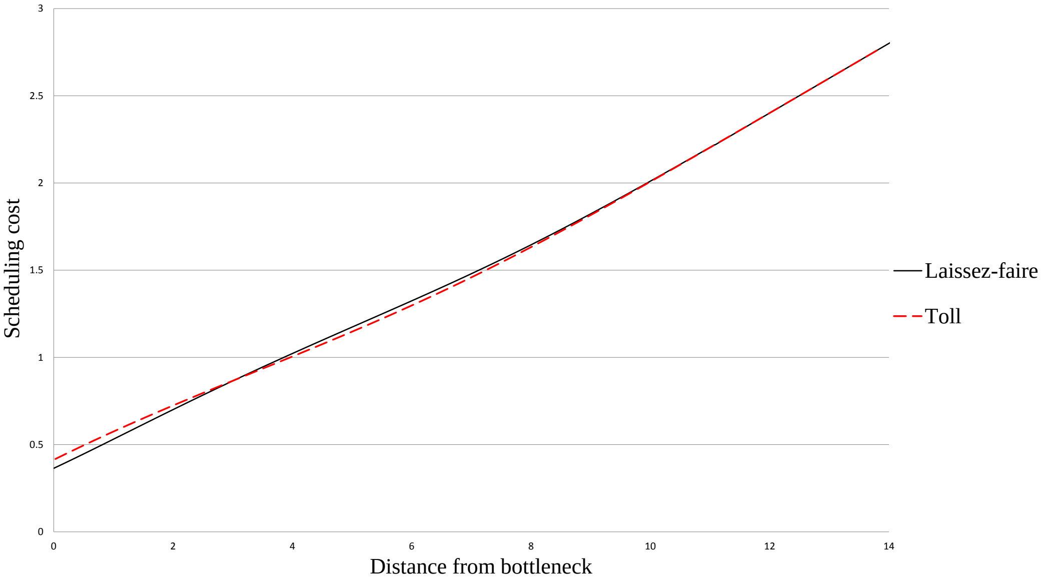


Figure 9: Density profiles with and without tolling

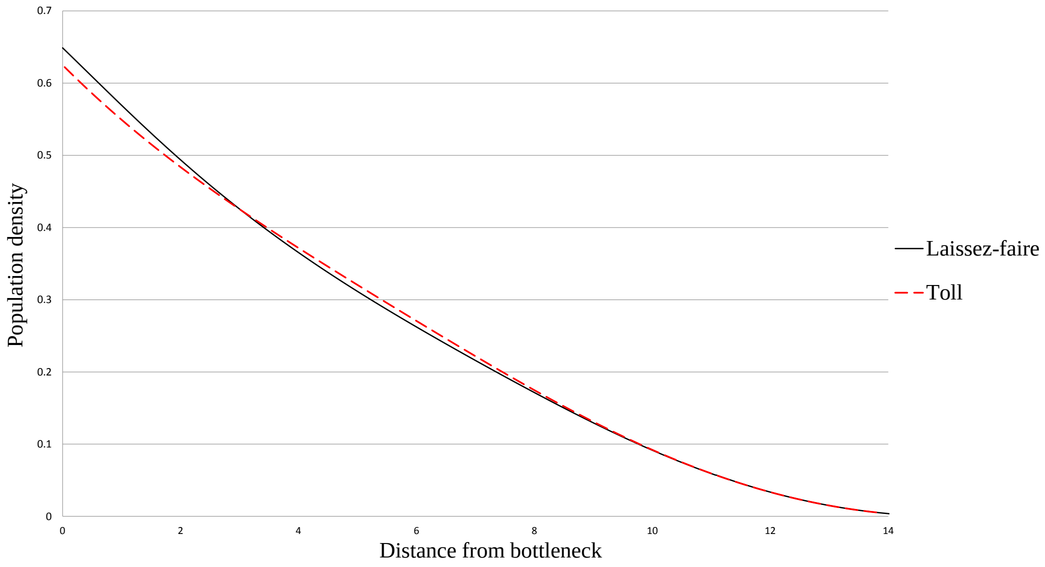
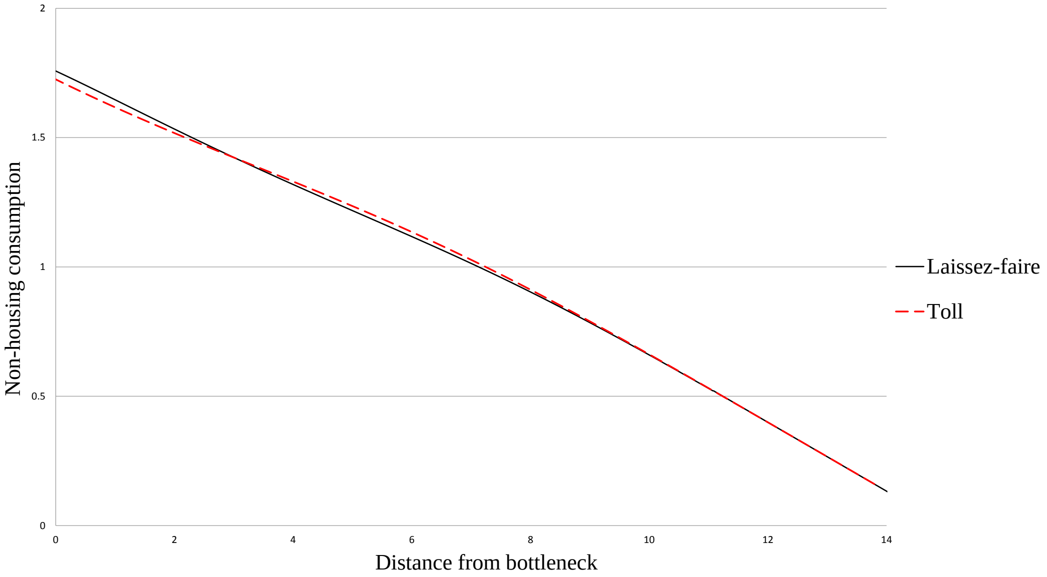


Figure 10: Non-housing consumption profiles with and without tolling



the interaction between congestion dynamics and urban spatial equilibrium. Most of all, unlike in the traditional congested-city models adopting the static congestion assumption, our model generates an optimal city that is less dense in the center and denser in the suburb than the city at the laissez-faire equilibrium. This result is derived analytically under fairly general conditions. It is qualitatively similar to the simulation results in Gubins and Verhoef (2014). Unlike in their model, the density effect of tolling in our model is solely a result of adjustments to commute scheduling and scheduling costs for residents at different locations, we do not require any additions to the specification of commuter preferences to link commute scheduling to space.

Our model generates a number of other implications, such as the relationship between commute scheduling and residential location and population density, the queuing time varying with location, and comparative static results with respect to model parameters.

Our theoretical model can potentially be extended in several ways. Perhaps the most interesting extension would be to incorporate the consumer choice between public transportation and car travel by adding public transportation to the current model. Also, since the present model allows congestion only to occur at the entrance to the CBD, a future work may allow suburban congestion in some form. A way would have to be found to avoid the issues encountered by Ross and Yinger (2000). Going further, it would be relevant to consider urban equilibrium in a model that allows for hypercongestion such as the bathtub model of Fosgerau (2015).

References

Alonso, W., 1964. *Location and Land Use.* Harvard University Press, Cambridge.

Anas, A., Rhee, H., 2006. Curbing excess sprawl with congestion tolls and urban boundaries. *Regional Science and Urban Economics* 36, 510-541.

Arnott, R.A., 1979. Unpriced transport congestion. *Journal of Economic Theory* 21, 294-316.

Arnott, R.A., 1998. Congestion tolling and urban spatial structure. *Journal of Regional Science* 38, 495-504.

Arnott, R.A., de Palma, A., Lindsey, R., 1990. Economics of a bottleneck. *Journal of Urban Economics* 27, 111-130.

Arnott, R.A., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *American Economic Review* 83, 161-179.

Arnott, R.A., DePalma, E., 2011. The corridor problem: Preliminary results on the no-toll equilibrium. *Transport Research Part B: Methodological* 45, 743-768.

Arnott, R.A., Pines, D., Sadka, E., 1986. The effects of an equiproportional transport improvement in a fully-closed Monocentric City. *Regional Science and Urban Economics* Vol. 16(3), pp. 387-406.

Brueckner, J.K., 1987. The structure of urban equilibria: A unified treatment of the Muth-Mills model. *Handbook of Regional and Urban Economics Vol. II*, 821-845.

Brueckner, J.K., 2007. Urban growth boundaries: An effective second-best remedy for unpriced traffic congestion? *Journal of Housing Economics* 16, 263-273.

Fosgerau, M., de Palma, A., 2012. Congestion in a city with a central bottleneck. *Journal of Urban Economics* 71, 269-277.

Fosgerau, M., de Palma, A., 2013. The dynamics of urban traffic congestion and the price of parking. *Journal of Public Economics* 105, 106-115.

Fosgerau, M., Engelson, L., 2011. The value of travel time variance. *Transportation Research Part B: Methodological* 45, 1-8.

Fosgerau, M., Small, K.A., 2013. Endogenous scheduling preferences and congestion. Forthcoming *International Economic Review.*

Fosgerau, M., 2015. Congestion in the bathtub. *Economics of Transportation* 4, 241-255.

Gubins, S., Verhoef, E.T., 2014. Dynamic bottleneck congestion and residential land use in the monocentric city. *Journal of Urban Economics* 80, 51-61.

Mills, E.S., 1972. *Urban Economics.* Glenview, Illinois: Scott Foresman.

Muth, R.F., 1969. *Cities and Housing.* University of Chicago Press, Chicago.

Riley, J.G., 1974. Optimal residential density and road transportation. *Journal of Urban Economics* 1, 230-249.

Ross, S.L., Yinger, J., 2000. Timing equilibria in an urban model with congestion. *Journal of Urban Economics* 47, 390-413.

Small, K.A., 1982. The scheduling of consumer activities: Work trips. *American Economic Review* 72 (3), 467-479.

Solow, R.M., Vickrey, W.S., 1971. Land use in a long narrow city. *Journal of Economic Theory* 3, 430-447.

Taylor, M.E., 2011. Introduction to differential equations. American Mathematical Society, Providence, Rhode Island.

Tsekeris, T., Geroliminis, N., 2013. City size, network structure and traffic congestion. *Journal of Urban Economics* 76, 1-14.

Tseng, Y.Y., Verhoef,, E.T., 2008. Value of Time by Time of Day: A Stated Preference Study. *Transportation Research Part B: Methodological* 42 (7-8): 607–18.

Vickrey, W.S., 1969. Congestion theory and transport investment. *American Economic Review* 59, 251-261.

Vickrey, W.S., 1973. Pricing, Metering, and Efficiently Using Urban Transportation Facilities. *Highway Research Record* 476: 36–48.

Wheaton, W.C., 1998. Land use and density in cities with congestion. *Journal of Urban Economics* 43, 258-272.

A Appendix

Proof of Proposition 2. Commuters arrive at the bottleneck in increasing sequence according to x. There is no queue prior to time a_0 and this implies that $a_0 \leq a_*(0)$ for all x since otherwise the commuter at location 0 could gain by arriving at the bottleneck at time $a_*(0)$.

Suppose that queuing does not begin at time a_0. Then $a(x) = a_*(x)$ in a neighborhood of 0. Due to sorting, $R(a(x)) = F(x)$ and hence $\rho(a_0) = f(0)/a'_*(0)$. Moreover, $a'_*(0) < 1$ by (7) and hence $\rho(a_0) > \psi$, which implies that queuing begins at time a_0, which is a contradiction. Hence, queuing begins at time a_0. ■

Proof of Proposition 3. The statements regarding the queueing commuters ($x < x_1$) is proved in Fosgerau and de Palma (2012) (Theorem 3). The statements regarding $a(x)$ and $t(x)$ for $x \geq x_1$ are immediate. So is continuity of $a(x)$ for $x < x_1$ and $x > x_1$. If $a()$ has a discontinuous jump at x_1, then $\lim_{x \to x_1^-} a(x) \neq a_*(x_1)$. If $\lim_{x \to x_1^-} a(x) < a_*(x_1)$, then commuters near x_1 can delay departure and gain. If $\lim_{x \to x_1^-} a(x) > a_*(x_1)$, then there must be queue at time x_1, which is also a contradiction. ■

Proof of Proposition 4. Evaluation of (10) at x_1 using $a(x_1) = a_*(x_1)$ and $c_1(a_*(x) - x, a_*(x)) + c_2(a_*(x) - x, a_*(x)) = 0$ yields

$$\begin{aligned} a'(x_1) &= -\frac{c_2(a(x_1) - x_1, a(x_1))}{c_1(a(x_1) - x_1, a(x_1))}\frac{f(x_1)}{\psi} \\ &= -\frac{c_2(a_*(x_1) - x_1, a_*(x_1))}{c_1(a_*(x_1) - x_1, a_*(x_1))}\frac{f(x_1)}{\psi} \\ &= \frac{f(x_1)}{\psi} = t'(x_1). \end{aligned} \tag{21}$$

To show that $t(x) > a_*(x)$ for $x < x_1$ near x_1, assume on the contrary that $t(x) \leq a_*(x)$ and recall that $a(x) < t(x)$. Define the function $\Delta(x) = c\left(a(x) - x, a_0 + \frac{F(x)}{\psi}\right) - c(a_*(x) - x, a_*(x))$ and note that $\Delta(x_1) = 0$ and $\Delta'(x_1) = 0$. For $x < x_1$ near x_1, we have $\Delta(x) > 0$ due to the queue and hence $\Delta'(x) < 0$. By enveloping, the derivative of

Δ is $\Delta'(x) = c_1(a_*(x) - x, a_*(x)) - c_1(a(x) - x, t(x))$ and hence

$$\begin{aligned} c_1(a_*(x) - x, a_*(x)) &< c_1(a(x) - x, t(x)) \\ &< c_1(t(x) - x, t(x)) \\ &= c_1(a_*(x) - x, a_*(x)) + (t - a_*(x))(c_{11} + c_{12}) \\ &\leq c_1(a_*(x) - x, a_*(x)), \end{aligned}$$

using that $a(x) < t(x)$, and $t(x) \leq a_*(x)$, the mean value theorem for the equality and Condition 1 for the last inequality. This is a contradiction and we conclude that $t(x) > a_*(x)$ as desired. ■

Proof of Proposition 5.

1. There is no queue, which implies that $t_\tau(a) = a$ always. Trip timing is optimally chosen, so we have the first-order condition with respect to a as

   $$\tau'(a) = -c_1(a - x, a) - c_2(a - x, a), \tag{22}$$

   which holds for $a = a_\tau(x)$.

2. Denoting the second-order condition $SOC \geq 0$, differentiation of the first-order condition (22) with respect to x shows that $0 = c_{11} + c_{12} + SOC \cdot a'_\tau$. Condition 1 then implies that $a'_\tau > 0$.

3. The sum of scheduling cost and toll increases with distance from the CBD,

   $$\frac{\partial}{\partial x}\left[c(a_\tau(x) - x, a_\tau(x)) + \tau(a_\tau(x))\right] = -c_1 > 0,$$

   which implies that the residential density is decreasing with x (see (5) and (6)).

4. Note that $R(a_\tau(x)) = F_\tau(x)$. Suppose tolling does not start at time $a_\tau(0)$. Then departures for residents near location 0 are not constrained by queueing and therefore they arrive at the bottleneck at time $a_*(x)$. This means that $\rho \cdot a'_* = f_\tau$. Also,

differentiation of the first-order condition, $c_1 + c_2 = 0$, with respect to x gives

$$0 < a'_* = \frac{c_{11} + c_{12}}{c_{11} + 2c_{12} + c_{22}} < 1,$$

which holds near location 0 under Conditions 1 and 2. Then,

$$\psi > f_\tau \frac{c_{11} + 2c_{12} + c_{22}}{c_{11} + c_{12}} > f_\tau.$$

But f_τ is decreasing. Then ρ is also decreasing at points with $a_\tau = a_*$, since $\rho \cdot a'_* = f_\tau$. Then commuters can arrive according to a_* throughout the day, which means the toll never becomes active. This is a contradiction. Thus, we conclude that $a_{\tau 0} = a_\tau(0)$. Furthermore, there must be some $x_{\tau 1}$ with $a_{\tau 1} = a_\tau(x_{\tau 1})$, since otherwise the whole city would have arrived at the bottleneck before time $a_{\tau 1}$, which is not possible in equilibrium, since residents far from the CBD would have a_* later than $a_{\tau 1}$ and so could delay departure and gain.

5. This follows since, by assumption, the bottleneck is fully utilized during $[a_{\tau 0}, a_{\tau 1}]$.

6. If $a_{\tau 1} < a_*(x_{\tau 1})$, then the last tolled resident could postpone departure slightly and gain. If the converse inequality holds, then the resident at $x_{\tau 1}$ would not be the last to be tolled, which is a contradiction.

7. If $\tau(a_{\tau 1}) > 0$, then residents just before x_1 could delay departure slightly to avoid the toll and gain a jump in utility. This is a contradiction with equilibrium and we conclude that $\tau(a_{\tau 1}) = 0$.

8. Assume on contrary that $\tau(a_{\tau 0}) > 0$. Then resident at location 0 could gain by arriving at the bottleneck slightly before $a_{\tau 0}$. This is a contradiction, and we conclude that $\tau(a_{\tau 0}) = 0$

9. This follows from the first-order condition (22) and the first-order condition for a_* by Conditions 1 and 2.

■

Lemma 1 *Assume Condition 4. Let Θ be parametrized in terms of $a_{\tau 0}$. Then $\frac{\partial p_\tau(x)}{\partial a_{\tau 0}} = -\frac{\partial \tau(a)}{\partial a_{\tau 0}}|_{a=a_\tau(x)} \cdot f_\tau(x)$.*

Proof. Utility remains constant for every toll in Θ, i.e.,

$$U(y - c(a_\tau(x) - x, a_\tau(x)) - \tau(a_\tau(x)) - p_\tau(x)q_\tau(x), q_\tau(x)) = \bar{U}. \tag{23}$$

From Condition 4, we know that each toll function τ is uniquely determined from $a_{\tau 1}$, and $a_{\tau 1}$ uniquely detemines $a_{\tau 0}$. Therefore each toll function is uniquely determined from $a_{\tau 0}$. Taking derivative of (23) with respect to $a_{\tau 0}$ (suppressing some notation), and using enveloping ((2) and (15)), we get

$$q_\tau(x)\frac{\partial p_\tau(x)}{\partial a_{\tau 0}} = -\frac{\partial \tau(a)}{\partial a_{\tau 0}}|_{a=a_\tau(x)}.$$

Then use $q_\tau = 1/f_\tau$ to obtain the conclusion. ■

Proof of Proposition 6. Θ is the set of tolls that removes queue on the finite interval $[a_{\tau 0}, a_{\tau 1}]$ and that lead to equilibrium with the entry rate at the bottleneck at capacity when tolling is active and below capacity otherwise. Taking derivative of the welfare function (14) with respect to $a_{\tau 0}$, we get

$$\frac{\partial W}{\partial a_{\tau 0}} = \int_0^\infty \frac{\partial p_\tau(x)}{\partial a_{\tau 0}} dx + \psi \frac{\partial}{\partial a_{\tau 0}} \int_{a_{\tau 0}}^{a_{\tau 1}} \tau(a) da. \tag{24}$$

Using Lemma 1, and substituting $q(x) = 1/f(x)$, we get

$$\begin{aligned}\frac{\partial W}{\partial a_{\tau 0}} &= -\int_0^\infty \frac{\partial \tau(a)}{\partial a_{\tau 0}}|_{a=a_\tau(x)} \cdot f_\tau(x) dx + \psi \frac{\partial}{\partial a_{\tau 0}} \int_{a_{\tau 0}}^{a_{\tau 1}} \tau(a) da. \\ &= -\psi \int_{a_{\tau 0}}^{a_{\tau 1}} \frac{\partial \tau(a)}{\partial a_{\tau 0}} da + \psi \left[\tau(a_{\tau 1}) \frac{\partial a_{\tau 1}}{\partial a_{\tau 0}} - \tau(a_{\tau 0}) + \psi \int_{a_{\tau 0}}^{a_{\tau 1}} \frac{\partial a_\tau}{\partial a_{\tau 0}} da \right] \\ &= \psi \left[\tau(a_{\tau 1}) \frac{\partial a_{\tau 1}}{\partial a_{\tau 0}} - \tau(a_{\tau 0}) \right].\end{aligned} \tag{25}$$

But Proposition 5 shows that $\tau(a_{\tau 1}) = \tau(a_{\tau 0}) = 0$. Therefore the welfare function is constant on Θ. ■

Proof of Proposition 7.

1. Define the function

$$h(x) \equiv c\left(a_\tau(x) - x, a_\tau(x)\right) + \tau(a_\tau(x)) - c\left(a(x) - x, t\left(x\right)\right). \tag{26}$$

By enveloping, we have

$$h'\left(x\right) = c_1\left(a(x) - x, t\left(x\right)\right) - c_1\left(a_\tau(x) - x, a_\tau(x)\right). \tag{27}$$

Note that when $h\left(x\right) > 0$, $f_\tau\left(x\right) < f\left(x\right)$ holds. Also note that $a\left(x\right) = a_*\left(x\right)$ and $a_\tau\left(x\right) \neq a_*\left(x\right)$ implies $h\left(x\right) > 0$. For $x \geq x_1$, we have $a\left(x\right) = a_*\left(x\right)$ and capacity is not utilized in this regime, which implies that $a_*'\left(x\right) > f\left(x\right)/\psi$ (seen by differentiation of $R(a(x)) = F(x)$ using $\psi > \rho$).

   (a) Consider the first case where $x_{\tau 1} > x_1$, and let us seek a contradiction with $a_0 < a_{\tau 0}$. There must be an interval during $]x_1, x_{\tau 1}[$ where $a_\tau\left(x\right) \neq a_*\left(x\right)$ and thus $f_\tau\left(x\right) < f\left(x\right)$. This implies that $a_\tau'\left(x\right) = f_\tau\left(x\right)/\psi < f\left(x\right)/\psi < a_*'\left(x\right)$. Hence if $a_\tau\left(x\right) < a_*\left(x\right)$ at some point, a_τ will remain below a_*. Since this is contradictory to $a_{\tau 1} = a_*(x_{\tau 1})$, we conclude that $a_\tau\left(x\right) \geq a_*\left(x\right)$ for all $x \in]x_1, x_{\tau 1}[$.

   For any x in this interval with $a_\tau\left(x\right) > a_*\left(x\right)$, we have $h\left(x\right) > 0$, $h'\left(x\right) < 0$ by $c_{11} + c_{12} > 0$ (Condition 1), and $a_\tau'\left(x\right) < a_*'\left(x\right)$ as shown just above.

   We are assuming $a_0 < a_{\tau 0}$ to establish its contradiction. Recall that $t\left(a_0\right) = a_0$. Note that $h\left(0\right) < 0$ and that $a\left(x\right) < a_\tau\left(x\right)$ implies $h'\left(x\right) < 0$ from (27) with $c_{11} > 0$ and $c_{22} > 0$ under $c_{12} = 0$. Then, $h < 0$ for some interval.

   During this interval, $f_\tau\left(x\right) > f\left(x\right)$ and hence $\frac{f_\tau(x)}{\psi} = a_\tau'\left(x\right) > t'\left(a\left(x\right)\right)a'\left(x\right) = \frac{f(x)}{\psi}$. This implies that $a_\tau\left(x\right) > t\left(a\left(x\right)\right) \geq a\left(x\right)$ and hence that $h' < 0$ at the end of the interval. This argument applies until $x = x_1$.

   In the interval $[x_1, x_{\tau 1}]$, we also have $h'\left(x\right) < 0$, which is a contradiction with $h\left(x_{\tau 1}\right) = 0$. This establishes that $a_{\tau 0} \leq a_0$.

(b) Consider the other case where $x_{\tau 1} < x_1$. We want to show a contradiction of $a_0 < a_{\tau 0}$. Note first that $a_\tau(x) = a_*(x)$ for an interval during $]x_{\tau 1}, x_1[$ and therefore $h(x) < 0$ and $f_\tau(x) > f(x)$ hold during the same interval. Also, the capacity is not fully utilized under the toll regime for $x \geq x_{\tau 1}$, which implies $a'_*(x) > f_\tau(x)/\psi$.

Therefore, $a'_*(x) > f_\tau(x)/\psi > f(x)/\psi = t'(a(x))a'(x)$ holds during $]x_{\tau 1}, x_1[$. Hence, if $a_\tau(x) > t(x)$ at $x_{\tau 1}$, $a_*(x) = a_\tau(x) > t(x) \geq a(x)$ holds for any x in $]x_{\tau 1}, x_1[$. Therefore, during this interval, $h'(x) < 0$.

As shown above, $h(x) < 0$ and $h'(x) < 0$ until $x_{\tau 1}$. Also, we've just showed $h'(x) < 0$ during $]x_{\tau 1}, x_1[$. Therefore, $h(x_1) = 0$ is impossible, which is contradictory.

The remaining issue is whether $a_\tau(x) \leq t(x)$ is possible at $x_{\tau 1}$. But, this is impossible, because as shown above $h < 0$ until $x_{\tau 1}$, which implies $f_\tau(x) > f(x)$. So, $a_\tau(x) = a_{\tau 0} + \frac{F_\tau(x)}{\psi} > a_0 + \frac{F(x)}{\psi} = t(x)$ holds until $x_{\tau 1}$.

Therefore, we conclude that $a_0 < a_{\tau 0}$ is contradictory, regardless of the relative sizes of $x_{\tau 1}$ and x_1, establishing that $a_{\tau 0} \leq a_0$.

2. This just follows from $a_0 \geq a_{\tau 0}$ and Condition 1.

■

Proof of Proposition 8.

1. The first result just follows from $a_{\tau 0} \leq a_0$, under which $h(0) \geq 0$ and $f_\tau(0) \leq f(0)$. By continuity, $f_\tau(x) \leq f(x)$ for x near 0.

2. (a)

   (b) To prove that there exist locations with $f_\tau(x) > f(x)$, consider first the case where $x_{\tau 1} < x_1$. Note that $a_\tau(x) = a_*(x)$ during the interval $]x_{\tau 1}, x_1[$. Then there must be an interval during $]x_{\tau 1}, x_1[$ where $a(x) \neq a_*(x)$ and thus $h(x) < 0$ and $f_\tau(x) > f(x)$.

(c) Consider instead the case where $x_1 < x_{\tau 1}$. Note that $a(x) = a_*(x)$ during the interval $]x_1, x_{\tau 1}[$. Since the residents at these locations do pay tolls under optimal tolling, $h(x) > 0$ and $f_\tau(x) < f(x)$ hold in this interval.

Also note that residents located at $]x_1, x_{\tau 1}[$ do not fully utilize capacity under laissez-faire, while capacity is fully utilized under tolling for those at $]x_1, x_{\tau 1}[$. Hence,

$$\frac{F(x_{\tau 1}) - F(x_1)}{a(x_{\tau 1}) - a(x_1)} < \psi = \frac{F_\tau(x_{\tau 1}) - F_\tau(x_1)}{a_\tau(x_{\tau 1}) - a_\tau(x_1)} < \frac{F(x_{\tau 1}) - F(x_1)}{a_\tau(x_{\tau 1}) - a_\tau(x_1)},$$

which leads to

$$a_\tau(x_{\tau 1}) - a_\tau(x_1) < a(x_{\tau 1}) - a(x_1).$$

Then

$$a_*(x_{\tau 1}) - a_\tau(x_1) < a_*(x_{\tau 1}) - a_*(x_1),$$

which implies that $a_*(x_1) < a_\tau(x_1)$, or equivalently,

$$a_*(x_1) = a_0 + \frac{F(x_1)}{\psi} < a_{\tau 0} + \frac{F_\tau(x_1)}{\psi} = a_\tau(x_1). \tag{28}$$

As shown above, $f_\tau(x) < f(x)$ for x near 0. Given $a_{\tau 0} \leq a_0$, for (28) to hold, there must be some locations where $f_\tau(x) > f(x)$ between 0 and x_1.

(d) Consider finally the case where $x_1 = x_{\tau 1}$. We have

$$\psi = \frac{F(x_1)}{a_*(x_1) - a(0)} = \frac{F_\tau(x_{\tau 1})}{a_*(x_{\tau 1}) - a_\tau(0)},$$

$a_\tau(0) \leq a(0)$, and $a_*(x_1) \leq a_*(x_{\tau 1})$, which implies that $F_\tau(x_1) \geq F(x_1)$. As $f_\tau(x) < f(x)$ for x near 0, we must have $f_\tau(x) > f(x)$ for x elsewhere in the interval $]0, x_1[$.

■