THEMA

théorie économique, modélisation et applications

THEMA Working Paper n°2024-04
CY Cergy Paris Université, France

Bottleneck congestion and urban spatial structure with heterogeneous households: Equilibrium, capacity expansion and congestion tolling

Zhi-Chun Li, De-Ping Yu, André de Palma

May 2024

CERGY PARIS UNIVERSITÉ

cnrs

Bottleneck congestion and urban spatial structure with heterogeneous households: Equilibrium, capacity expansion and congestion tolling

Zhi-Chun Li [a,*], De-Ping Yu [a], André de Palma [b]

[a] *School of Management, Huazhong University of Science and Technology, Wuhan 430074, China*

[b] *CY Cergy-Paris University, Thema, France*

Abstract

We propose an analytical solvable model for household residential location choice in a linear monocentric city corridor with bottleneck congestion. Households are heterogeneous in terms of their income. The bottleneck is located between central downtown and adjacent suburb. The urban equilibrium is formulated as the solution of differential equations. We analytically explore the distributional effects of bottleneck capacity expansion on households and the bottleneck capacity investment issues under no toll and first-best and second-best tolls. The results show that the benefits of different-income households from bottleneck capacity expansion change with toll schemes. Specifically, under the no toll and first-best toll, those who gain most are the mid-income households residing at the bottleneck and in a suburban location (close to the bottleneck) respectively, whereas those who gain least are the poorest or richest households. Under the second-best toll, there are two possible cases: the poorest households gain most while the richest households gain least, or the mid-income households residing at the bottleneck gain most while the richest or poorest households gain least. With constant return to scale for capacity investment, self-financing principle still holds for the first-best and second-best tolling in the urban spatial context. Ignoring the changes in urban spatial structure due to household relocation may cause overinvestment or underinvestment in optimal bottleneck capacity under the no toll, but definitely underinvestment under the first-best and second-best tolls.

Keywords: Bottleneck congestion; urban spatial structure; heterogeneous households; capacity expansion; distributional effect; congestion tolling; self-financing.

JEL classification: R13, R14, R41, R42

[*]Tel.: +86-27-8755-6490; Fax: +86-27-8755-6437.

E-mail addresses: smzcli@gmail.com (Z.-C. Li), andre.de-palma@cyu.fr (A. de Palma).

1. Introduction

Traffic congestion during peak periods has become increasingly severe in many densely populated megacities worldwide, mainly due to the imbalance between traffic demand and infrastructure supply. Traffic congestion dynamics directly affect commuting costs, which significantly influence households' residential location choices, especially for low-income households. It is, therefore, essential to consider the interactions among traffic congestion dynamics, residential location choices, and household income heterogeneity in urban models, as pointed out in previous studies like Takayama and Kuwahara (2017) and Takayama (2020).

Traffic congestion is a deadweight loss and resource wastage for society. A possible solution to mitigating the growing traffic congestion may be to expand the capacity of traffic bottleneck (a supply-side measure). Unfortunately, bottleneck capacity expansion may induce new traffic demand and thus cause further congestion during peak periods. Furthermore, the funds available for capacity expansion remain limited in most urban areas. Congestion tolling, as a demand-side measure, has been widely suggested as a viable alternative to capacity expansion because of its potential to internalize congestion externalities. Typical examples include the tolling schemes adopted in London, Singapore, Stockholm, Oslo, and Hong Kong (de Palma and Lindsey, 2011). Naturally, implementations of capacity expansion and congestion tolling raise some important issues in terms of welfare, equity, and public finance: How does bottleneck capacity expansion under tolling affect the residential location choices of households with different income levels? Who gains and who loses? Does the self-financing principle (revenue from congestion tolls covers cost of capacity expansion) hold? What happens if the changes in urban spatial structure due to household relocation under bottleneck capacity expansion and/or congestion tolling are ignored. This paper aims to address these important social issues.

We begin by establishing the urban equilibrium with dynamic congestion and heterogeneous households. We consider a continuous linear monocentric city corridor, with a bottleneck located in the corridor. Residents are heterogeneous in terms of their values of time (VOT), following a continuous distribution. They decide where to live along the corridor and when to depart from home for commuting trips. Based on the residential location theory (see e.g., Alonso, 1964; Muth, 1969; Mills, 1972; Fujita, 1989) and the dynamic bottleneck congestion theory (see e.g., Vickrey, 1969; Arnott et al., 1990, 1993, 1994; Li et al., 2020), we first derive the household residential distribution along the corridor at equilibrium, formulated as the solution to differential equations. We extend the studies of Takayama and Kuwahara (2017) and Takayama (2020) from a discrete VOT case to a continuous VOT case. To our knowledge, this paper is the first to obtain a closed-form solution for the combined problem of heterogeneous households' residential location choices and dynamic bottleneck congestion.

Building upon the established urban equilibrium, we address the bottleneck capacity expansion issues under three scenarios: no toll, first-best (dynamic) toll, and second-best (flat) toll. The flat toll means that each driver passing through the bottleneck faces the same toll level.[1] Our investigation focuses on three key aspects: (i) the distributional effects of bottleneck capacity expansion on heterogeneous households, (ii) the self-financing principle for bottleneck capacity investment with congestion tolling in the urban spatial context, and (iii) the effects of ignoring urban spatial changes on capacity investment decisions.

We find that bottleneck capacity expansion benefits all households in the corridor regardless of whether congestion tolling is implemented or not. However, such benefits are differentiated across different-income households. The mid-income households residing at the bottleneck gain most from capacity expansion under no toll, whereas the mid-income households residing in a suburban location (close to the bottleneck) gain most under the first-best toll. For both of them, the poorest households residing at the city boundary or the richest households residing at the CBD (central business district) gain least. Under the second-best toll, two possible cases occur: the poorest households gain most while the richest households gain least, or the mid-income households residing at the bottleneck gain most while the richest or poorest households gain least. These results imply that congestion tolling significantly affects the distributional effects of bottleneck capacity expansion on households. Our results are comparable to previous studies. Arnott et al. (1994) showed that bottleneck capacity investment may benefit higher- or lower-income commuters more, without urban spatial

[1] Flat (or uniform) toll means an identical toll over a day, and thus all commuters passing through the bottleneck face the same toll, regardless of their arrival time at the bottleneck. Therefore, its implementation does not change commuters' departure pattern. Different from the flat toll, coarse tolling means step tolling, which needs to determine the toll level and the start and end times of tolling for each step. Implementation of coarse tolling will affect commuters' departure pattern. For these definitions, readers can refer to Arnott et al. (1990, 1993), and Tabuchi (1993).

dimension consideration. Takayama (2020) showed that with the bottleneck located at the CBD entrance (which is a special case of any bottleneck locations as addressed in this paper), commuters residing closer to the CBD gain but those residing farther from the CBD lose. Our study extends these studies by incorporating endogenous urban spatial structure and any bottleneck locations.

Our work also extends the famous self-financing principle for facility improvement (Mohring and Harwitz, 1962; de Palma and Lindsey, 2007) to urban spatial context, in which urban spatial structure is endogenously determined. We demonstrate that under congestion tolling (first-best and second-best), the ratio of toll revenue to optimal capacity investment cost is equal to the capacity elasticity of the investment cost. Particularly, as the capacity cost exhibits a constant return to scale, the optimal capacity investment is exactly self-financing. Arnott et al. (1993) presented the elasticity-related self-financing principle for the bottleneck capacity expansion under the assumption of homogeneous commuters and absent urban spatial structure. Arnott and Kraus (1995) further extended the analysis to the case of heterogeneous commuters, showing that the self-financing property holds for the first-best dynamic toll but not for the uniform (or flat) toll. This paper finds that in the context of heterogeneous commuters and endogenous urban spatial structure, the self-financing principle remains valid for toll-funded bottleneck capacity investment issue regardless of toll types (first-best dynamic or second-best flat toll).

In addition, we disclose the effects of ignoring the interplays between bottleneck capacity expansion and urban spatial structure. We find that under no toll, ignoring the urban spatial effects (i.e., change in urban spatial structure due to capacity expansion) may lead to an overinvestment or underinvestment in the capacity. However, under congestion tolling (first-best and second-best), it will cause an underinvestment. These results are different from those of traditional urban models with homogeneous households and static congestion (i.e., congestion level depends simply on traffic volume, regardless of time-of-use pattern). For example, Solow (1973) numerically showed that with no toll, the naïve cost-benefit analysis (with the rule that the marginal saving in transportation cost equals the marginal cost of road capacity expansion) tended to overinvest in road capacity. Kanemoto (1977) argued that the naïve method caused overinvestment in capacity near the CBD, but either overinvestment or underinvestment near the city boundary.

Our work is closely related to the studies combining dynamic bottleneck congestion and residential location choice. Arnott (1998) first introduced an urban model in this regard, but assumed homogeneous households and a downtown-suburb discrete city. Fosgerau and de Palma (2012) extended it to consider a continuous space but with an exogenous residential distribution. Gubins and Verhoef (2014) assumed that the marginal utility of households spending time at home depends on housing size to link household preferences for housing consumption and for departure time. Such an assumption is pivotal in their model, but it requires further validation and justification, as highlighted in Fosgerau and Kim (2019). Fosgerau et al. (2018) considered more general scheduling preferences to avoid such an assumption. Further, Fosgerau and Kim (2019) considered a discrete city with a central residential area and a suburban residential area, each having a bottleneck. These aforementioned studies assumed homogeneity in household income and a bottleneck located at the entrance to the CBD. Recently, Takayama and Kuwahara (2017) and Takayama (2020) made a major contribution to the issues of heterogeneous households' residential location choices and bottleneck congestion. In their studies, heterogeneous households were discretized into limited income groups. However, they could not obtain closed-form solutions. Moreover, they focused on the demand-side congestion pricing, and paid little attention to the supply-side bottleneck capacity expansion.

This paper presents an analytical model for heterogeneous households' residential location choices and bottleneck capacity expansion in a linear monocentric city with any bottleneck locations, yielding significant insights into congestion dynamics, household heterogeneity, and urban spatial structure. The proposed model can serve as a useful tool for modeling household residential location choices and evaluating various urban policies.

The remainder of this paper is organized as follows. In the next section, the urban equilibrium with bottleneck congestion and heterogeneous households is formulated. Section 3 addresses the bottleneck capacity expansion issues under no toll. Section 4 further addresses the bottleneck capacity expansion issues under the first-best and second-best tolls. Section 5 concludes this paper and provides suggestions for further studies. Some proofs, mathematical derivations, and numerical examples are given in the appendices.

2. Urban equilibrium with bottleneck congestion and heterogeneous households

2.1. Basic setup

Consider a transportation corridor located in a closed, linear, and monocentric city, with a population size of N. The city's residents continuously distribute along the corridor. They have different income levels and thus different VOTs. In this paper, we assume VOT is positively related to income. We represent α as resident's VOT, with $\alpha \in [\underline{\alpha}, \overline{\alpha}]$ in which $\underline{\alpha}$ and $\overline{\alpha}$ are the lower and upper bounds, respectively. We use the distance from the CBD to denote a location. A bottleneck with capacity q is located at location a, as shown in Fig. 1. The bottleneck location divides the corridor into two areas: downtown (i.e., $[0, a)$) and suburb (i.e., $[a, B]$). Traffic congestion during the commuting peak period occurs at the bottleneck due to its limited capacity.[2] The length of the corridor (or city boundary) is B, endogenously determined by the model. The words "commuter", "resident" and "household" are interchangeable in this paper.

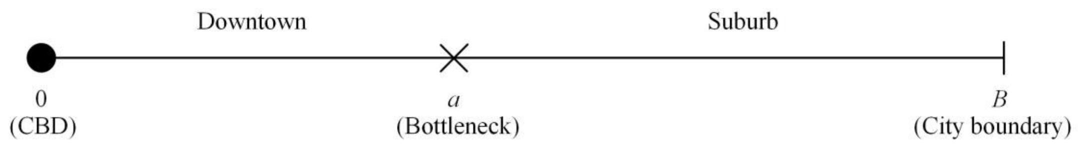

Fig. 1. A linear city corridor with a bottleneck.

All job opportunities are located in the CBD. Every morning, commuters travel from home to the CBD along the bottleneck-constrained corridor. Commuters' travel costs depend on their home locations. The commuters originating in the downtown do not pass through the bottleneck, and thus incur no congestion. All of them prefer punctual arrivals at the workplace without causing any schedule delay. Their commuting costs thus include only the free-flow travel time cost. Let $c_D(x, \alpha)$ be the (one-way) commuting cost of the downtown commuters with VOT α and residential location x ($0 \le x < a$), defined as

---

[2] This paper focuses on the interactions between dynamic congestion and urban spatial structure. For analytical tractability, we assume that congestion occurs only at the bottleneck of the corridor, and other locations are congestion-free.

$$c_D(x,\alpha) = t_0 x\alpha\,, \tag{1}$$

where the subscript "*D*" represents the downtown area, and t_0 denotes the free-flow travel time per unit of distance, and thus $t_0 x$ is the free-flow travel time from *x* to CBD. The VOT α is used to convert time units into equivalent monetary units.

By contrast, commuters residing in the suburb have to traverse the bottleneck on their way to work, thus incurring a queuing delay. Similar to most bottleneck congestion studies, the queue at the bottleneck is vertical and has no physical length.[3] For simplicity, we assume that the penalty for being late is infinite, and thus no late arrivals are permitted. Such an assumption has been adopted in some revious realted studies, such as Newell (1988), DePalma and Arnott (2012), Xiao and Zhang (2014), and Wu and Huang (2015). Therefore, commuters' departure time choices will be based on a trade-off between the bottleneck queuing delay and early-arrival schedule delay. Let $c_S(x,\alpha)$ be the equilibrium commuting cost of suburban commuters with VOT α and location *x* ($a \le x \le B$). According to the bottleneck theory with continuous VOT distribution, we have

$$c_S(x,\alpha) = \frac{\hat{N}_S}{q}\beta + t_0 x\alpha = \frac{\hat{N}_S}{q}\eta\alpha + t_0 x\alpha, \tag{2}$$

where the subscript "*S*" represents the suburban area. $\hat{N}_S$ is the total number of suburban residents (i.e., bottleneck users), determined endogenously later. *q* is the bottleneck capacity, β is the value of early-arrival time, and η is the ratio of the value of early-arrival time to VOT, i.e., $\eta = \beta/\alpha$. In this paper, we assume η is a constant across residents, as in some bottleneck models (see e.g., Vickrey, 1973; Arnott et al., 1994; Xiao and Zhang, 2014). This means that households differ in both their VOTs and values of early-arrival time. The first term on the right-hand side (RHS) of Eq. (2) is the equilibrium bottleneck congestion cost (including the queuing delay cost and the early-arrival schedule delay cost). The second term is the free-flow travel time cost. The detailed derivation of Eq. (2) is provided in Appendix A.

To sum up, the commuting cost $c(x,\alpha)$ can be represented as

$$c(x,\alpha) = \begin{cases} c_D(x,\alpha), & x \in [0,a), \\ c_S(x,\alpha), & x \in [a,B]. \end{cases} \tag{3}$$

[3] For the model with physical vehicle queue length consideration, readers can refer to Mun (1999).

2.2. Equilibrium household residential distribution

Traditional urban models usually assume that urban households are homogeneous in their income levels or VOTs. However, in reality income across households varies, depending on their occupations and skills. This leads to heterogeneity in residents' VOTs, explicitly treated in this paper. In the following, we explore the difference in the residential location choices of households due to their VOT heterogeneity.

Households obtain utility from land (or equivalently housing) consumption and non-land (numéraire) goods consumption. Households' preferences for land and numéraire goods are quasi-linear, given by

$$u(x,\alpha) = z(x,\alpha) - \frac{k}{2h(x,\alpha)}, \tag{4}$$

where $u(x,\alpha)$ is utility of households with location x and VOT α, $z(x,\alpha)$ is numéraire goods consumption (measured in monetary units), and $h(x,\alpha)$ is land consumption (measured in land areas).[4] The positive constant k represents households' preferences for land (a larger value indicates a stronger preference, and vice versa). The second term on the RHS of Eq. (4) represents the households' utility derived from land consumption, measured in monetary units. Such a hyperbolic utility function has been adopted in some previous residential location models (see e.g., Mossay and Picard, 2011; Picard and Tabuchi, 2013; Blanchet et al., 2016; Akamatsu et al., 2017; Picard and Tran, 2021).[5] Eq. (4) measures the household utility in monetary units, i.e., cardinal utility, which facilitates the comparison between utilities of different households and the calculation of social surplus (or welfare).

In this paper, the price of the numéraire goods is normalized to one, but the land price is

[4] In this paper, it is assumed that utility u is twice continuously differentiable with regard to z and h, the distribution function of VOT α is continuous, and the commuting cost $c(x,\alpha)$ is piecewise continuous with the bottleneck location as the discontinuous point.

[5] The hyperbolic and logarithmic preferences (see e.g., Beckmann, 1976; Fujita and Thisse, 2002) for the land are two frequent instances of the same class of preferences $(h^{1-\rho}-1)/(1-\rho)$ where $\rho = 2$ and $\rho \to 1$ respectively, which yield iso-elastic demands for residential space with price elasticity equal to 1/2 and 1, respectively. Therefore, the present hyperbolic preference represents an intermediate case between Beckmann's demand and the inelastic demand for residential space that is standard in urban economics.

endogenously determined by the model. Households' income is spent on numéraire goods, land, and commuting, and their budget constraint is given as

$$w(\alpha) = z(x,\alpha) + p(x)h(x,\alpha) + c(x,\alpha), \tag{5}$$

where $w(\alpha)$ is the income of the household with VOT α. $p(x)$ is land rental price, and thus $p(x)h(x,\alpha)$ is land expenditure for households with location *x* and VOT α.

Each household chooses a residential location, land area, and quantity of numéraire goods to maximize utility subject to the budget constraint. From Eqs. (4) and (5), the utility maximization problem for households at a given location *x* can be expressed as

$$\max_h \ u(x,\alpha) = w(\alpha) - p(x)h(x,\alpha) - c(x,\alpha) - \frac{k}{2h(x,\alpha)}. \tag{6}$$

From the first-order optimality condition $du/dh = 0$, we have

$$h(x) = \sqrt{\frac{k}{2p(x)}}. \tag{7}$$

Eq. (7) shows that at a given location, household land consumption is only related to the land price but not to their income, because household preference is quasi-linear.

Substituting Eq. (7) into Eq. (6) yields the household indirect utility as

$$u(x,\alpha) = w(\alpha) - c(x,\alpha) - \sqrt{2kp(x)}. \tag{8}$$

At equilibrium, no household has an incentive to unilaterally change its residential location. That is, the utility has been maximized for the chosen location, which corresponds to the condition $\partial u(x,\alpha)/\partial x = 0$.[6] Combing it and Eq. (8) yields

$$\frac{dp(x)}{dx} = -\alpha t_0 \sqrt{\frac{2p(x)}{k}} < 0. \tag{9}$$

It shows that as the distance from the CBD increases, the land rental price decreases. There is thus a trade-off between commuting cost and land rental price when choosing residential locations. Heterogeneous residents with different income levels / VOTs exhibit different attitudes towards such a trade-off.

[6] Since $u(x,\alpha)$ is continuously differentiable (except at the bottleneck), the utility maximization condition for the chosen location $x \in (0, B)$ is equivalent to the first-order and second-order conditions: $\partial u(x,\alpha)/\partial x = 0$ and $\partial^2 u(x,\alpha)/\partial x^2 < 0$.

Since households' utility has been maximized under equilibrium location choices, exchanging households' locations would reduce their total utility. That is, for any two residential locations x_1 and x_2 (without loss generality, we assume $x_1 > x_2$), with corresponding VOTs α_1 and α_2, $u(x_1,\alpha_1)+u(x_2,\alpha_2) > u(x_1,\alpha_2)+u(x_2,\alpha_1)$ should hold. Since $u(x,\alpha)$ in Eq. (8) is submodular in x and α, we immediately obtain $\alpha_1 < \alpha_2$.[7] This means that households residing farther from the CBD have a smaller VOT than those closer to the CBD. We have thus the following proposition.

Proposition 1. At equilibrium, households spatially sort themselves in a descending order of VOTs from the CBD outward, i.e., households with higher VOTs reside closer to the CBD, while those with lower VOTs reside closer to the suburb.

Proposition 1 indicates that a household's location is determined by its VOT or income. Households with higher VOTs prefer to live closer to the CBD due to their greater aversion to high commuting costs over high land prices.[8] A similar residential sorting has also been shown in Takayama and Kuwahara (2017), but with a discrete households' VOT distribution.

2.3. Equilibrium household residential density

We have derived the residential sorting along the corridor in the previous section. In this section, we further derive the household residential density, i.e., the number of households per unit of land area. According to Eqs. (1) and (2), the commuting cost is discontinuous at the bottleneck (an upward jump). Therefore, in the following we in turn derive the downtown and suburban residential densities.

Let $n_D(x)$ be the household residential density at downtown location x, and $N_D(x)$ be the

---

[7] It is easy to show that $\partial^2 u(x,\alpha)/\partial x \partial \alpha < 0$ holds, and thus $u(x,\alpha)$ is submodular in x and α (see e.g., Topkis, 1998; Fujishige, 2005).

[8] It should be pointed out that the residential sorting presented in Proposition 1 depends on the quasi-linear utility function, due to its inherent property of income-inelastic land consumption (see Eq. (7)). The quasi-linear utility function may apply to Chinese and European cities where the rich live in the city central areas, like Beijing, Shanghai, Paris or London. We admit that other utility functions, such as Cobb-Douglas functions, can be adopted for a further study (see e.g., Takayama, 2020), but may incur analytical intractability.

cumulative number of households from the CBD to *x*. Analogously, $n_S(x)$ and $N_S(x)$ are the residential density and cumulative number of households in a suburban location *x*, respectively. Obviously, $dN_i(x)/dx = n_i(x),\ i = D, S$ holds. Land supply is uniform across the city, and normalized to 1 at any location of the corridor without loss of generality. Housing consumption per household equals land supply divided by the number of households, i.e., $h_i(x) = 1/n_i(x),\ i = D, S$. Substituting it into Eq. (7), the land rental price can be expressed as a function of residential density:

$$p_i(x) = \frac{k}{2}\left(n_i(x)\right)^2,\ i = D, S. \tag{10}$$

Plugging it into Eq. (9) yields the following first-order ordinary differential equation regarding residential density:

$$t_0\alpha + k\frac{dn_i(x)}{dx} = 0,\ i = D, S. \tag{11}$$

For analytical tractability, we assume that households' VOTs are uniformly distributed, i.e., $\alpha \sim U[\underline{\alpha}, \overline{\alpha}]$. Since households are spatially distributed in a descending order of VOTs outward (see Proposition 1), there is a one-to-one correspondence between location *x* and VOT α. The VOTs at the CBD and the city boundary are $\overline{\alpha}$ and $\underline{\alpha}$, respectively. Let α^* be the household VOT at the bottleneck (also called critical VOT), endogenously determined later. For clarity, we introduce the concept of VOT mass, denoted by b, as the number of households per unit of VOT. Given the uniformly distributed VOT, b is a constant, i.e., $b = N/\left(\overline{\alpha} - \underline{\alpha}\right)$. Let $\hat{N}_D$ and $\hat{N}_S$ be the total number of households in the downtown and suburb, respectively. We thus have $\hat{N}_D = b(\overline{\alpha} - \alpha^*)$ and $\hat{N}_S = b(\alpha^* - \underline{\alpha})$.

In the downtown, owing to the one-to-one correspondence between location *x* and VOT α, $N_D(x) = b(\overline{\alpha} - \alpha)$ holds. Substituting it into Eq. (11) to remove variable α yields the following vital second-order ordinary differential equation for $N_D(x)$:

$$\frac{d^2 N_D(x)}{dx^2} - \frac{t_0}{kb} N_D(x) = -\frac{t_0\overline{\alpha}}{k},\ x \in [0, a). \tag{12}$$

Note that the cumulative number of households at the CBD is 0, and that from the CBD to the bottleneck is the downtown population. Hence, the boundary conditions for Eq. (12) are

$N_D(0)=0$ and $N_D(a)=\hat{N}_D$. The analytical solution for $N_D(x)$ can thus be derived as

$$N_D(x)=c_1e^{rx}+c_2e^{-rx}+b\bar{\alpha},\ x\in[0,a), \tag{13}$$

where r, c_1, and c_2 are constants, with

$$r=\sqrt{\frac{t_0}{kb}},\quad c_1=\frac{b\bar{\alpha}-e^{ra}b\alpha^*}{e^{2ra}-1},\ \text{and}\ \ c_2=\frac{e^{ra}(-e^{ra}b\bar{\alpha}+b\alpha^*)}{e^{2ra}-1}. \tag{14}$$

From Eq. (14), $r>0$ and $c_2<0$ hold, while the sign of c_1 is undetermined. From Eq. (13), the residential density in the downtown is

$$n_D(x)=c_1re^{rx}-c_2re^{-rx},\ x\in[0,a). \tag{15}$$

Similarly, in the suburb, $N_S(x)=b(\bar{\alpha}-\alpha)$ holds due to the one-to-one correspondence between location and VOT. Combining it with Eq. (11) yields the following differential equation for $N_S(x)$:

$$\frac{d^2N_S(x)}{dx^2}-\frac{t_0}{kb}N_S(x)=-\frac{t_0\bar{\alpha}}{k},\ x\in[a,B], \tag{16}$$

subject to the boundary conditions: $N_S(a)=\hat{N}_D$, $N_S(B)=N$, and $p_S(B)=r_A$. Herein, r_A is the exogenous agricultural land rent, and $p_S(B)$, given by Eq. (11), is the land rental price at the city boundary *B*. These boundary conditions mean that the cumulative number of households from the CBD to the bottleneck is the downtown population, that from the CBD to the city boundary is the total city population, and the land rental price at the city edge equals the exogenous agricultural land rent.

Based on Eq. (16) and the associated boundary conditions, one can solve $N_S(x)$ as

$$N_S(x)=c_3e^{rx}+c_4e^{-rx}+b\bar{\alpha},\ \ x\in[a,B], \tag{17}$$

where c_3 and c_4 are constants, given by

$$c_3=e^{-ra}\frac{-b\alpha^*+\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}}{2}\ \ \text{and}\ \ c_4=e^{ra}\frac{-b\alpha^*-\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}}{2}. \tag{18}$$

From Eq. (18), $c_4<0$ holds, while the sign of c_3 is undetermined. Meanwhile, the city boundary *B* is solved as

$$B=\frac{1}{r}\ln\left(\frac{-b\underline{\alpha}+\sqrt{2r_A/(kr^2)}}{2c_3}\right). \tag{19}$$

One can thus obtain the suburban residential density as

$$n_S(x) = c_3 re^{rx} - c_4 re^{-rx},\ x \in [a, B]. \tag{20}$$

Thus far, we have solved the equilibrium household residential density in the downtown and suburban areas for a given critical VOT α^* at the bottleneck. In the following, we determine α^*. Substituting Eq. (11) into Eq. (8), one can rewrite the indirect utility function as

$$u_i(x,\alpha) = w(\alpha) - c_i(x,\alpha) - kn_i(x),\ i = D, S. \tag{21}$$

Note that at the equilibrium, the household utility should be continuous at the bottleneck, i.e., $u_D(a,\alpha^*) = u_S(a,\alpha^*)$. According to Eqs. (1), (2), (15), (20), and (21), the equilibrium condition for VOT α^* is

$$kr\frac{e^{2ra}+1}{e^{2ra}-1}b\alpha^* + kr\sqrt{(b\alpha^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)} + \frac{\eta}{q}b(\alpha^* - \underline{\alpha})\alpha^* = kr\frac{2e^{ra}b\bar{\alpha}}{e^{2ra}-1}. \tag{22}$$

Eq. (22) has a unique solution (if it exists) regarding α^*. This is because the left-hand side of Eq. (22) increases with α^* while the RHS is a constant.[9] Once α^* is solved by Eq. (22), the equilibrium cumulative number of households and residential density can then be determined by Eqs. (13), (15), (17), and (20).

Owing to the one-on-one correspondence between VOT α and residential location *x*, i.e., $N_i(x) = b(\bar{\alpha} - \alpha),\quad i = D, S$, we can determine households' residential locations as

$$\begin{cases} x_D(\alpha) = \dfrac{1}{r}\ln\left(\dfrac{-b\alpha + \sqrt{(b\alpha)^2 - 4c_1c_2}}{2c_1}\right), \alpha \in (\alpha^*, \bar{\alpha}], \\ x_S(\alpha) = \dfrac{1}{r}\ln\left(\dfrac{-b\alpha + \sqrt{(b\alpha)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}}{2c_3}\right), \alpha \in [\underline{\alpha}, \alpha^*], \end{cases} \tag{23}$$

where $x_D(\alpha)$ and $x_S(\alpha)$ are the residential locations of downtown and suburban households with VOT α, respectively.

From Eqs. (10), (15), (20), and (22), it is easy to show that both the residential density and the land rental price monotonically decrease with the distance from the CBD, and are

9 Eq. (22) might have no solution if the city population *N* is small enough or the bottleneck location *a* is too far from the CBD such that all population of the city is within the downtown area $[0, a]$. Such an extreme case is unrealistic and meaningless, and thus is omitted here.

discontinuous at the bottleneck (a downward jump). In contrast to the traditional monocentric urban model using static congestion with continuous residential density and land rental price, the bottleneck congestion here does have an important impact on household residential distribution and land market. The bottleneck congestion reduces the accessibility to the CBD for suburban households, thus lowering their willingness to pay for suburban land. Such discontinuity of the accessibility at the bottleneck causes the discontinuity of the land/housing price and residential density. Similar discontinuity has also been observed under cordon tolling schemes. For example, Mun et al. (2003) and De Lara et al. (2013) showed that the cordon toll causes a discontinuity in residential density and land rent through analytical or numerical methods.

So far, we have established the urban equilibrium, providing a firm basis for the investigation of bottleneck capacity expansion and congestion tolling issues in the next sections.

3. Bottleneck capacity expansion with no toll

Bottleneck congestion is a deadweight loss, and expanding bottleneck capacity is an efficient measure to reduce such a loss. In the long run, bottleneck capacity expansion may alter households' commuting schedules and residential location choices, thereby impacting urban spatial structure. In this section, we examine the distributional effects of bottleneck capacity expansion on different-income households and determine the optimal capacity under no toll. We also compare the results with and without considering the urban spatial effects.

3.1. Distributional effects of bottleneck capacity expansion on households

We first look at the distributional effects of bottleneck capacity expansion on different-income households, i.e., who gains (more) and who loses (more). According to Eqs. (1), (15), and (21), the equilibrium utility for downtown households with VOT α can be given by

$$u_D(\alpha) = w(\alpha) - t_0 \alpha x_D(\alpha) - kr\left(c_1 e^{r x_D(\alpha)} - c_2 e^{-r x_D(\alpha)}\right),\ \alpha \in (\alpha^*, \bar{\alpha}]. \tag{24}$$

From Eq. (24), one can derive

$$\frac{du_D(\alpha)}{dq} = \left(\frac{-b\alpha + \sqrt{(b\alpha)^2 - 4c_1c_2}}{2c_1} + \frac{2c_1}{-b\alpha + \sqrt{(b\alpha)^2 - 4c_1c_2}} \right) \frac{e^{ra}krb}{e^{2ra} - 1} \frac{d\alpha^*}{dq} > 0, \tag{25}$$

$$\frac{d^2u_D(\alpha)}{dqd\alpha}<0, \tag{26}$$

where $d\alpha^*/dq$, derived from Eq. (22), is larger than 0. Eq. (25) shows that all downtown households can gain from the bottleneck capacity expansion, but the marginal gains for the residents with higher VOTs are less than those with lower VOTs according to Eq. (26).

Similarly, the equilibrium utility for suburban households can be rewritten as

$$u_S(\alpha)=w(\alpha)-t_0\alpha x_S(\alpha)-\frac{\eta b(\alpha^*-\underline{\alpha})}{q}\alpha-kr\left(c_3e^{rx_S(\alpha)}-c_4e^{-rx_S(\alpha)}\right),\ \alpha\in[\underline{\alpha},\alpha^*]. \tag{27}$$

From Eq. (27), we have

$$\frac{du_S(\alpha)}{dq}=\left(\frac{e^{2ra}+1}{e^{2ra}-1}krb+\frac{\eta b(\alpha^*-\underline{\alpha})}{q}\right)\frac{\alpha}{\alpha^*}\frac{d\alpha^*}{dq}>0, \tag{28}$$

$$\frac{d^2u_S(\alpha)}{dqd\alpha}>0. \tag{29}$$

Eqs. (28) and (29) imply that all suburban households can gain from bottleneck capacity expansion, but the marginal gains increase linearly with VOT, i.e., the richer households gain more. The derivations of Eqs. (25) and (28) are tedious, and omitted here to save paper space.

Summarizing the above results leads to the following proposition.

Proposition 2. With the bottleneck capacity expansion,
(i) Each household benefits due to increased utility;
(ii) The marginal benefit is highest for mid-income households living at the bottleneck with VOT α^*, while it decreases for households with VOT farther away from α^*. Consequently, the marginal benefit is lowest for the richest households residing in the CBD or the poorest households residing at the city boundary.

Proposition 2 shows that the bottleneck capacity expansion leads to differential effects on heterogeneous households.[10] This can be explained as follows. From utility function Eq. (21), changes in household utility depend on changes in commuting costs and residential density

[10] Proposition 2 works when household VOT α is continuously distributed so that equilibrium utility $u(\alpha)$ is continuously differentiable with regard to α (except for the critical VOT). For a discrete VOT distribution, this conclusion may not hold and deserves a further investigation.

(or equivalently, land/housing area) for a given household income level. Bottleneck capacity expansion induces households to migrate outward due to reduced congestion, and thus households enjoy larger housing space. In the downtown, the lower-income households near the bottleneck can benefit more than the higher-income households near the CBD due to a greater increase in housing space. Conversely, in the suburb, the higher-income households near the bottleneck benefit more than the lower-income households near the city boundary. This is because although all suburban households benefit from increased housing space and reduced congestion costs after bottleneck capacity expansion, the higher-income households experience a greater reduction in congestion costs, leading to greater overall benefits.

The results of Proposition 2 are comparable to Arnott et al. (1994) and Takayama (2020). Arnott et al. (1994) showed that bottleneck capacity investment may benefit either higher-income or lower-income commuters more, without spatial dimension consideration. Takayama (2020) found that commuters nearer to the CBD gain but those farther away the CBD lose, with assumptions of Cobb-Douglas preference (for housing and numéraire goods) and a bottleneck location at the entrance of CBD. In our model, the bottleneck is at an arbitrary location of the corridor, resulting in greater gains for the households near the bottleneck. These results highlight the importance of considering interactions among bottleneck location, household preference, and urban spatial structure when evaluating the distributional impacts of bottleneck capacity expansion.

However, whether the richest or the poorest households benefit the least remains unclear from Proposition 2. For further check, we set $\alpha = \bar{\alpha}$ in Eq. (25) and $\alpha = \underline{\alpha}$ in Eq. (28), yielding

$$\frac{du_D(\alpha)}{dq}\Big|_{\alpha=\bar{\alpha}} = \frac{2e^{ra}krb}{e^{2ra}-1}\frac{d\alpha^*}{dq}, \tag{30}$$

$$\frac{du_S(\alpha)}{dq}\Big|_{\alpha=\underline{\alpha}} = \left(\frac{e^{2ra}+1}{e^{2ra}-1}krb + \frac{\eta b(\alpha^* - \underline{\alpha})}{q}\right)\frac{\underline{\alpha}}{\alpha^*}\frac{d\alpha^*}{dq}. \tag{31}$$

From Eqs. (30) and (31), the bottleneck location a and the current bottleneck capacity q play vital roles. Specifically, given other parameters, a small a results in a big α^* and thus a small $\underline{\alpha}/\alpha^*$, thus causing the RHS value of Eq. (31) to be smaller than that of Eq. (30). That is, the marginal gains for the poorest households are less than those for the richest households. Contrarily, if a is enough big, i.e., the bottleneck is close to the city boundary, then $\underline{\alpha}/\alpha^*$ approaches 1, and thus the RHS of Eq. (31) may be larger than that of Eq. (30), implying that the richest households gain less than the poorest households. On the other hand, the current bottleneck capacity q is also vital for the comparison of Eqs. (30) and (31), because the coefficient of $d\alpha^*/dq$ in Eq. (30) keeps unchanged with q, whereas the coefficient in Eq. (31) decreases with q. In Appendix G, we provide numerical examples to illustrate how these two factors affect the households' marginal benefits from capacity expansion. It turns out that if the bottleneck is located quite close to (far from) CBD, then the poorest (richest) households marginally benefit the least. If the bottleneck is positioned relatively in the middle of the corridor, then the present bottleneck capacity dominates the results: the marginal benefit of the richest (poorest) households is the lowest when the capacity level is low (high).

3.2. Optimal bottleneck capacity investment

In this section, we discuss the optimal investment in the bottleneck capacity to maximize social surplus, which is defined as the social benefit minus the bottleneck capacity investment cost. The social benefit includes the total utility of all households and the aggregate net land rent received by the local government (landlord). The social surplus and its components are in turn defined as follows.

Let TU be the total utility of all households, calculated as

$$TU = \hat{N}_D \int_{\alpha^*}^{\bar{\alpha}} \frac{1}{\bar{\alpha} - \alpha^*} u_D(\alpha) d\alpha + \hat{N}_S \int_{\underline{\alpha}}^{\alpha^*} \frac{1}{\alpha^* - \underline{\alpha}} u_S(\alpha) d\alpha \,, \tag{32}$$

where $u_D(\alpha)$ and $u_S(\alpha)$ are given by Eqs. (24) and (27), respectively. Note that $1/(\bar{\alpha} - \alpha^*)$ and $1/(\alpha^* - \underline{\alpha})$ are the probability density functions of VOT α in downtown and suburb, respectively. Therefore, the two terms on the RHS of Eq. (32) are the total utility of downtown and suburban households, respectively.

We assume that the local government owns the urban land and collects the land rent, but needs to pay the agricultural rent (or land opportunity cost) to the central government, as assumed in Kanemoto (1977). Let LR be the aggregate net land rent collected by the local government, expressed as

$$LR = \int_0^a (p_D(x) - r_A)dx + \int_a^B (p_S(x) - r_A)dx, \tag{33}$$

where r_A is the exogenous agricultural rent, land rent $p_i(x)$, $i = D, S$ is given by Eq. (10), and city boundary B is determined by Eq. (19).

Denote Φ as the social benefit, which is the sum of the total utility of all households and the aggregate net land rent, i.e.,

$$\Phi = TU + LR. \tag{34}$$

Let IC be the bottleneck capacity investment cost (including the construction and operating costs, etc.), and SS the social surplus. The optimal bottleneck capacity investment problem can then be represented as

$$\max_q \ SS = \Phi - IC. \tag{35}$$

In this paper, we assume that IC depends on q, and $dIC/dq > 0$ holds, i.e., a larger capacity requires a higher investment cost, and vice versa. From Eq. (35), the optimal bottleneck capacity can be obtained by first-order optimality condition $d\Phi/dq = dIC/dq$.

The total household utility increases as the bottleneck capacity increases since all households gain after bottleneck capacity expansion according to Proposition 2. We can further derive $dB/dq > 0$, $dLR/dq < 0$, and $d\Phi/dq > 0$ (see Appendix B). That is to say, as the bottleneck capacity increases, the city boundary expands, the aggregate net land rent decreases, but the social benefit (i.e., the sum of total household utility and aggregate net land rent) increases. These results mean that bottleneck capacity expansion poses different effects on different stakeholders: residents gain, while the local government loses due to the decreased net land rent besides the capacity investment cost. This occurs because some downtown households migrate to the suburb due to improved accessibility to the CBD. As a result, the city expands outward and land rent decreases in response to the increased bottleneck capacity. Similar phenomenon has also been observed in empirical studies such as Baum-Snow (2007), which found that new highways passing through the city reduce central-city population and make city more spread-out.

In addition, we can obtain the upper and lower bounds of $d\Phi/dq$ as

$$\frac{\eta b^2}{4q^2}\left((\alpha^*)^2 - \underline{\alpha}^2\right)\left(\alpha^* - \underline{\alpha}\right) < \frac{d\Phi}{dq} < \frac{\eta b^2}{2q^2}\left((\alpha^*)^2 - \underline{\alpha}^2\right)\left(\alpha^* - \underline{\alpha}\right). \tag{36}$$

The derivation of (36) is relegated to Appendix B. To explain the economic implications behind Eq. (36), we compute total bottleneck congestion cost, denoted as TC. According to Eq. (2), it can be calculated as

$$TC = \int_{\underline{\alpha}}^{\alpha^*} \frac{\eta b^2}{q}\left(\alpha^* - \underline{\alpha}\right)\alpha d\alpha = \frac{\eta b^2}{2q}\left((\alpha^*)^2 - \underline{\alpha}^2\right)\left(\alpha^* - \underline{\alpha}\right). \quad (37)$$

The marginal effect of capacity expansion on TC is

$$\frac{dTC}{dq} = -\frac{\eta b^2}{2q^2}\left((\alpha^*)^2 - \underline{\alpha}^2\right)\left(\alpha^* - \underline{\alpha}\right) + \frac{\eta b^2}{2q}\frac{d\left(\left((\alpha^*)^2 - \underline{\alpha}^2\right)\left(\alpha^* - \underline{\alpha}\right)\right)}{d\alpha^*}\frac{d\alpha^*}{dq}, \quad (38)$$

where $d\alpha^*/dq > 0$. The first term on the RHS represents the direct marginal reduction in TC caused by the increased capacity, while the second term represents the indirect marginal increase in TC caused by changes in urban spatial structure (i.e., the increased suburban population reflected by α^*). Eqs. (36) and (38) demonstrate that the marginal effect of capacity expansion on social benefit is less than its direct effect on total bottleneck congestion cost, but larger than half of that direct effect.

Summarizing the above, we have the following proposition.

Proposition 3. With the bottleneck capacity expansion,
(i) The city is expanded outward. The total utility of all households is increased, the aggregate net land rent is decreased, and the social benefit is increased.
(ii) The marginal increase in social benefit is less than the direct marginal reduction in total bottleneck congestion cost, but greater than its half.

3.3. Comparison of results with and without urban spatial structure consideration

Traditional bottleneck studies typically determine optimal bottleneck capacity by a naïve criterion: the marginal capacity expansion cost equals the marginal reduction in bottleneck congestion cost. Besides, they implicitly assumed household residential locations are exogenously given and thus the urban spatial structure remains unchanged as the capacity expands. However, such a naïve optimality condition is invalidated when considering urban spatial structure. This is because bottleneck capacity expansion affects household residential location (thus commuting distance), household utility, and land rent, in addition to the bottleneck congestion cost. Next, we assess the distorted bottleneck capacity investment due to ignoring changes in urban spatial structure.

To do so, we represent $\tilde{q}$ as the initial bottleneck capacity from which it is expanded, and $\tilde{\alpha}^*$ as the initial critical VOT at the bottleneck. If the change of urban spatial structure is ignored, then the critical VOT does not vary with bottleneck capacity expansion and thus remains a constant $\tilde{\alpha}^*$. From Eq. (38), the marginal reduction in total bottleneck congestion cost, denoted by Ω, is

$$\Omega = \frac{\eta b^2}{2q^2}\left((\tilde{\alpha}^*)^2 - \underline{\alpha}^2\right)\left(\tilde{\alpha}^* - \underline{\alpha}\right), \text{ for } q \geq \tilde{q}. \tag{39}$$

The marginal reduction Ω measures the marginal benefit of capacity expansion to society without considering urban spatial structure.

If urban spatial structure is explicitly treated as in this paper, the marginal benefit of capacity expansion to society will be $d\Phi/dq$. Combining Eqs. (36) and (39), one can judge that for any initial capacity $\tilde{q}$, there must exist a $\Delta\tilde{q}$, satisfying

$$\frac{d\Phi}{dq} < \Omega, \text{ for } q \in \left[\tilde{q}, \tilde{q} + \Delta\tilde{q}\right]. \tag{40}$$

Eq. (40) shows that at the initial stage of bottleneck capacity expansion (i.e., $q \in \left[\tilde{q}, \tilde{q} + \Delta\tilde{q}\right]$), ignoring change in urban spatial structure will overestimate the marginal benefit of capacity expansion to society. However, since α^* in $d\Phi/dq$ increases with q (see Eq. (B.17)), $d\Phi/dq > \Omega$ may also occur when capacity has been expanded to a quite high level (i.e., $q \in \left[\tilde{q} + \Delta\tilde{q}, \infty\right]$), meaning that ignoring change in urban spatial structure may also underestimate the marginal benefit to society. As a result, ignoring urban spatial structure may cause underinvestment or overinvestment in bottleneck capacity, depending on the marginal cost of capacity expansion.

Proposition 4. Ignoring the effects of the bottleneck capacity expansion on urban spatial structure may overestimate or underestimate its marginal benefit to society, depending on the bottleneck capacity expanded. Consequently, it may lead to overinvestment or underinvestment in the optimal bottleneck capacity.

Proposition 4 is further illustrated in Appendix G by a numerical example. It can be explained as follows. On the one hand, with capacity expansion, ignoring the change in urban spatial structure will overestimate its benefit to land market, because capacity expansion actually reduces aggregate net land rent (see Proposition 3(i)). On the other hand, ignoring urban spatial structure will underestimate the benefit to residents, especially for downtown residents since their utility is increased after relocation (see Proposition 2(i)). In the initial stage of capacity expansion, the overestimated effect on land market dominates the underestimated effect on households' utility. This result will be reversed when the capacity is significantly expanded compared to the initial state. It should be mentioned that the land ownership plays an important role in the capacity investment decision through affecting the social benefit.[11]

Capacity investment decision biases have also been studied in traditional urban models with static congestion and homogeneous households. Our results are comparable to these studies. For example, Kanemoto (1977) showed that the naïve cost-benefit rule (i.e., the marginal saving in transportation cost equals the marginal cost of road expansion) tends to overinvest in road capacity near the CBD, but may overinvest or underinvest near the city boundary. With dynamic congestion and heterogeneous households, we find that if the naïve investment rule is adopted and change in urban spatial structure is ignored, either overinvestment or underinvestment in bottleneck capacity occurs.

4. Bottleneck capacity expansion with congestion tolling

In the previous section, we analyzed the impact of bottleneck capacity expansion without toll, demonstrating its ability to enhance household utility and overall social welfare. However, capacity expansion cannot eliminate congestion externalities, leading to resource wastage. Well-designed congestion tolling schemes, as implemented in Singapore, London, and Hong Kong, can (partially) mitigate these externalities. Naturally, a hybrid approach combining capacity expansion (a supply-side measure) and congestion tolling (a demand-side measure)

[11] It is assumed in this paper that the land is owned by the local government and all net land rents go to social benefit Φ. If the land is owned by absentee landlords and a proportion θ $(0<\theta\leq 1)$ of net land rents are kept by them, then only a proportion $1-\theta$ of net land rents are included in the social benefit, i.e., $\Phi = TU + (1-\theta)LR$. In this case, $d\Phi/dq$ will be larger than the value obtained in this paper since dLR/dq is negative. However, whether Eq. (40) and Proposition 4 still hold is not sure.

is expected to address the congestion issue efficiently. Moreover, toll revenue can be used for capacity expansion, alleviating the government's financial burden. Therefore, this section investigates bottleneck capacity expansion problems under congestion tolling, including the first-best (dynamic) toll and the second-best (flat) toll.

4.1. Bottleneck capacity expansion under first-best toll

4.1.1. First-best toll and critical VOT at bottleneck

The first-best dynamic toll is imposed at the bottleneck since congestion externality only occurs here. Downtown commuters are exempt from congestion tolls because they don't pass through the bottleneck, and thus their commuting cost is given by Eq. (1). Denote $\tau_{FB}(\alpha)$ as the first-best toll imposed on the suburban commuters with VOT α, where the subscript "*FB*" means the first-best, and $c_B(\alpha)$ as the bottleneck congestion cost (queuing delay plus schedule delay) under the first-best toll. The commuting cost for suburban households includes bottleneck congestion cost, first-best toll, and free-flow travel time cost, given as

$$c_{S,FB}(x,\alpha) = c_B(\alpha) + \tau_{FB}(\alpha) + t_0 x\alpha . \tag{41}$$

Under the first-best toll, it is easy to show that households still reside along the corridor in a decreasing VOT order from the CBD outward. Given the critical VOT α^*_{FB} at the bottleneck, residential density and household utility in downtown and suburb can be similarly derived as in the no-toll case (referring to Section 2.3). Since household utility is continuous at the bottleneck, the equilibrium condition used to determine α^*_{FB} can be given by

$$kr\frac{e^{2ra}+1}{e^{2ra}-1}b\alpha^*_{FB} + kr\sqrt{(b\alpha^*_{FB})^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)} + c_B(\alpha^*_{FB}) + \tau_{FB}(\alpha^*_{FB}) = kr\frac{2e^{ra}b\bar{\alpha}}{e^{2ra}-1}. \tag{42}$$

We now determine the first-best toll $\{\tau_{FB}(\alpha) \mid \alpha^*_{FB} \le \alpha \le \bar{\alpha}\}$. Denote Φ_{FB} as the social benefit under the first-best toll, including the total household utility TU_{FB}, aggregate net land rent LR_{FB}, and toll revenue TR_{FB} collected by the government. Given the bottleneck capacity, the first-best toll aims to maximize the social benefit, formulated as

$$\max_{\{\tau_{FB}(\alpha)\}} \Phi_{FB} = TU_{FB} + LR_{FB} + TR_{FB} . \tag{43}$$

Based on the maximization problem (43), we can obtain the following proposition (its proof is relegated to Appendix C).

Proposition 5. The social optimum is achieved exactly under the first-best toll that completely eliminates the queues behind the bottleneck.

Obviously, the urban system is distorted under the unpriced bottleneck congestion due to two important sources: bottleneck congestion externality caused by suburban commuters and mismatched population allocation between downtown and suburb. Proposition 5 implies that these distortions can be simultaneously eliminated by the first-best toll.

From Proposition 5 and its proof, the first-best toll $\tau_{FB}(\alpha)$ and the corresponding bottleneck congestion cost $c_B(\alpha)$ can, respectively, be given by

$$\tau_{FB}(\alpha) = \frac{\eta b}{2q}(\alpha^2 - \underline{\alpha}^2),\ \alpha \in [\underline{\alpha}, \alpha^*_{FB}], \tag{44}$$

$$c_B(\alpha) = \frac{\eta b}{q}(\alpha^*_{FB} - \alpha)\alpha,\ \alpha \in [\underline{\alpha}, \alpha^*_{FB}]. \tag{45}$$

Plugging these equations into Eq. (42) yields the equilibrium condition that defines the critical VOT α^*_{FB} at the bottleneck.

4.1.2. Distributional effects of bottleneck capacity expansion on households

We now investigate the distributional effects of bottleneck capacity expansion on different households under the first-best toll. Following the same procedure as in the no-toll case, we can show that under the first-best toll, all downtown households benefit from the bottleneck capacity expansion, but higher-income households benefit less. We now examine the effects of capacity expansion on suburban households. According to Eqs. (20), (21), (44), and (45), the equilibrium utility of suburban households is

$$u_{S,FB}(\alpha) = w(\alpha) - t_0\alpha x_S(\alpha) - \frac{\eta b}{q}(\alpha^*_{FB} - \alpha)\alpha - \frac{\eta b}{2q}(\alpha^2 - \underline{\alpha}^2) - kr\left(c_3 e^{rx_S(\alpha)} - c_4 e^{-rx_S(\alpha)}\right), \tag{46}$$

where $x_S(\alpha)$ is the residential location of suburban households, given by Eq. (23).

Based on Eq. (46), we can derive

$$\frac{du_{S,FB}(\alpha)}{dq} > 0, \tag{47}$$

$$\frac{d^2u_{S,FB}(\alpha)}{dqd\alpha} = \left(\frac{2\alpha_{FB}^* - 2\alpha}{(\alpha_{FB}^*)^2 - \underline{\alpha}^2} \left(krb\frac{e^{2ar}+1}{e^{2ar}-1} + \frac{krb^2\alpha_{FB}^*}{\sqrt{(b\alpha_{FB}^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}} + \frac{\eta b}{q}\alpha_{FB}^* \right) - \frac{krb^2}{\sqrt{(b\alpha_{FB}^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}} - \frac{\eta b}{q} \right) \frac{d\alpha_{FB}^*}{dq}, \tag{48}$$

where $d\alpha_{FB}^*/dq$ is greater than 0 and independent of α (see Eq. (D.3)).

Eq. (47) shows that all suburban commuters gain from bottleneck capacity expansion. However, who gains more is unclear since the sign of Eq. (48) is ambiguous. Note that the RHS of Eq. (48) decreases with α, and is positive when $\alpha = \underline{\alpha}$ but negative when $\alpha = \alpha_{FB}^*$. Thereby, there exists a unique VOT $\hat{\alpha}_{FB}(q) \in (\underline{\alpha}, \alpha_{FB}^*)$ such that $d^2u_{S,FB}/dqd\alpha > 0$ for $\alpha \in [\underline{\alpha}, \hat{\alpha}_{FB})$ and $d^2u_{S,FB}/dqd\alpha < 0$ for $\alpha \in (\hat{\alpha}_{FB}, \alpha_{FB}^*]$. This implies that among relatively low-income suburban households (i.e., $\alpha \in [\underline{\alpha}, \hat{\alpha}_{FB}]$), higher-VOT households marginally benefit more from the bottleneck capacity expansion, while among relatively high-income suburban households (i.e., $\alpha \in [\hat{\alpha}_{FB}, \alpha_{FB}^*]$), lower-VOT households marginally benefit more. As a result, relatively mid-income households with VOT $\hat{\alpha}_{FB}$ benefit most. Summarizing the above, we have the following proposition.

Proposition 6. With the bottleneck capacity expansion subject to the first-best toll,
(i) Each household benefits due to increased utility.
(ii) The marginal benefit is highest for relatively mid-income households living at a suburban location with VOT $\hat{\alpha}_{FB}$, while it decreases for the households with VOT farther from $\hat{\alpha}_{FB}$. Consequently, the marginal benefit is lowest for the richest households residing in the CBD or the poorest households residing at the city boundary.

Proposition 6 reveals that under the first-best toll, the households living in a suburban location (not the bottleneck) gain the most from the capacity expansion, which significantly differs from the no-toll case where households living at the bottleneck gain the most. This is because by comparing Eqs. (2), (44), and (45), among different-income suburban households, the difference in marginal reduction of their congestion costs due to capacity expansion (i.e., $\partial^2 c_S(x,\alpha)/\partial q\partial\alpha$) becomes smaller after implementing the first-best toll. The utility increment from the decreased bottleneck congestion cost no more dominates that from the increased house size. Suburban households with larger VOTs benefit more from the decreased bottleneck congestion cost, but less from the increased house size than those with lower VOTs. The trade-off between these two factors leads the relatively mid-income suburban households living at a suburban location to benefit the most from capacity expansion. Numerical examples in Appendix G confirms these findings. It further discloses that if the bottleneck location is quite far from (close to) the CBD, the marginal benefit is lowest for the richest (poorest) households. If the bottleneck location is relatively in the middle of the city, the result depends on the current bottleneck capacity.

4.1.3. Optimal bottleneck capacity investment and self-financing property

Under the first-best toll, the optimal bottleneck capacity investment aims to maximize social surplus, formulated as

$$\max_{q} SS_{FB} = \Phi_{FB} - IC = TU_{FB} + LR_{FB} + TR_{FB} - IC, \tag{49}$$

where SS_{FB} is the social surplus under the first-best toll. The total household utility TU_{FB} and aggregate net land rent LR_{FB} are defined by Eqs. (32) and (33), respectively. According to Eq. (44), toll revenue TR_{FB} collected by the local government can be given by

$$TR_{FB} = \int_{\underline{\alpha}}^{\alpha_{FB}^*} \frac{\eta b^2}{2q}(\alpha^2 - \underline{\alpha}^2) d\alpha = \frac{\eta b^2}{q}\left(\frac{1}{6}(\alpha_{FB}^*)^3 - \frac{1}{2}\underline{\alpha}^2 \alpha_{FB}^* + \frac{1}{3}\underline{\alpha}^3\right). \tag{50}$$

Similar to the no-toll case, we can derive that with the bottleneck capacity expansion subject to the first-best toll, the city boundary moves outward, the aggregate net land rent decreases, and the total household utility increases. Besides, we can derive

$$\frac{d\Phi_{FB}}{dq} = \frac{\eta b^2}{q^2}\left(\frac{1}{6}(\alpha_{FB}^*)^3 - \frac{1}{2}\alpha_{FB}^* \underline{\alpha}^2 + \frac{1}{3}\underline{\alpha}^3\right) > 0\,. \tag{51}$$

The detailed derivation of Eq. (51) is relegated to Appendix D. Eq. (51) shows that the social benefit increases after the capacity expansion under the first-best toll. To further disclose the economic implication behind Eq. (51), we define TC_{FB} as the total bottleneck congestion cost (excluding the congestion toll since it is merely a payment transfer within the city). From Eq. (45), TC_{FB} can be given by

$$TC_{FB} = \int_{\underline{\alpha}}^{\alpha_{FB}^*} \frac{\eta b^2}{q}\left(\alpha_{FB}^* - \alpha\right) \alpha\, d\alpha = \frac{\eta b^2}{q}\left(\frac{1}{6}(\alpha_{FB}^*)^3 - \frac{1}{2}\alpha_{FB}^* \underline{\alpha}^2 + \frac{1}{3}\underline{\alpha}^3\right). \tag{52}$$

The marginal reduction in TC_{FB} after the capacity expansion is thus derived as

$$\frac{dTC_{FB}}{dq} = -\frac{\eta b^2}{q^2}\left(\frac{1}{6}(\alpha_{FB}^*)^3 - \frac{1}{2}\alpha_{FB}^*\underline{\alpha}^2 + \frac{1}{3}\underline{\alpha}^3\right) + \frac{\eta b^2}{q}\frac{1}{6}\frac{d\left((\alpha_{FB}^*)^3 - 3\alpha_{FB}^*\underline{\alpha}^2 + 2\underline{\alpha}^3\right)}{d\alpha_{FB}^*}\frac{d\alpha_{FB}^*}{dq}, \quad (53)$$

where $d\alpha_{FB}^*/dq > 0$. The first term on the RHS of Eq. (53) is the direct marginal reduction in TC_{FB} caused by the increased capacity, while the second term represents the indirect marginal increase in TC_{FB} caused by the change in urban spatial structure. Eqs. (51) and (53) demonstrate that under the first-best toll, the marginal effect of capacity expansion on social benefit is exactly equal to its direct effect on total congestion cost.

On the other hand, Eq. (51) can be rewritten as $d\Phi_{FB}/dq = TR_{FB}/q$ according to Eq. (50). Note that with optimal bottleneck capacity investment, the first-order optimality condition for optimization problem (49) requires $d\Phi_{FB}/dq = dIC/dq$. Combining them yields

$$\frac{TR_{FB}}{IC} = \frac{dIC}{dq}\frac{q}{IC}. \quad (54)$$

Eq. (54) reveals a striking result about the bottleneck capacity investment. It shows that under the first-best toll, the ratio of toll revenue to the optimal bottleneck capacity investment cost is equal to the elasticity of capacity investment cost with respect to capacity.

In light of the above, we have the following proposition.

Proposition 7. With the bottleneck capacity expansion subject to the first-best toll,
(i) The city is expanded outward. The total household utility is increased, the aggregate net land rent is decreased, and the social benefit is increased.
(ii) The marginal increase in the social benefit is equal to the direct marginal reduction in total bottleneck congestion cost.
(iii) With the optimal capacity, the ratio of the toll revenue to the capacity investment cost equals the elasticity of investment cost with respect to capacity. With increasing (decreasing) return to scale, the capacity investment financed by the toll revenue yields a surplus (a deficit). With constant return to scale, self-financing holds.

Proposition 7(i) is straightforward, similar to the no-toll case. However, Proposition 7(ii) shows a distinct result from the no-toll case. Specifically, under the first-best toll, the increase in the social benefit due to capacity expansion is exactly equal to the direct reduction in total bottleneck congestion cost. However, under the no toll, the former is less than the latter (see Proposition 3). Proposition 7(iii) presents a remarkable result regarding the self-financing property when the effects of the bottleneck capacity expansion on the urban spatial structure are taken into account. The seminal work on bottleneck capacity investment by Arnott et al. (1993) showed that the ratio of the first-best toll revenue to the optimal capacity construction cost equals the elasticity of construction cost with respect to capacity. However, they assumed homogeneous commuters and did not consider the urban spatial dimension. Arnott and Kraus (1995) further showed that such a self-financing principle holds for the first-best toll with heterogeneous commuters but an exogenous urban spatial structure. This paper further extends such an important property to the context of heterogeneous commuters and endogenous urban spatial structure.

It should be pointed out that the self-financing result has a close relationship with the Henry George Theorem. In this paper, we achieve the self-financing outcome within an urban spatial framework. This result stems from implementing a constant return to scale, characterized by the homogeneity of degree zero, in the bottleneck congestion function (see Eq. (2)). If the congestion function exhibits variable returns to scale, the Henry George Theorem is applicable, as detailed in Berglas and Pines (1981), Kanemoto (1984), and Arnott (2004).

4.1.4. Comparison of results with and without urban spatial structure consideration

We now look at the investment decision biases of ignoring urban spatial structure. Similar to the no-toll case, we consider an initial urban equilibrium state from which the bottleneck capacity is expanded, with initial capacity $\tilde{q}$ and critical VOT $\tilde{\alpha}^*_{FB}$ at the bottleneck. If the urban spatial effect is ignored, then the marginal benefit of capacity expansion to society Ω_{FB} (i.e., the marginal reduction in total bottleneck congestion cost) can be given by

$$\Omega_{FB} = \frac{\eta b^2}{q^2}\left(\frac{1}{6}(\tilde{\alpha}^*_{FB})^3 - \frac{1}{2}\tilde{\alpha}^*_{FB}\underline{\alpha}^2 + \frac{1}{3}\underline{\alpha}^3\right),\ q \geq \tilde{q} \quad (55)$$

where $\tilde{\alpha}^*_{FB}$ remains invariable with regard to capacity q because the urban spatial structure is exogenously given.

If the spatial effects are considered, the marginal benefit of the capacity expansion to society is $d\Phi_{FB}/dq$ (see Eq. (51)), in which α^*_{FB} increases with q. Specifically, $\alpha^*_{FB}(q) = \tilde{\alpha}^*_{FB}$ when $q = \tilde{q}$, and $\alpha^*_{FB}(q) > \tilde{\alpha}^*_{FB}$ when $q > \tilde{q}$. Combining Eqs. (51) and (55) yields

$$\frac{d\Phi_{FB}}{dq} \geq \Omega_{FB}, \text{ for } q \geq \tilde{q}. \tag{56}$$

Inequality (56) takes "=" when $q = \tilde{q}$. Eq. (56) implies that the marginal benefit of capacity expansion to society with urban spatial structure consideration is larger than that without such a consideration. Besides, it is easy to show that Φ_{FB} is concave with respect to q, and $d\,\Omega_{FB}/dq < 0$. We immediately have the following proposition.

Proposition 8. Under the first-best toll, ignoring the effects of bottleneck capacity expansion on urban spatial structure underestimates its marginal benefit to society, thus leading to underinvestment in the optimal bottleneck capacity.

Proposition 8 shows that under the first-best toll, ignoring urban spatial effects and using the naïve cost-benefit method lead to a biased assessment of the benefit of capacity expansion and underinvestment in the bottleneck capacity, causing losses to society.[12] This occurs because the underestimated household utility outperforms the overestimated land rent revenue. This result differs from the no-toll case, in which ignoring urban spatial effects may cause overinvestment or underinvestment in the bottleneck capacity, as presented in Proposition 5.

4.2. Bottleneck capacity expansion under second-best toll

4.2.1. Second-best toll and critical VOT at bottleneck

Due to political and technical issues as pointed out by some previous studies (van den Berg and Verhoef, 2011; Lindsey, 2012), the first-best dynamic toll is hardly put into operation in practice. As the second-best choice, a flat toll scheme is more prevalent in real applications (Mun et al., 2003; De Lara et al., 2013; Li et al., 2014).

---

[12] If the land is owned by absentee landlords and a proportion $\theta\ (0 < \theta \leq 1)$ of the land rents are kept by them, then social benefit $\Phi_{FB} = TU_{FB} + (1-\theta)LR_{FB}$. In this case, $d\Phi_{FB}/dq$ is larger than the RHS of Eq. (51), and thus Eq. (56) and Proposition 8 still hold.

Denote τ_{SB} as the second-best flat toll, and α^*_{SB} as the critical VOT at the bottleneck, where the subscript "*SB*" means the second-best. Under the second-best toll, the commuting cost for downtown households is given by Eq. (1). According to the bottleneck theory, given the residential distribution of suburban commuters, a flat toll will not change their departure patterns. Therefore, the commuting cost for suburban households is given by Eq. (2) with an additional flat toll τ_{SB}. Again, it can be shown that residents sort along the corridor outward in a decreasing order of VOT. Given the value of α^*_{SB}, the residential density and household utility can be obtained similar to the no-toll case (see Section 2.3). From the continuity of household utility, α^*_{SB} can be determined by Eq. (22) with an additional flat toll τ_{SB} on its left-hand side.

We now determine the second-best toll τ_{SB}. Denote Φ_{SB} as the social benefit under the second-best toll. It consists of total household utility TU_{SB}, aggregate net land rent LR_{SB}, and toll revenue TR_{SB} (i.e., the flat toll multiplied by the suburban population). Given the bottleneck capacity *q*, the second-best toll aims to maximize social benefit, expressed as

$$\max_{\tau_{SB}} \Phi_{SB} = TU_{SB} + LR_{SB} + TR_{SB}. \quad (57)$$

The optimality condition of the maximization problem (57) yields

$$\tau_{SB} = \frac{\eta b}{2q}\left((\alpha^*_{SB})^2 - \underline{\alpha}^2\right) = \frac{\eta b(\alpha^*_{SB} - \underline{\alpha})}{q}\frac{\alpha^*_{SB} + \underline{\alpha}}{2}. \quad (58)$$

The derivation of Eq. (58) is relegated to Appendix E. The implications of Eq. (58) are explained below. At equilibrium, the household VOT in the suburb follows a uniform distribution over $[\underline{\alpha}, \alpha^*_{SB}]$ because the household VOT in the entire city follows a uniform distribution over $[\underline{\alpha}, \overline{\alpha}]$. Therefore, $(\alpha^*_{SB} + \underline{\alpha})/2$ represents the mean value of suburban VOTs. From Eq. (2), the bottleneck congestion cost (excluding toll) for a suburban commuter with VOT α is $\eta b(\alpha^*_{SB} - \underline{\alpha})\alpha/q$. Hence, Eq. (58) implies that at equilibrium, the optimal flat toll for suburban commuters is equal to their average congestion cost.

Based on Eq. (58), the total toll revenue can thus be calculated as

$$TR_{SB} = \hat{N}_S \tau_{SB} = \frac{\eta b^2}{2q}\left((\alpha^*_{SB})^2 - \underline{\alpha}^2\right)\left(\alpha^*_{SB} - \underline{\alpha}\right). \quad (59)$$

It should be mentioned that unlike the first-best toll which can completely eliminate congestion externality and population distortion between downtown and suburb, the second-best flat toll can partially lower congestion externality through population adjustment between downtown and suburban areas.

4.2.2. Distributional effects of bottleneck capacity expansion on households

Again, similar to the no-toll case, one can show that all the downtown households benefit from bottleneck capacity expansion, but the marginal benefit for richer households is lower. We now look at the utility changes of suburban households after the capacity expansion. We can easily obtain the equilibrium utility $u_{S,SB}(\alpha)$ for suburban households under the second-best toll, and derive $du_{S,SB}(\alpha)/dq > 0$ and

$$\frac{d^2 u_{S,SB}(\alpha)}{d\alpha dq} = \frac{2\xi_{SB}}{3\alpha_{SB}^* + \underline{\alpha}} \frac{d\alpha_{SB}^*}{dq}, \tag{60}$$

where $d\alpha_{SB}^*/dq > 0$, and ξ_{SB} is given by

$$\xi_{SB} = krb\frac{e^{2ar}+1}{e^{2ar}-1} - \frac{1}{2}\frac{krb^2(\alpha_{SB}^* + \underline{\alpha})}{\sqrt{(b\alpha_{SB}^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}} + \frac{3}{2}\frac{\eta b}{q}\left(\alpha_{SB}^* - \underline{\alpha}\right). \tag{61}$$

It shows that under the second-best toll, all suburban households gain from bottleneck capacity expansion. However, the comparison of the marginal benefits among them is ambiguous since the sign of ξ_{SB} is indefinite. Note that ξ_{SB} is independent of VOT α. Therefore, Eq. (60) shows that the marginal benefit for suburban households may increase or decrease linearly with VOT α, relying on the sign of ξ_{SB}. Summarizing these results, we have the following proposition.

Proposition 9. With the bottleneck capacity expansion subject to the second-best toll,
(i) Each household benefits due to increased utility.
(ii) The marginal benefits are differentiated across different-income households: as $\xi_{SB} > 0$, the marginal benefit is highest for relatively mid-income households living at the bottleneck with VOT α_{SB}^*, whereas it decreases for households with VOT farther away from α_{SB}^*, and the richest or poorest households benefit least; as $\xi_{SB} < 0$, households' marginal benefits

monotonically decrease with their VOTs, i.e., the poorest households benefit most while the richest households benefit least.

As stated in the no-toll case, among the suburban households, who benefits more from the capacity expansion depends on the tug-of-war between the increased housing space and the decreased bottleneck congestion cost. Proposition 9 indicates that such a tug-of-war depends on the sign of ξ_{SB}. Unlike the no toll and first-best toll cases, the poorest households may gain most from the capacity expansion under the second-best toll. Numerical study in Appendix G further illustrates that if the bottleneck is located quite far from the CBD, then ξ_{SB} takes a negative value. As a result, households' marginal benefits from capacity expansion decrease with their VOTs, i.e., the poorest (richest) households gain the most (least).

4.2.3. Optimal bottleneck capacity investment and self-financing property

We now determine the optimal bottleneck capacity investment and check the self-financing property under the second-best toll. The optimal capacity investment problem to maximize the social surplus can be formulated as

$$\max_{q} SS_{SB} = \Phi_{SB} - IC, \tag{62}$$

where SS_{SB} is the social surplus under the second-best toll, and the social benefit Φ_{SB} is defined by Eq. (57).

Similarly, one can show that under the second-best toll, bottleneck capacity expansion leads to an outward movement of the city boundary, a decrease in the aggregate net land rent, but an increase in the total household utility and the social benefit. The marginal increase in social benefit can be calculated as

$$\frac{d\Phi_{SB}}{dq} = \frac{\eta b^2}{2q^2}\left((\alpha_{SB}^*)^2 - \underline{\alpha}^2\right)\left(\alpha_{SB}^* - \underline{\alpha}\right) > 0\,. \tag{63}$$

Define TC_{SB} as the total bottleneck congestion cost (excluding congestion toll). It can easily be derived that the marginal reduction in TC_{SB} after capacity expansion is

$$\frac{dTC_{SB}}{dq} = -\frac{\eta b^2}{2q^2}\left((\alpha_{SB}^*)^2 - \underline{\alpha}^2\right)\left(\alpha_{SB}^* - \underline{\alpha}\right) + \frac{\eta b^2}{2q}\frac{d\left((\alpha_{SB}^*)^2 - \underline{\alpha}^2\right)\left(\alpha_{SB}^* - \underline{\alpha}\right)}{d\alpha_{SB}^*}\frac{d\alpha_{SB}^*}{dq}\,. \tag{64}$$

The two terms on the RHS of Eq. (64) are, respectively, the direct marginal reduction in total congestion cost caused by the increased capacity and the indirect marginal increase in total congestion cost due to the changed urban spatial structure. The first term on the RHS of Eq. (64) and the term on the RHS of Eq. (63) are the same, meaning that under the second-best toll, the direct effect of bottleneck capacity expansion on total bottleneck congestion cost is exactly equal to its marginal effect on social benefit.

With the optimal capacity investment, the first-order condition for the maximization problem (62) necessitates $d\Phi_{SB}/dq = dIC/dq$. Combining it with Eqs. (59) and (63) yields

$$\frac{TR_{SB}}{IC} = \frac{dIC}{dq}\frac{q}{IC}, \tag{65}$$

which demonstrates that under the second-best toll, the ratio of toll revenue to the optimal capacity investment cost still equals the capacity elasticity of the investment cost.

We summarize the above results and obtain the following proposition.

Proposition 10. With the bottleneck capacity expansion subject to the second-best toll,
(i) The city is expanded outward. The total household utility is increased, the aggregate net land rent is decreased, and the social benefit is increased.
(ii) The marginal increase in the social benefit is equal to the direct marginal reduction in total bottleneck congestion cost.
(iii) With the optimal capacity, the ratio of the toll revenue to the capacity investment cost equals the elasticity of investment cost with respect to capacity. With increasing (decreasing) return to scale, the capacity investment financed by the toll revenue yields a surplus (a deficit). With constant return to scale, self-financing holds.

Proposition 10 shows that the self-financing principle remains valid under the second-best context. This finding extends the study of Arnott et al. (1993), in which self-financing property holds under the flat toll, but for homogeneous commuters without urban spatial structure consideration. Arnott and Kraus (1995) further considered household heterogeneity in absence of urban spatial effects and concluded that the self-financing property no more applied to the uniform (flat) toll. By contrast, our work indicates that if the urban spatial structure (or household relocation behavior) and household heterogeneity are explicitly treated, the self-financing principle still holds for the optimal flat toll.

4.2.4. Comparison of results with and without urban spatial structure consideration

We now briefly discuss the capacity investment biases when ignoring urban spatial structure. Let $\tilde{q}$ represent the initial bottleneck capacity, and $\tilde{\omega}^{*}_{SB}$ the critical VOT at the bottleneck. From Eq. (64), we can easily obtain the marginal benefit Ω_{SB} of capacity investment to society when the urban spatial structure is ignored. Further, we can show $d\Phi_{SB}/dq \geq \Omega_{SB}$ for $q \geq \tilde{q}$, similar to the procedure adopted in the first-best toll case. These findings can be summarized as follows.

Proposition 11. Under the second-best toll, ignoring the urban spatial effects will underestimate the marginal benefit of bottleneck capacity expansion to society, leading to underinvestment in the optimal bottleneck capacity.

Propositions 8 and 11 show that regardless of the first-best dynamic toll or the second-best flat toll, underinvestment in bottleneck capacity would occur due to ignoring the urban spatial structure.[13]

4.3. Comparison of urban spatial structures with optimal bottleneck capacities under no toll, first-best, and second-best tolls

Thus far, we have discussed the optimal bottleneck capacity investment issues under no toll, first-best toll, and second-best toll. For ease of presentation, the cities with optimal capacities under no toll, first-best toll, and second-best toll are referred to as no-toll optimal, first-best, and second-best cities, respectively. Comparing their urban spatial structures, we obtain the following proposition, and its proof is given in Appendix F.

Proposition 12. With constant return to scale for bottleneck capacity investment,
(i) For a given household, its residential location in the first-best city is farther from the CBD than in the second-best city, making the first-best city longer than the second-best city.
(ii) The downtown residential density and land rent under the first-best city are lower than

[13] Similar to the first-best toll (see Footnote 12), if a portion of land rents go to the absentee landlords under the second-best toll, Proposition 11 still holds.

those under the second-best city. The results are reversed for the suburban area.

(iii) As $\underline{\alpha} = 0$, the no-toll optimal city lies between the first-best and second-best cities, in terms of residential location, city length, residential density, and land rent.

The results of Proposition 12 can be explained as follows. Given the suburban population and bottleneck capacity, the first-best dynamic toll can eliminate bottleneck queues and thus reduce the commuting cost of suburban commuters. Hence, the first-best toll attracts some households to live in the suburban area. As a result, the city boundary expands. By contrast, the second-best flat toll does not change the departure patterns of bottleneck users but increases their commuting costs. In essence, the second-best toll alleviates the bottleneck congestion by driving some suburban commuters to live downtown, and thus the city shrinks. As a result, the first-best city is longer than the second-best city, albeit they have different optimal capacities. For a special case of $\underline{\alpha} = 0$, we analytically prove that the no-toll optimal city lies between the first-best and second-best cities in terms of residential location, city length, residential density, and land rent. Numerical results in Appendix G indicate that the no-toll optimal city still lies between the first-best and second-best cities when $\underline{\alpha}$ is relatively small, which is consistent with the special case of $\underline{\alpha} = 0$. However, when $\underline{\alpha}$ is relatively large (approaching $\bar{\alpha}$), the no-toll optimal city has the lowest (highest) downtown (suburban) residential density, and the farthest residential location for a given household among the three optimal cities.

Finally, we numerically explore the relationships of optimal capacity investments among no-toll, first-best, and second-best cities (see also Appendix G). It shows that such relationships are ambiguous, depending on the model parameters, such as the marginal cost δ of capacity expansion. As δ is small, the no-toll optimal city requires the largest capacity investment; otherwise, the largest investment takes place in the first-best city.

5. Conclusion and further studies

We presented a novel model combining household residential location choice and bottleneck congestion. Residents are heterogeneous in terms of income, and continuously distributed along a linear city corridor. A bottleneck with limited capacity in the corridor causes traffic congestion during commutes. We analytically derive urban equilibrium solution by well-defined differential equations. We examine the distributional effects of bottleneck capacity expansion on heterogeneous households, and investigate the bottleneck capacity investment issues under the no toll, first-best toll, and second-best toll, as well as the self-financing issue. The biases in bottleneck capacity investment due to ignoring its effects on urban spatial structure are also examined.

The following insightful findings are obtained. First, regardless of whether the toll is levied or not, all households benefit from bottleneck capacity expansion. However, the marginal benefit is differential across different-income households. In particular, under no toll, the mid-income households residing at the bottleneck location gain the most, while the richest or the poorest households gain the least. Under the first-best toll, the households residing in a suburban location (close to the bottleneck location) gain the most. Under the second-best toll, there is a critical condition leading households' marginal benefit from capacity expansion to decrease with their income. Numerical results show that such differential marginal benefit depends on the bottleneck location and current bottleneck capacity level. Second, under the first-best or second-best toll, the ratio of toll revenue to optimal capacity investment cost is exactly equal to the elasticity of investment cost with respect to capacity. With constant return to scale for capacity investment, the self-financing principle holds exactly. Third, ignoring the change in urban spatial structure causes distorted capacity investment. Specifically, under no toll, overinvestment or underinvestment in capacity may occur, depending on capacity expansion cost. However, under the first-best and second-best tolls, only underinvestment takes place. The proposed model elucidates the interplay among bottleneck congestion, capacity expansion, and urban spatial structure, and can serve as a useful tool to efficiently evaluate and design anti-congestion policies from demand side, supply side, or both.

Some extensions can be envisaged as follows. First, household utility follows a quasi-linear function, which should be justified using real survey data. Empirical calibrations of the utility function are beneficial for the model application in realistic cases. Second, this paper assumes a uniform household income distribution. Considering other income distributions, such as lognormal distribution, is more realistic. However, deriving a closed-form solution for a general distribution could be challenging, and a simulation method may be necessary in such cases. Third, this paper assumes a linear monocentric urban structure. However, many cities in reality have radial and/or circular structures (Li et al., 2013; Li and Wang, 2018) or multiple business centers (Anas and Kim, 1996; Anas and Xu, 1999). Therefore, there is a need to extend the proposed model to account for other urban forms. Finally, this paper focuses on the case of fixed bottleneck location, like bridges and tunnels. However, real road bottlenecks can arise from various random factors like traffic accidents, adverse weather, road works, and lane changes. The bottleneck locations may thus stochastically vary by time of day, day of week, and season. Therefore, extending the proposed model to consider the case of stochastic bottleneck locations is meaningful.

Acknowledgments

We are grateful to the editor Kristian Behrens, Erik Verhoef, Hai Yang, Hai-Jun Huang, Robin Lindsey and two anonymous referees for their helpful comments and suggestions. This study was supported for the first two authors by grants from the National Natural Science Foundation of China (72131008) and the Fundamental Research Funds for the Central Universities (2021GCRC014) and for the last author by AFFINITE, ANR grant, France.

CRediT authorship contribution statement

Zhi-Chun Li: Conceptualization, Methodology, Investigation, Resources, Supervision, Writing – original draft, Writing – review & editing. **De-Ping Yu:** Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **André de Palma:** Conceptualization, Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing.

References

Alonso, W., 1964. Location and Land Use: Toward a General Theory of Land Rent. Harvard University Press, Cambridge, Massachusetts.

Akamatsu, T., Fujishima, S., Takayama, Y., 2017. Discrete-space agglomeration model with social interactions: Multiplicity, stability, and continuous limit of equilibria. Journal of Mathematical Economics 69, 22-37.

Anas, A., Kim, I., 1996. General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration. Journal of Urban Economics 40 (2), 232-256.

Anas, A., Xu, R., 1999. Congestion, land use, and job dispersion: A general equilibrium model. Journal of Urban Economics 45 (3), 451-473.

Arnott, R., 1998. Congestion tolling and urban spatial structure. Journal of Regional Science 38 (3), 495-504.

Arnott, R., 2004. Does the Henry George Theorem provide a practical guide to optimal city size? The American Journal of Economics and Sociology 63 (5), 1057-1090.

Arnott, R., de Palma, A., Lindsey, R., 1990. Economics of a bottleneck. Journal of Urban Economics 27 (1), 111-130.

Arnott, R., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. American Economic Review 83 (1), 161.

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

Arnott, R., Kraus, M., 1995. Financing capacity in the bottleneck model. Journal of Urban Economics 38 (3), 272-290.

Baum-Snow, N., 2007. Did highways cause suburbanization? Quarterly Journal of Economics 122 (2), 775-805.

Beckmann, M.J., 1976. Spatial equilibrium and the dispersed city. In: Papageorgiou, Y.Y. (ed.), Mathematical Land Use Theory, Lexington Books, Lexington, MA, pp. 117-125.

Berglas, E., Pines, D., 1981. Clubs, local public goods and transportation models: A synthesis. Journal of Public Economics 15 (2), 141-162.

Blanchet, A., Mossay, P., Santambrogio, F., 2016. Existence and uniqueness of equilibrium for a spatial model of social interactions. International Economic Review 57 (1), 31-60.

De Lara, M., de Palma, A., Kilani, M., Piperno, S., 2013. Congestion pricing and long term urban form: Application to Paris region. Regional Science and Urban Economics 43 (2), 282-295.

de Palma, A., Lindsey, R., 2007. Transport user charges and cost recovery. In: de Palma, A., Lindsey, R., Proost, S. (Eds.), Investment and the Use of Tax and Toll Revenues in the Transport Sector, Research in Transportation Economics 19, 29-57.

de Palma, A., Lindsey, R., 2011. Traffic congestion pricing methodologies and technologies. Transportation Research Part C 19 (6), 1377-1399.

de Palma, E., Arnott, R., 2012. Morning commute in a single-entry traffic corridor with no late arrivals. Transportation Research Part B 46 (1), 1-29.

Fosgerau, M., de Palma, A., 2012. Congestion in a city with a central bottleneck. Journal of Urban Economics 71 (3), 269-277.

Fosgerau, M., Kim, J., 2019. Commuting and land use in a city with bottlenecks: Theory and

evidence. Regional Science and Urban Economics 77, 182-204.
Fosgerau, M., Kim, J., Ranjan, A., 2018. Vickrey meets Alonso: Commute scheduling and congestion in a monocentric city. Journal of Urban Economics 105, 40-53.
Fujishige, S., 2005. Submodular Functions and Optimization. Elsevier, Amsterdam.
Fujita, M., 1989. Urban Economic Theory: Land Use and City Size. Cambridge University Press.
Fujita, M., Thisse, J.F., 2002. Economics of Agglomeration: Cities, Industrial Location, and Regional Growth. Cambridge University Press.
Gubins, S., Verhoef, E.T., 2014. Dynamic bottleneck congestion and residential land use in the monocentric city. Journal of Urban Economics 80, 51-61.
Hall, J.D., 2024. Inframarginal travelers and transportation policy. International Economic Review. https://doi.org/10.1111/iere.12692.
Kanemoto, Y., 1977. Cost-benefit analysis and the second best land use for transportation. Journal of Urban Economics 4 (4), 483-503.
Kanemoto, Y., 1984. Pricing and investment policies in a system of competitive commuter railways. Review of Economic Studies 51 (4), 665-681.
Li, Z.C., Chen, Y.J., Wang, Y.D., Lam, W.H.K., Wong, S.C., 2013. Optimal density of radial major roads in a two-dimensional monocentric city with endogenous residential distribution and housing prices. Regional Science and Urban Economics 43 (6), 927-937.
Li, Z.C., Huang, H.J., Yang, H., 2020. Fifty years of the bottleneck model: A bibliometric review and future research directions. Transportation Research Part B 139, 311-342.
Li, Z.C., Wang, Y.D., 2018. Analysis of multimodal two-dimensional urban system equilibrium for cordon toll pricing and bus service design. Transportation Research Part B 111, 244-265.
Li, Z.C., Wang, Y.D., Lam, W.H.K., Sumalee, A., Choi, K., 2014. Design of sustainable cordon toll pricing schemes in a monocentric city. Networks and Spatial Economics 14 (2), 133-158.
Lindsey, R., 2012. Road pricing and investment. Economics of Transportation 1 (1-2), 49-63.
Mills, E. S., 1972. Urban Economics, Scott Foresman, Glenview, Illinois.
Mohring, H., Harwitz, M., 1962. Highway Benefits: An Analytical Framework. Northwestern University Press.
Mossay, P., Picard, P.M., 2011. On spatial equilibria in a social interaction model. Journal of Economic Theory 146 (6), 2455-2477.
Mun, S.I., 1999. Peak-load pricing of a bottleneck with traffic jam. Journal of Urban

Economics 46 (3), 323-349.
Mun, S.I., Konishi, K.J., Yoshikawa, K., 2003. Optimal cordon pricing. Journal of Urban Economics 54 (1), 21-38.
Muth, R. F., 1969. Cities and Housing. University of Chicago Press, Chicago.
Newell, G.F., 1988. Traffic flow for the morning commute. Transportation Science 22 (1), 47-58.
Picard, P.M., Tabuchi, T., 2013. On microfoundations of the city. Journal of Economic Theory 148 (6), 2561-2582.
Picard, P.M., Tran, T.T.H., 2021. Small urban green areas. Journal of Environmental Economics and Management 106, 102418.
Pines, D., Sadka, E., 1981. Optimum, second-best, and market allocations of resources within an urban area. Journal of Urban Economics 9 (2), 173-189.
Solow, R.M., 1973. Congestion cost and the use of land for streets. The Bell Journal of Economics and Management Science 4 (2), 602.
Tabuchi, T., 1993. Bottleneck congestion and modal split. Journal of Urban Economics 34 (3), 414-431.
Takayama, Y., 2020. Who gains and who loses from congestion pricing in a monocentric city with a bottleneck? Economics of Transportation 24, 100189.
Takayama, Y., Kuwahara, M., 2017. Bottleneck congestion and residential location of heterogeneous commuters. Journal of Urban Economics 100, 65-79.
Topkis, D.M., 1998. Supermodularity and Complementarity. Princeton University Press, Princeton.
Van den Berg, V.A.C., Verhoef, E.T., 2011. Winning or losing from dynamic bottleneck congestion pricing? The distributional effects of road pricing with heterogeneity in values of time and schedule delay. Journal of Public Economics 95 (7-8), 983-992.
Vickrey, W.S., 1969. Congestion theory and transport investment. American Economic Review 59 (2), 251-260.
Vickrey, W.S., 1973. Pricing, metering, and efficiently using urban transportation facilities. Highway Research Record 476, 36-48.
Wu, W.X., Huang, H.J., 2015. An ordinary differential equation formulation of the bottleneck model with user heterogeneity. Transportation Research Part B 81, 34-58.
Xiao, F., Zhang, H.M., 2014. Pareto-improving and self-sustainable pricing for the morning commute with nonidentical commuters. Transportation Science 48 (2), 159-169.

Appendix A: Derivation of Eq. (2)

For a resident at suburban location x $(x \geq a)$ with VOT α, the commuting cost $c_S(x,\alpha)$ is composed of queuing delay cost at the bottleneck, the early-arrival penalty, and the free-flow travel time cost, expressed as

$$c_S(x,\alpha) = \alpha m(t) + \beta(\alpha)(t_e - t) + t_0 x\alpha, \tag{A.1}$$

where $\beta(\alpha)$ is the value of early-arrival time for the resident with VOT α, t is the arrival time at the CBD, t_e is the desired arrival time or work start time, and $m(t)$ is the queuing delay time at the bottleneck. It is assumed that residents with different VOTs have the same ratio, η, of value of early-arrival time to VOT, i.e.,

$$\eta = \frac{\beta(\alpha)}{\alpha}, \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]. \tag{A.2}$$

At equilibrium, any resident in the suburb cannot unilaterally change his/her schedule to reduce commuting cost, i.e., $dc_S(x,\alpha)/dt = 0$. Combining it with Eq. (A.1) yields

$$\alpha \frac{dm(t)}{dt} - \beta(\alpha) = 0. \tag{A.3}$$

Let t_s be the arrival time of the first commuter at the CBD. Obviously, there is no queue for the first commuter, i.e., $m(t_s) = 0$. Combining it with Eqs. (A.2) and (A.3) yields

$$m(t) = \eta(t - t_s). \tag{A.4}$$

Since late arrival is not allowed, the work start time t_e is also the arrival time of the last commuter at the CBD, and thus the peak period lasts $t_e - t_s$ units of time. The bottleneck runs at capacity during the peak period. We thus have

$$\frac{\hat{N}_S}{q} = t_e - t_s. \tag{A.5}$$

Substituting Eqs. (A.2), (A.4), and (A.5) into (A.1), the commuting cost for the suburban residents residing at x with VOT α can be obtained as

$$c_S(x,\alpha) = \frac{\hat{N}_S}{q}\eta\alpha + t_0 x\alpha. \tag{A.6}$$

This completes the derivation of Eq. (2).

Appendix B. Derivation of Eq. (36)

Bottleneck capacity expansion changes household residential distribution. From Eq. (22), we can obtain

$$\frac{d\alpha^*}{dq}=\left(kr\frac{e^{2ra}+1}{e^{2ra}-1}b+kr\frac{b^2\alpha^*}{\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}}+\frac{\eta b(2\alpha^*-\underline{\alpha})}{q}\right)^{-1}\frac{\eta b(\alpha^*-\underline{\alpha})\alpha^*}{q^2}>0\,. \quad \text{(B.1)}$$

$$\frac{d\left(b(\alpha^*-\underline{\alpha})/q\right)}{dq}=-\frac{1}{\eta\alpha^*}\left(kr\frac{e^{2ra}+1}{e^{2ra}-1}b+kr\frac{b^2\alpha^*}{\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}}+\frac{\eta}{q}b(\alpha^*-\underline{\alpha})\right)\frac{d\alpha^*}{dq}<0. \quad \text{(B.2)}$$

From Eq. (19), we can derive

$$\frac{dB}{dq}=\frac{b}{r\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}}\frac{d\alpha^*}{dq}>0\,, \quad \text{(B.3)}$$

which means that the capacity expansion leads the city to expand outward.

From Eq. (11), the aggregate net land rent in Eq. (33) can be rewritten as

$$LR=\int_0^a\frac{k}{2}n_D^2(x)dx+\int_a^B\frac{k}{2}n_S^2(x)dx-r_AB\,. \quad \text{(B.4)}$$

From Eqs. (14) and (15), we have

$$\begin{aligned}&\int_0^a\frac{k}{2}n_D^2(x)dx=\int_0^a\frac{k}{2}r^2\left(c_1^2e^{2rx}+c_2^2e^{-2rx}-2c_1c_2\right)dx\\&=\frac{kr}{4}\frac{\left(e^{2ra}+1\right)\left((b\overline{\alpha})^2+(b\alpha^*)^2\right)-4e^{ra}b^2\overline{\alpha}\alpha^*}{e^{2ra}-1}+kr^2a\frac{e^{2ra}(b\overline{\alpha})^2-\left(e^{3ra}+e^{ra}\right)b^2\overline{\alpha}\alpha^*+e^{2ra}(b\alpha^*)^2}{\left(e^{2ra}-1\right)^2}.\end{aligned} \quad \text{(B.5)}$$

From Eqs. (18)-(20), we have

$$\begin{aligned}&\int_a^B\frac{k}{2}n_S^2(x)dx=\int_a^B\frac{k}{2}r^2\left(c_3^2e^{2rx}+c_4^2e^{-2rx}-2c_3c_4\right)dx\\&=\frac{kr}{4}\left(-b\underline{\alpha}\sqrt{2r_A/(kr^2)}\right)+\frac{kr}{4}b\alpha^*\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}-kr^2\frac{(b\underline{\alpha})^2-2r_A/(kr^2)}{4}\left(B-a\right).\end{aligned} \quad \text{(B.6)}$$

Substituting Eqs. (B.5) and (B.6) into (B.4), we can derive

$$\begin{aligned}\frac{dLR}{dq}&=\left(\frac{krb}{2}\frac{\left(e^{2ra}+1\right)b\alpha^*-2e^{ra}b\overline{\alpha}}{e^{2ra}-1}+\frac{krb\left((b\alpha^*)^2-(b\underline{\alpha})^2\right)}{2\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}}+kr^2a\frac{-e^{ra}\left(e^{2ra}+1\right)b^2\overline{\alpha}+2e^{2ra}b^2\alpha^*}{\left(e^{2ra}-1\right)^2}\right)\frac{d\alpha^*}{dq}\\&<\left(\frac{krb}{2}\frac{\left(e^{2ra}+1\right)b\alpha^*-2e^{ra}b\overline{\alpha}}{e^{2ra}-1}+\frac{krb}{2}\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}+kr^2a\frac{-e^{ra}\left(e^{2ra}+1\right)b^2\overline{\alpha}+2e^{2ra}b^2\alpha^*}{\left(e^{2ra}-1\right)^2}\right)\frac{d\alpha^*}{dq}.\end{aligned} \quad \text{(B.7)}$$

According to equilibrium condition Eq. (22), one can judge

$$\frac{krb}{2}\frac{\left(e^{2ra}+1\right)b\alpha^*-2e^{ra}b\overline{\alpha}}{e^{2ra}-1}+\frac{krb}{2}\sqrt{(b\alpha^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}<0\,. \quad \text{(B.8)}$$

Combining Eqs. (B.7) and (B.8), one obtains $dLR/dq < 0$, meaning that bottleneck capacity expansion leads to a decrease in the aggregate net land rent.

Note that VOT α and location x have a one-to-one correspondence, i.e., $N_i(x) = b(\bar{\alpha} - \alpha),\ i = D, S$. According to Eqs. (13) and (17), we can obtain the VOT of a household at any location x as

$$\alpha(x) = \begin{cases} -\dfrac{c_1 e^{rx} + c_2 e^{-rx}}{b},\ x \in [0, a), \\ -\dfrac{c_3 e^{rx} + c_4 e^{-rx}}{b},\ x \in [a, B]. \end{cases} \tag{B.9}$$

Hence, the utility of the households at location x can be written as

$$u(x) = w(\alpha(x)) - c(x, \alpha(x)) - kn(x)\,, \tag{B.10}$$

where $n(x) = n_D(x)$ for $x \in [0, a)$, and $n(x) = n_S(x)$ for $x \in [a, B]$.

From (B.10), the total utility of all households in the city can be expressed as

$$TU = \int_0^B u(x)n(x)dx = \int_0^B w(\alpha(x))n(x)dx - \int_0^B c(x, \alpha(x))n(x)dx - \int_0^B kn^2(x)dx. \tag{B.11}$$

Combining it with Eq. (B.4), we have

$$\Phi = TU + LR = \int_0^B w(\alpha(x))n(x)dx - \int_0^B c(x, \alpha(x))n(x)dx - LR - 2r_A B\,. \tag{B.12}$$

From Eqs. (1), (2), (15), and (20), we have

$$\begin{aligned} &\int_0^B c(x, \alpha(x))n(x)dx \\ &= \int_0^a \frac{t_0 r}{b} x\left(c_1 e^{rx} + c_2 e^{-rx}\right)\left(c_2 e^{-rx} - c_1 e^{rx}\right)dx + \int_a^B \frac{r}{b}\left(t_0 x + \frac{b(\alpha^* - \underline{\alpha})\eta}{q}\right)\left(c_3 e^{rx} + c_4 e^{-rx}\right)\left(c_4 e^{-rx} - c_3 e^{rx}\right)dx \\ &= -\left(\begin{aligned} &\frac{t_0 a}{2b}\frac{2e^{2ra}(b\bar{\alpha})^2 + (e^{4ra}+1)(b\alpha^*)^2 - 2e^{ra}(e^{2ra}+1)b\bar{\alpha}b\alpha^*}{\left(e^{2ra}-1\right)^2} + \frac{t_0}{4br}\frac{4e^{ra}b\bar{\alpha}b\alpha^* - (e^{2ra}+1)\left((b\bar{\alpha})^2 + (b\alpha^*)^2\right)}{e^{2ra}-1} \\ &+ \frac{t_0}{4b}\left((b\underline{\alpha})^2 + 2r_A/(kr^2)\right)B - \frac{t_0 a}{4b}\left(2(b\alpha^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)\right) + \frac{(\alpha^* - \underline{\alpha})\eta}{2q}\left((b\underline{\alpha})^2 - (b\alpha^*)^2\right) \\ &+ \frac{t_0}{4br}\left(b\underline{\alpha}\sqrt{2r_A/(kr^2)} - b\alpha^*\sqrt{(b\alpha^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}\right) \end{aligned}\right). \end{aligned} \tag{B.13}$$

From Eqs. (B.2), (B.3), and (B.13), we can derive

$$\frac{d\left(\int_0^B c(x, \alpha(x))n(x)dx\right)}{dq} = -\left(\begin{aligned} &\frac{krb}{2}\frac{e^{2ra}+1}{e^{2ra}-1}\left(b\alpha^* - \frac{b\underline{\alpha}^2}{\alpha^*}\right) + \frac{\eta\left((b\alpha^*)^2 - (b\underline{\alpha})^2\right)\left(\alpha^* - \underline{\alpha}\right)}{2\alpha^* q} - \frac{b\left(\alpha^* - \underline{\alpha}\right)\eta}{q}b\alpha^* \\ &+ t_0 a\frac{2e^{2ra}b\alpha^* - e^{ra}(e^{2ra}+1)b\bar{\alpha}}{\left(e^{2ra}-1\right)^2} + \frac{krb}{2}\frac{-\left(e^{2ra}+1\right)b\alpha^* + 2e^{ra}b\bar{\alpha}}{e^{2ra}-1} \end{aligned}\right)\frac{d\alpha^*}{dq}. \tag{B.14}$$

In addition, from Eq. (B.3), we have

$$2r_A \frac{dB}{dq} = \frac{krb}{2} \frac{4 r_A/(kr^2)}{\sqrt{(b\alpha^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}} \frac{d\alpha^*}{dq}. \tag{B.15}$$

Note that the first integral term on the RHS of Eq. (B.12) represents the total income of all households, and is independent of capacity q. We thus have

$$\frac{d\Phi}{dq} = -\frac{d\left(\int_0^B c(x,\alpha(x))n(x)dx\right)}{dq} - \frac{dLR}{dq} - 2r_A \frac{dB}{dq}. \tag{B.16}$$

Substituting Eqs. (B.7), (B.14), and (B.15) into (B.16), and using the equilibrium condition Eq. (22) yield

$$\frac{d\Phi}{dq} = \left(\frac{krb}{2} \frac{e^{2ra}+1}{e^{2ra}-1} \left(b\alpha^* - \frac{b\underline{\alpha}^2}{\alpha^*} \right) + \frac{krb\left((b\alpha^*)^2 - (b\underline{\alpha})^2\right)}{2\sqrt{(b\alpha^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}} + \frac{\eta\left((b\alpha^*)^2 - (b\underline{\alpha})^2\right)\left(\alpha^* - \underline{\alpha}\right)}{2\alpha^* q} \right) \frac{d\alpha^*}{dq} > 0, \tag{B.17}$$

where $\frac{d\alpha^*}{dq}$ is given by Eq. (B.1).

Note that the following relationships hold

$$\frac{\eta b}{4q}(2\alpha^* - \underline{\alpha})(b\alpha^* - \frac{b\underline{\alpha}^2}{\alpha^*}) < \frac{\eta\left((b\alpha^*)^2 - (b\underline{\alpha})^2\right)\left(\alpha^* - \underline{\alpha}\right)}{2\alpha^* q} < \frac{\eta b}{2q}(2\alpha^* - \underline{\alpha})(b\alpha^* - \frac{b\underline{\alpha}^2}{\alpha^*}). \tag{B.18}$$

Combining Eqs. (B.1), (B.17), and (B.18), we can obtain

$$\frac{\eta b^2}{4q^2}\left((\alpha^*)^2 - \underline{\alpha}^2\right)\left(\alpha^* - \underline{\alpha}\right) < \frac{d\Phi}{dq} < \frac{\eta b^2}{2q^2}\left((\alpha^*)^2 - \underline{\alpha}^2\right)\left(\alpha^* - \underline{\alpha}\right). \tag{B.19}$$

This completes the derivation Eq. (36).

Appendix C: Proof of Proposition 5

As proven in Takayama and Kuwahara (2017), given the critical household VOT at the bottleneck, the land market is efficient in both the downtown and the suburb. Therefore, the distortion in our model arises only from two sources: the bottleneck congestion externality incurred by suburban commuters, and the mismatch in population allocation between the downtown and suburb.

First, we justify that the first-best dynamic congestion toll on the suburban households with VOT α should take the following form:

$$\tau_{FB}(\alpha) = \tau^*(\alpha) + \lambda, \ \alpha \in [\underline{\alpha}, \alpha_{FB}^*] \tag{C.1}$$

where $\tau^*(\alpha)$ is the dynamic bottleneck congestion toll that exactly eliminates queues, and parameter λ is a constant toll used to balance the population allocation between the downtown and suburb. To confirm this, we assume that there is another toll scheme $\hat{\tau}_{FB}(\alpha)$, with the critical VOT at the bottleneck $\hat{\alpha}_{FB}^*$. In toll scheme $\tau_{FB}(\alpha)$, we can always achieve the same critical VOT $\hat{\alpha}_{FB}^*$ by adjusting the value of λ, such that the population allocation between downtown and suburb under toll scheme $\tau_{FB}(\alpha)$ is the same as that under toll scheme $\hat{\tau}_{FB}(\alpha)$. Since household preference is quasi-linear, toll schemes $\tau_{FB}(\alpha)$ and $\hat{\tau}_{FB}(\alpha)$ have the same residential density and land rent across the city. Therefore, the difference in social surplus between these two toll schemes lies in the total bottleneck congestion cost (excluding toll). Apparently, the social surplus under toll scheme $\tau_{FB}(\alpha)$ is larger than that under toll scheme $\hat{\tau}_{FB}(\alpha)$ because the former minimizes the total congestion cost according to the bottleneck theory.

Since suburban commuters' VOTs follow a uniform distribution $U(\underline{\alpha}, \alpha_{FB}^*)$, the optimal bottleneck dynamic toll $\tau^*(\alpha)$ that eliminates queues in Eq. (C.1) can be derived as

$$\tau^*(\alpha) = \frac{\eta b}{2q}(\alpha^2 - \underline{\alpha}^2), \ \alpha \in [\underline{\alpha}, \alpha_{FB}^*]. \tag{C.2}$$

The derivation of Eq. (C.2) is similar to the procedure in Appendix A, and thus omitted here. Readers can refer to Xiao and Zhang (2014) and Wu and Huang (2015). Meanwile, the corresponding bottleneck congestion cost $c_B(\alpha)$ (i.e., early-arrival schedule delay cost) after the toll can be derived as

$$c_B(\alpha)=\frac{\eta b}{q}(\alpha_{FB}^{*}-\alpha)\alpha,\ \alpha\in[\underline{\alpha},\alpha_{FB}^{*}]. \tag{C.3}$$

We next solve the optimal constant toll λ that maximizes the social benefit. From Eq. (42), the equilibrium condition determining the critical VOT α_{FB}^{*} can be rewritten as

$$kr\frac{e^{2ra}+1}{e^{2ra}-1}b\alpha_{FB}^{*}+kr\sqrt{(b\alpha_{FB}^{*})^2-(b\underline{\alpha})^2+2r_A/(kr^2)}+\frac{\eta b}{2q}\left((\alpha_{FB}^{*})^2-\underline{\alpha}^2\right)+\lambda=kr\frac{2e^{ra}b\overline{\alpha}}{e^{2ra}-1}. \tag{C.4}$$

Referring to Eq. (B.12), the optimization problem Eq. (43) can be reformulated as

$$\max_{\lambda}\Phi_{FB}=TU_{FB}+LR_{FB}+TR_{FB}=\int_0^B w(\alpha(x))n(x)dx-\int_0^B\tilde{c}(x,\alpha(x))n(x)dx-LR_{FB}-2r_AB, \tag{C.5}$$

where $\alpha(x)$ is given by Eq. (B.9), and $\tilde{c}(x,\alpha(x))$ is defined as

$$\tilde{c}(x,\alpha)=\begin{cases}t_0x\alpha, & \text{for } 0\le x<a,\\ \dfrac{\eta b}{q}(\alpha_{FB}^{*}-\alpha)\alpha+t_0x\alpha, & \text{for } x\ge a.\end{cases} \tag{C.6}$$

From Eqs. (1), (2), (15), (20), and (B.9), we have

$$\begin{aligned}&\int_0^B\tilde{c}(x,\alpha(x))n(x)dx\\&=\int_0^a\frac{t_0r}{b}x\left(c_1e^{rx}+c_2e^{-rx}\right)\left(c_2e^{-rx}-c_1e^{rx}\right)dx+\int_a^B\left(t_0x+\frac{\eta b}{q}(\alpha_{FB}^{*}+\frac{c_3e^{rx}+c_4e^{-rx}}{b})\right)\frac{r}{b}\left(c_3e^{rx}+c_4e^{-rx}\right)\left(c_4e^{-rx}-c_3e^{rx}\right)dx\\&=-\frac{t_0a}{2b}\frac{1}{(e^{2ar}-1)^2}\left(2e^{2ra}(b\overline{\alpha})^2+(e^{4ra}+1)(b\alpha_{FB}^{*})^2-2e^{ra}(e^{2ra}+1)b^2\overline{\alpha}\alpha_{FB}^{*}\right)+\frac{t_0}{4br}\frac{1}{e^{2ar}-1}\left((e^{2ra}+1)\left((b\overline{\alpha})^2+(b\alpha_{FB}^{*})^2\right)-4e^{ra}b^2\overline{\alpha}\alpha_{FB}^{*}\right)\\&-\frac{t_0}{2b}\left(\begin{aligned}&\frac{B}{2}\left((b\underline{\alpha})^2+2r_A/(kr^2)\right)-\frac{a}{2}\left(2(b\alpha_{FB}^{*})^2-(b\underline{\alpha})^2+2r_A/(kr^2)\right)\\&+\frac{1}{2r}\left(b\underline{\alpha}\sqrt{2r_A/(kr^2)}-b\alpha_{FB}^{*}\sqrt{(b\alpha_{FB}^{*})^2-(b\underline{\alpha})^2+2r_A/(kr^2)}\right)\end{aligned}\right)+\frac{\eta b}{6q}\left(b\alpha_{FB}^{*}\left((\alpha_{FB}^{*})^2-\underline{\alpha}^2\right)-2b\underline{\alpha}^2\left(\alpha_{FB}^{*}-\underline{\alpha}\right)\right).\end{aligned} \tag{C.7}$$

Based on Eq. (C.7), we can derive

$$\begin{aligned}&\frac{d\int_0^B\tilde{c}(x,\alpha(x))n(x)dx}{d\lambda}\\&=\left(\begin{aligned}&t_0a\frac{e^{ra}(e^{2ra}+1)b\overline{\alpha}-(e^{4ra}+1)b\alpha_{FB}^{*}}{\left(e^{2ra}-1\right)^2}+\frac{krb}{2}\frac{\left(e^{2ra}+1\right)b\alpha_{FB}^{*}-2e^{ra}b\overline{\alpha}}{e^{2ra}-1}+\frac{\eta}{2q}\left((b\alpha_{FB}^{*})^2-(b\underline{\alpha})^2\right)\\&-\frac{krb}{4}\frac{-(b\alpha_{FB}^{*})^2+(b\underline{\alpha})^2+2r_A/(kr^2)}{\sqrt{(b\alpha_{FB}^{*})^2-(b\underline{\alpha})^2+2r_A/(kr^2)}}+\frac{krb}{4}\sqrt{(b\alpha_{FB}^{*})^2-(b\underline{\alpha})^2+2r_A/(kr^2)}+t_0ab\alpha_{FB}^{*}\end{aligned}\right)\frac{d\alpha_{FB}^{*}}{d\lambda},\end{aligned} \tag{C.8}$$

where $d\alpha_{FB}^{*}/d\lambda$ is less than 0 and can be determined by Eq. (C.4).

Besides, from Eqs. (B.7) and (B.15), we have

$$\frac{dLR}{d\lambda}=\left(\frac{krb}{2}\frac{\left(e^{2ra}+1\right)b\alpha_{FB}^{*}-2e^{ra}b\overline{\alpha}}{e^{2ra}-1}+\frac{krb\left((b\alpha_{FB}^{*})^{2}-(b\underline{\alpha})^{2}\right)}{2\sqrt{(b\alpha_{FB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}}+kr^{2}a\frac{-e^{ra}\left(e^{2ra}+1\right)b^{2}\overline{\alpha}+2e^{2ra}b^{2}\alpha_{FB}^{*}}{\left(e^{2ra}-1\right)^{2}}\right)\frac{d\alpha_{FB}^{*}}{d\lambda}, \tag{C.9}$$

$$2r_{A}\frac{dB}{d\lambda}=\frac{krb}{2}\frac{4r_{A}/(kr^{2})}{\sqrt{(b\alpha_{FB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}}\frac{d\alpha_{FB}^{*}}{d\lambda}. \tag{C.10}$$

Taking the first-order derivative of Φ_{FB} with respect to λ in Eq. (C.5), and combining Eqs. (C.4), (C.8), (C.9), and (C.10) yield

$$\frac{d\Phi_{FB}}{d\lambda}=-\frac{d\int_{0}^{B}\mathcal{C}(x,\alpha(x))n(x)dx}{d\lambda}-\frac{dLR}{d\lambda}-2r_{A}\frac{dB}{d\lambda}=\lambda\frac{d\alpha_{FB}^{*}}{d\lambda}. \tag{C.11}$$

Since $d\alpha_{FB}^{*}/d\lambda<0$, Eq. (C.11) implies the social benefit is maximized when $\lambda=0$. Combing it with Eq. (C.2), one obtains that the social optimal toll is exactly the optimal bottleneck dynamic toll that eliminates bottleneck queues, i.e.,

$$\tau_{FB}(\alpha)=\frac{\eta b}{2q}(\alpha^{2}-\underline{\alpha}^{2}),\ \alpha\in[\underline{\alpha},\alpha_{FB}^{*}]. \tag{C.12}$$

This completes the proof of Proposition 5.

Appendix D: Derivation of Eq. (51)

Referring to Eq. (C.5), we have

$$\frac{d\Phi_{FB}}{dq} = -\frac{d\int_0^B \mathscr{C}(x,\alpha(x))n(x)dx}{dq} - \frac{dLR_{FB}}{dq} - 2r_A\frac{dB}{dq}. \tag{D.1}$$

From Eqs. (C.4) and (C.7), we can derive

$$\frac{d\int_0^B \mathscr{C}(x,\alpha(x))n(x)dx}{dq} = \left(\begin{aligned} & t_0 a\frac{e^{ra}(e^{2ra}+1)b\overline{\alpha} - 2e^{2ra}b\alpha_{FB}^*}{\left(e^{2ra}-1\right)^2} + \frac{krb}{2}\frac{\left(e^{2ra}+1\right)b\alpha_{FB}^* - 2e^{ra}b\overline{\alpha}}{e^{2ra}-1} + \frac{krb}{2}\frac{-(b\alpha_{FB}^*)^2-(b\underline{\alpha})^2}{\sqrt{(b\alpha_{FB}^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}} \\ & +\frac{\eta}{2q}\left((b\alpha_{FB}^*)^2-(b\underline{\alpha})^2\right) - \frac{e^{2ra}+1}{e^{2ra}-1}krb^2\alpha_{FB}^* - \frac{1}{3}\frac{\eta b^2}{q}\left(\alpha_{FB}^*-\underline{\alpha}\right)\left(2\alpha_{FB}^*+\underline{\alpha}\right) \\ & +\frac{2}{3}\left(b\alpha_{FB}^* + \frac{(b\underline{\alpha})^2}{\alpha_{FB}^*+\underline{\alpha}}\right)\left(krb\frac{e^{2ra}+1}{e^{2ra}-1} + \frac{krb^2\alpha_{FB}^*}{\sqrt{(b\alpha_{FB}^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}} + \frac{\eta b(\alpha_{FB}^*-\underline{\alpha})}{2q}\right) \end{aligned} \right)\frac{d\alpha_{FB}^*}{dq}, \tag{D.2}$$

where $\frac{d\alpha_{FB}^*}{dq}$ can be obtained from Eq. (C.4) as

$$\frac{d\alpha_{FB}^*}{dq} = \left(krb\frac{e^{2ra}+1}{e^{2ra}-1} + \frac{krb^2\alpha_{FB}^*}{\sqrt{(b\alpha_{FB}^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}} + \frac{\eta b}{q}\alpha_{FB}^*\right)^{-1}\frac{\eta b}{2q^2}\left((\alpha_{FB}^*)^2-\underline{\alpha}^2\right) > 0. \tag{D.3}$$

Substituting Eqs. (D.2), (B.7), and (B.15) into (D.1) yields

$$\begin{aligned} \frac{d\Phi_{FB}}{dq} &= \left(\frac{1}{3}b\alpha_{FB}^* - \frac{2}{3}\frac{(b\underline{\alpha})^2}{\alpha_{FB}^*+\underline{\alpha}}\right)\left(krb\frac{e^{2ra}+1}{e^{2ra}-1} + \frac{krb^2\alpha_{FB}^*}{\sqrt{(b\alpha_{FB}^*)^2-(b\underline{\alpha})^2+2r_A/(kr^2)}} + \frac{\eta b\alpha_{FB}^*}{q}\right)\frac{d\alpha_{FB}^*}{dq} \\ &= \frac{\eta b^2}{q^2}\left(\frac{1}{6}(\alpha_{FB}^*)^3 - \frac{1}{2}\alpha_{FB}^*\underline{\alpha}^2 + \frac{1}{3}\underline{\alpha}^3\right) > 0. \end{aligned} \tag{D.4}$$

This completes the derivation of Eq. (51).

Appendix E: Derivation of Eq. (58)

Similar to Eq. (22), under the second-best toll τ_{SB}, we can write the equilibrium condition that determines critical VOT α_{SB}^{*} as

$$kr\frac{e^{2ra}+1}{e^{2ra}-1}b\alpha_{SB}^{*}+kr\sqrt{(b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}+\frac{\eta}{q}b(\alpha_{SB}^{*}-\underline{\alpha})\alpha_{SB}^{*}+\tau_{SB}=kr\frac{2e^{ra}b\overline{\alpha}}{e^{2ra}-1}. \quad \text{(E.1)}$$

From Eq. (E.1), one can obtain

$$\frac{d\alpha_{SB}^{*}}{d\tau_{SB}}=-\left(krb\frac{e^{2ra}+1}{e^{2ra}-1}+krb\frac{b\alpha_{SB}^{*}}{\sqrt{(b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}}+\frac{\eta}{q}b(2\alpha_{SB}^{*}-\underline{\alpha})\right)^{-1}<0. \quad \text{(E.2)}$$

According to Eq. (B.12), the social benefit Φ_{SB} under the second-best toll can be written as

$$\Phi_{SB}=TU_{SB}+LR_{SB}+TR_{SB}=\int_{0}^{B}w(\alpha(x))n(x)dx-\int_{0}^{B}c(x,\alpha(x))n(x)dx-LR_{SB}-2r_{A}B, \quad \text{(E.3)}$$

where $c(x,\alpha(x))$ is obtained by replacing α^{*} in Eq. (3) with α_{SB}^{*}. From Eqs. (B.3) and (B.7), we have

$$\frac{dB}{d\tau_{SB}}=\frac{b}{r}\frac{1}{\sqrt{(b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}}\frac{d\alpha_{SB}^{*}}{d\tau_{SB}}. \quad \text{(E.4)}$$

$$\frac{dLR_{SB}}{d\tau_{SB}}=\left(\frac{krb}{2}\frac{\left(e^{2ra}+1\right)b\alpha_{SB}^{*}-2e^{ra}b\overline{\alpha}}{e^{2ra}-1}+\frac{krb\left((b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}\right)}{2\sqrt{(b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}}+kr^{2}a\frac{-e^{ra}\left(e^{2ra}+1\right)b^{2}\overline{\alpha}+2e^{2ra}b^{2}\alpha_{SB}^{*}}{\left(e^{2ra}-1\right)^{2}}\right)\frac{d\alpha_{SB}^{*}}{d\tau_{SB}}. \quad \text{(E.5)}$$

Referring to $\int_{0}^{B}c(x,\alpha(x))n(x)dx$ in Eq. (B.13), we can derive

$$\frac{d\left(\int_{0}^{B}c(x,\alpha(x))n(x)dx\right)}{d\tau_{SB}}=\begin{pmatrix}-\frac{t_{0}a}{2}\frac{2(e^{4ra}+1)b\alpha_{SB}^{*}-2e^{ra}(e^{2ra}+1)b\overline{\alpha}}{(e^{2ra}-1)^{2}}-\frac{t_{0}}{2r}\frac{2e^{ra}b\overline{\alpha}-(e^{2ra}+1)b\alpha_{SB}^{*}}{e^{2ra}-1}\\ -\frac{t_{0}}{4r}\frac{(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}{\sqrt{(b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}}+t_{0}ab\alpha_{SB}^{*}-\frac{\eta\left((b\underline{\alpha})^{2}-(b\alpha_{SB}^{*})^{2}\right)-2\eta\left(b\alpha_{SB}^{*}-b\underline{\alpha}\right)b\alpha_{SB}^{*}}{2q}\\ +\frac{t_{0}}{4r}\sqrt{(b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}+\frac{t_{0}}{4r}\frac{(b\alpha_{SB}^{*})^{2}}{\sqrt{(b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}}\end{pmatrix}\frac{d\alpha_{SB}^{*}}{d\tau_{SB}}. \quad \text{(E.6)}$$

The optimal second-best toll τ_{SB} requires $d\Phi_{SB}/d\tau_{SB}=0$. From Eqs. (E.3)-(E.6), we have

$$\frac{d\Phi_{SB}}{d\tau_{SB}}=\left(krb\frac{2e^{ra}b\overline{\alpha}-\left(e^{2ra}+1\right)b\alpha_{SB}^{*}}{e^{2ra}-1}-krb\sqrt{(b\alpha_{SB}^{*})^{2}-(b\underline{\alpha})^{2}+2r_{A}/(kr^{2})}+\frac{\eta b^{2}}{2q}\left(\alpha_{SB}^{*}-\underline{\alpha}\right)\left(-3\alpha_{SB}^{*}-\underline{\alpha}\right)\right)\frac{d\alpha_{SB}^{*}}{d\tau_{SB}}=0. \quad \text{(E.7)}$$

Combining Eqs. (E.1) and (E.7) yields

$$\tau_{SB}=\frac{\eta b}{2q}\left((\alpha_{SB}^{*})^{2}-\underline{\alpha}^{2}\right). \quad \text{(E.8)}$$

This completes the derivation of Eq. (58).

Appendix F: Proof of Proposition 11

Note that given the critical VOT at the bottleneck, the city length, residential density, and residential location with no toll are the same as those with congestion tolling, which can be obtained by Eqs. (15), (19), (20), and (23). This means that the critical VOT at the bottleneck governs the urban spatial structure. In the following, we look at the effects of the critical VOT on the urban spatial structure.

From Eq. (B.3), we have

$$\frac{dB}{d\alpha^*} = \frac{b}{r\sqrt{(b\alpha^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}} > 0, \tag{F.1}$$

which means that the city boundary increases with the critical VOT at the bottleneck.

From Eqs. (14), (15), (18) and (20), we can derive

$$\frac{dn_D(x)}{d\alpha^*} = \left(-\frac{e^{ra}b}{e^{2ra}-1}re^{rx} - \frac{e^{ra}b}{e^{2ra}-1}re^{-rx}\right) < 0, \tag{F.2}$$

and

$$\frac{dn_S(x)}{d\alpha^*} = \frac{-br}{\sqrt{(b\alpha^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}}\left(c_3e^{rx} + c_4e^{-rx}\right) > 0. \tag{F.3}$$

Eqs. (F.2) and (F.3) indicate that the residential density of households in the downtown (or in the suburb) decreases (or increases) as the critical VOT at the bottleneck increases.

From Eqs. (14) and (23), we can derive

$$\frac{dx_D(\alpha)}{d\alpha^*} = \frac{e^{ra}b}{e^{2ra}-1}\frac{-2c_1c_2 - 2c_1^2 + (b\alpha)^2 - b\alpha\sqrt{(b\alpha)^2 - 4c_1c_2}}{rc_1\left(-b\alpha + \sqrt{(b\alpha)^2 - 4c_1c_2}\right)\sqrt{(b\alpha)^2 - 4c_1c_2}} > 0. \tag{F.4}$$

Similarly, from Eqs. (18) and (23), we have

$$\frac{dx_S(\alpha)}{d\alpha^*} = \frac{1}{r}\frac{b}{\sqrt{(b\alpha^*)^2 - (b\underline{\alpha})^2 + 2r_A/(kr^2)}} > 0. \tag{F.5}$$

Eqs. (F.4) and (F.5) mean that the residential location of households increases as the critical VOT increases.

In order to compare the spatial structure differences (city length, residential density, and residential location) among the no-toll optimal, first-best, and second-best cities, we only need to compare the values of α^*, α^*_{FB}, and α^*_{SB}.

We first prove $\alpha^*_{FB} > \alpha^*_{SB}$ by contradiction method. The capacity investment cost is assumed to satisfy a constant return to scale $IC = \delta q$. Denote q^*_{FB} and q^*_{SB} as the optimal capacities in the first-best city and the second-best city, respectively. We assume that $\alpha^*_{FB} < \alpha^*_{SB}$ holds. Comparing equilibrium conditions Eqs. (42) and (E.1), one can obtain

$$\frac{1}{q^*_{SB}}(\alpha^*_{SB} - \underline{\alpha})(\alpha^*_{SB} + \underline{\alpha}) < \frac{1}{2q^*_{SB}}(\alpha^*_{SB} - \underline{\alpha})(3\alpha^*_{SB} + \underline{\alpha}) < \frac{1}{2q^*_{FB}}\left((\alpha^*_{FB})^2 - \underline{\alpha}^2\right), \tag{F.6}$$

$$\frac{\alpha^*_{SB} - \underline{\alpha}}{q^*_{SB}} < \frac{1}{2}\frac{\alpha^*_{FB} - \underline{\alpha}}{q^*_{FB}}. \tag{F.7}$$

On the other hand, since the capacity investment cost exhibits a constant return to scale, the optimality condition implies $\frac{d\Phi_{FB}}{dq} = \frac{d\Phi_{SB}}{dq} = \delta$. From Eqs. (51) and (63), one obtains

$$\frac{\left((\alpha^*_{FB})^2 - \underline{\alpha}^2\right)}{2{q^*_{FB}}^2}\left(\frac{1}{3}b\alpha^*_{FB} - \frac{2}{3}\frac{b\underline{\alpha}^2}{\alpha^*_{FB} + \underline{\alpha}}\right) = \frac{\left((\alpha^*_{SB})^2 - \underline{\alpha}^2\right)}{2{q^*_{SB}}^2}\left(\alpha^*_{SB} - \underline{\alpha}\right). \tag{F.8}$$

From Eqs. (F.6) and (F.8), one obtains

$$\frac{\alpha^*_{FB} - \underline{\alpha}}{3}\frac{\left((\alpha^*_{FB})^2 - \underline{\alpha}^2\right)}{2{q^*_{FB}}^2} < \left(\frac{1}{3}\alpha^*_{FB} - \frac{2}{3}\frac{\underline{\alpha}^2}{\alpha^*_{FB} + \underline{\alpha}}\right)\frac{\left((\alpha^*_{FB})^2 - \underline{\alpha}^2\right)}{2{q^*_{FB}}^2} < \frac{\alpha^*_{SB} - \underline{\alpha}}{2q^*_{SB}}\frac{\left((\alpha^*_{FB})^2 - \underline{\alpha}^2\right)}{2q^*_{FB}}. \tag{F.9}$$

Combining the leftmost and rightmost terms in Eq. (F.9) yields

$$\frac{\alpha^*_{FB} - \underline{\alpha}}{q^*_{FB}} < \frac{3}{2}\frac{\alpha^*_{SB} - \underline{\alpha}}{q^*_{SB}}. \tag{F.10}$$

Clearly, Eq. (F.10) contradicts (F.7). Therefore, $\alpha^*_{FB} > \alpha^*_{SB}$ holds. According to Eqs. (F.1)-(F.5), one can obtain that the first-best city is longer than the second-best city, and any resident with a given VOT in the first-best city lives farther from the CBD than in the second-best city. Besides, the residential density and thus the land rent in the downtown area (or the suburban area) in the first-best city are lower (or higher) than in the second-best city. Similarly, one can prove $\alpha^*_{SB} < \alpha^* < \alpha^*_{FB}$ for $\underline{\alpha} = 0$ by using the contradiction method. This completes the proof of Proposition 11.

Appendix G: Numerical examples

Numerical examples are used to further illustrate the properties of the model and gain additional insights. Specifically, we aim to (i) investigate how the bottleneck location and current bottleneck capacity affect households' marginal benefits from capacity expansion, (ii) compare the marginal benefit of capacity expansion to society with and without urban spatial consideration when there is no toll, and (iii) examine the differences in both spatial structure and optimal capacity investment among the no-toll optimal, first-best, and second-best cities.

G1. Parameter specifications

Consider one monocentric linear transportation corridor, with a population size of N = 20000. The bottleneck is located 10 km from the CBD, i.e., $a = 10$ km. The agricultural rent r_A is \$100 per day. The average auto free-flow travel speed is 40 km per hour, i.e., $t_0 = 1/40$ h/km. The ratio, η, of the value of early-arrival time to the VOT is 0.3, as calibrated by Hall (2024). The lower and upper bounds of VOT α are $\underline{\alpha} = \$40$ and $\bar{\alpha} = \$120$ per hour, respectively. Besides, we suppose that residents' VOT is proportional to their wage, and $w(\alpha) = 8\alpha$ with a consideration that the work time is about 8 hours per day. The parameter k in the hyperbolic utility function (see Eq. (4)) is 0.05. The bottleneck capacity investment cost IC follows a power function, $IC = \delta q^{\rho}$, where δ is the coefficient and ρ represents the elasticity of capacity investment cost regarding capacity. The base values of δ and ρ are set as \$20 and 1, respectively.

G2. Distributional effects of bottleneck capacity expansion

Analytical results in the previous sections have shown that differential marginal benefits across different-income households due to bottleneck capacity expansion may depend on the bottleneck location and the current bottleneck capacity level. To detect their relationships, we consider three bottleneck locations of $a = 5$, 10, 25 km from the CBD, and two current bottleneck capacity levels of $q = 2000$ and 10000 veh/h. The marginal benefits of capacity expansion under different toll schemes are depicted in Fig. G.1.

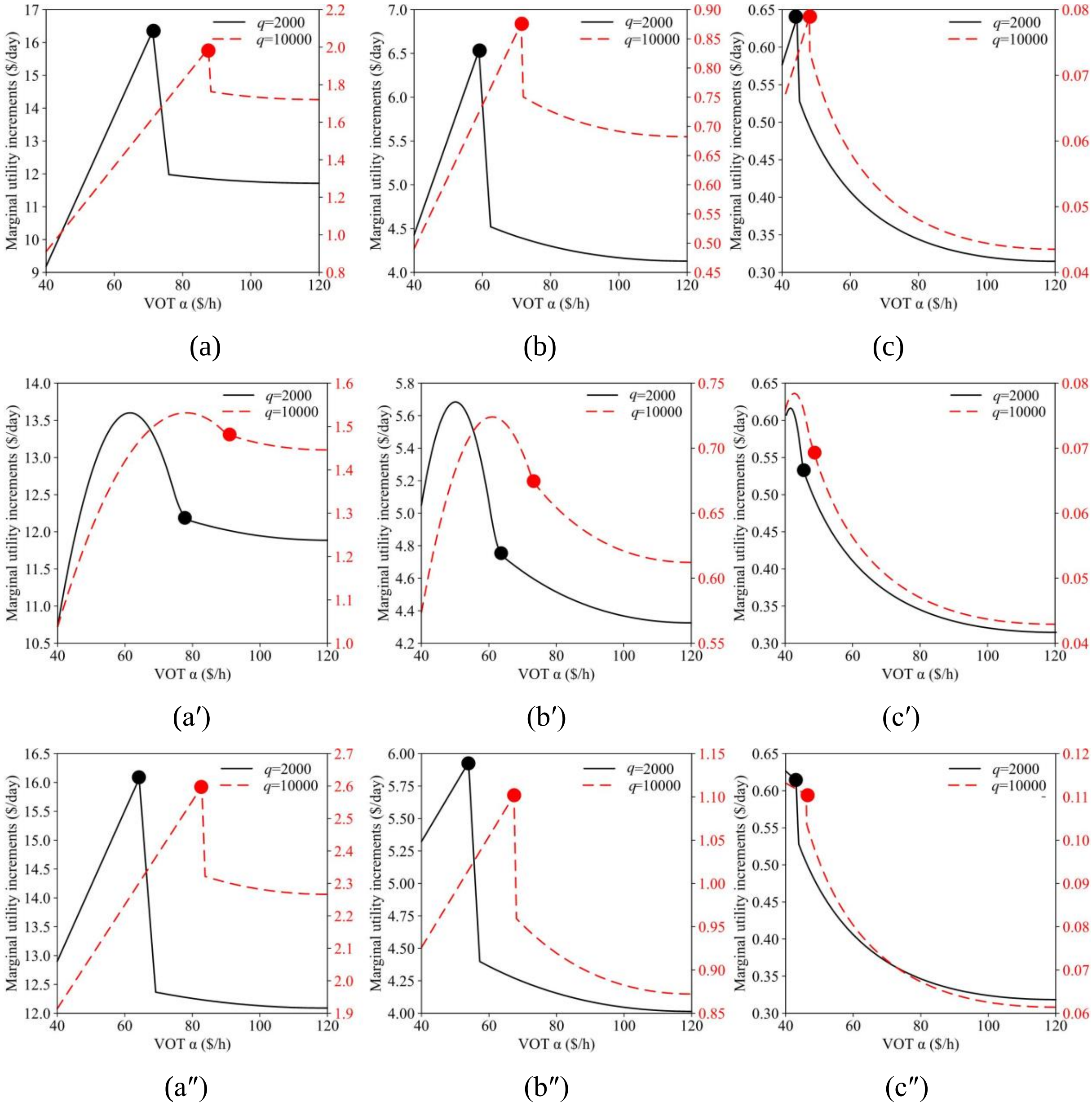

Fig. G.1. Marginal benefit of households from bottleneck capacity expansion. (a)-(c): no toll with a = 5, 10, 25km; (a′)-(c′): first-best toll with a = 5, 10, 25km; (a″)-(c″): second-best toll with a = 5, 10, 25km. The round points represent the households at the bottleneck.

It can be seen in Fig. G.1(a)-(c) that under no toll, the (relatively) mid-income households residing at the bottleneck location gain the most from the capacity expansion. The richest or the poorest gain the least: when the bottleneck is located very close to the CBD (i.e., $a = 5$), the poorest households gain the least; when it is located very far from the CBD (i.e., $a = 25$), the richest households gain the least; and when the bottleneck is positioned relatively in the middle of the city (i.e., $a = 10$), the current bottleneck capacity dominates the result: the richest households gain the least if the current capacity is small (i.e., $q = 2000$); otherwise, the poorest households gain the least if the capacity is large (i.e., $q = 10000$).

Under the first-best toll (see Fig. G.1(a′)-(c′)), the pattern of household benefits from capacity expansion is similar to that in the no-toll scenario, with the main difference being that households living at a suburban location (not the bottleneck) gain the most. Under the second-best toll (see Fig. G.1(a″)-(c″)), when the bottleneck location is very far from the CBD (i.e., $a = 25$), the household marginal benefit from capacity expansion decreases with their VOTs, i.e., the richer households benefit less. When the bottleneck location is not very far away from the CBD (i.e., $a = 5,\ 10$), the mid-income households at the bottleneck benefit the most, whereas the poorest or richest households benefit the least, depending on the bottleneck location and current capacity level.

G3. Marginal benefit of bottleneck capacity expansion with and without urban spatial structure consideration under no toll

To further confirm Proposition 4, Fig. G.2 presents the marginal benefit of bottleneck capacity expansion to society with and without urban spatial consideration under no toll. The initial capacity is 2000 veh/h, from which it is expanded. It shows that at the initial stage of capacity expansion (i.e., $q < 3300$ veh/h), ignoring urban spatial structure will overestimate the marginal benefit. As the capacity becomes large (i.e., $q > 3300$ veh/h), an underestimate occurs. As a result, ignoring urban spatial structure may cause overinvestment or underinvestment in the optimal bottleneck capacity, depending on the marginal cost of capacity expansion.

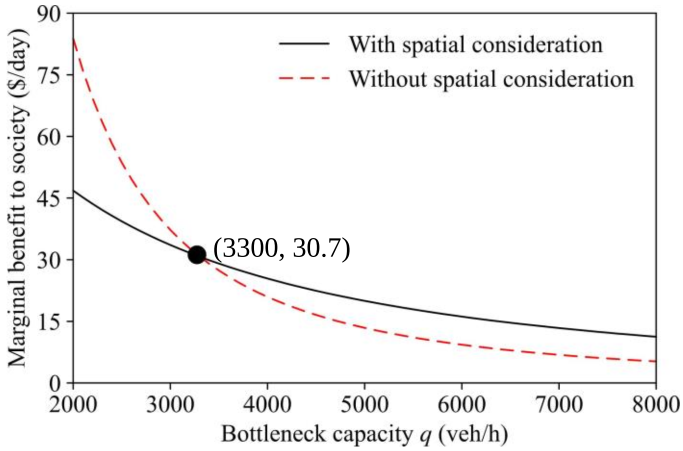

Fig. G.2. Marginal benefit of bottleneck capacity expansion under no toll.

G4. Difference in urban spatial structure and optimal capacity among no-toll optimal, first-best, and second-best cities

In Section 4.3, we have analytically proved that with constant return to scale for capacity investment (i.e., $\rho = 1$), the first-best city has a lower (higher) residential density in downtown (suburb), and a farther residential location than the second-best city for a given household (see Proposition 12). As the lower bound of household VOT $\underline{\alpha} = 0$, the no-toll optimal city lies between the first-best and second-best cities in terms of residential density and location. To check whether this conclusion is robust for $\underline{\alpha} \neq 0$, we conduct numerical analysis of $\underline{\alpha}$, with $\underline{\alpha}$ being \$20 and \$80 per hour, respectively. The results are shown in Fig. G.3.

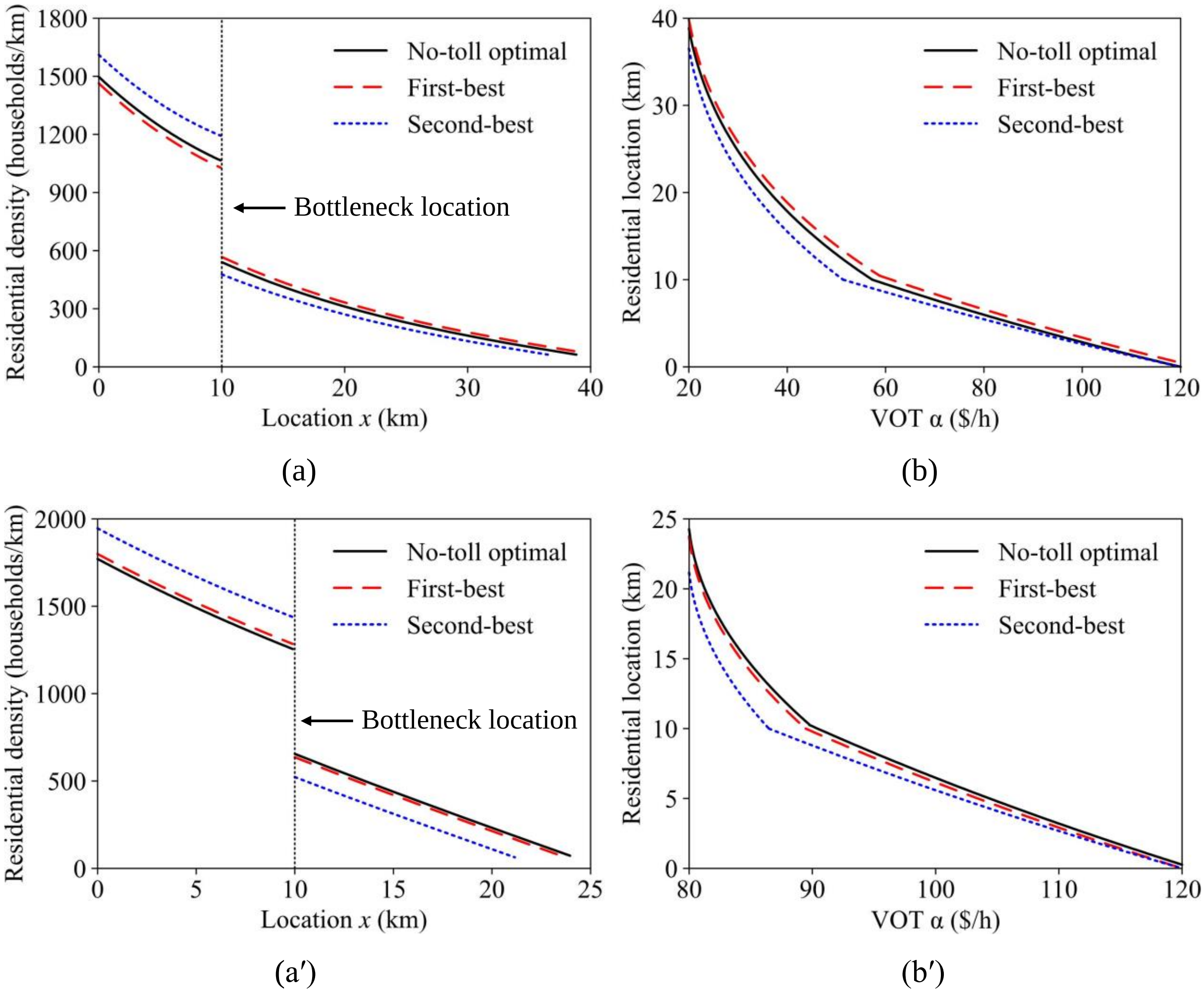

Fig. G.3. Residential density and residential location for no-toll optimal, first-best, and second-best cities. (a) and (b): $\underline{\alpha} = \$20/\text{h}$; (a′) and (b′): $\underline{\alpha} = \$80/\text{h}$.

Fig. G.3 shows that the first-best city always has a lower downtown residential density but a higher suburban suburban residential density, and a farther residential location than the second-best city, regardless of $\underline{\alpha}$. If the lower bound of household VOT $\underline{\alpha}$ is small (i.e., $\underline{\alpha} = \$20/\text{h}$), the no-toll optimal city lies between the first- and second-best cities in terms of residential density and location (see Fig. G.3(a) and (b)), which is consistent with Proposition 12. However, if $\underline{\alpha}$ is large (i.e., $\underline{\alpha} = \$80/\text{h}$), the no-toll optimal city has the lowest (highest) downtown (suburban) residential density, and the farthest residential location from the CBD among the three cities (see Fig. G.3(a′) and (b′)). With non-constant return to scale for capacity investment cost (i.e, $\rho \neq 1$), similar results can be obtained, but not shown here for saving paper space.

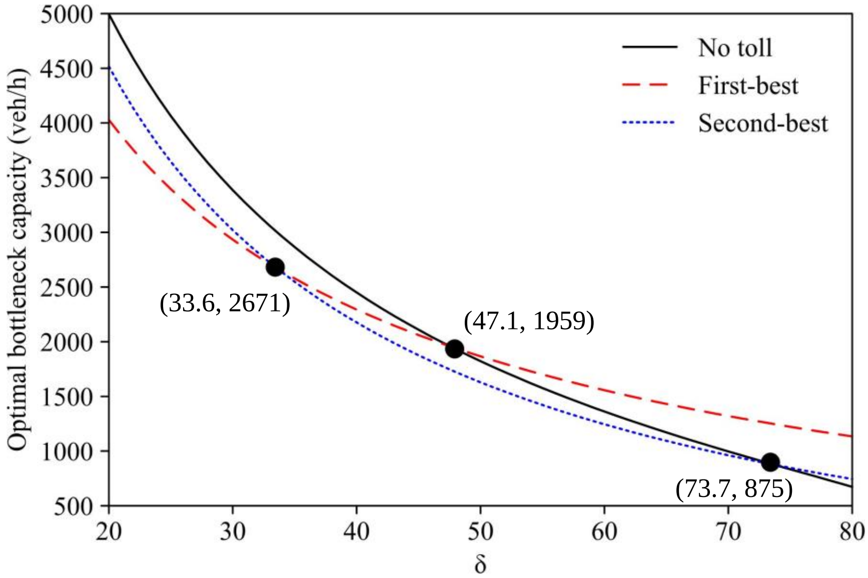

Fig. G.4. Optimal bottleneck capacity vs. marginal investment cost δ under no toll, first-best toll, and second-best toll.

In addition, it is difficult to analytically determine the relationships of optimal capacity investments among the no-toll optimal, first-best, and second-best cities. As an example, Fig. G.4 shows the changes of optimal capacity investment with the marginal capacity investment cost δ numerically. It turns out that as δ is small ($\delta < 47.1$), the no-toll optimal city requires the largest capacity investment; Otherwise, the first-best city requires the largest investment. The lowest capacity investment may occur in any one of the three optimal cities, depending on the value of δ.