TI 2024-011/VIII
Tinbergen Institute Discussion Paper

Pricing in the Stochastic Bottleneck Model with Price-Sensitive Demand

Revision: February 2025

Qiumin Liu[1]
Vincent A.C. van den Berg[2]
Erik T. Verhoef[3]
Rui Jiang[4]

1 Beijing Transport Institute, Beijing Jiaotong University and VU Amsterdam
2 VU Amsterdam and Tinbergen Institute
3 VU Amsterdam and Tinbergen Institute
4 Beijing Jiaotong University

Pricing in the Stochastic Bottleneck Model with Price-Sensitive Demand

Qiumin Liu [a,b,c], Vincent A.C. van den Berg [c,d,#], Erik T. Verhoef [c,d], Rui Jiang [b]

[a] *Beijing Transport Institute, Beijing 100073, China*

[b] *School of Systems Science, Beijing Jiaotong University, Beijing 100044, China*

[c] *Department of Spatial Economics, VU Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands*

[d] *Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS Amsterdam, The Netherlands*

[#] Corresponding author: v.a.c.vanden.berb@vu.nl

Abstract

We analyse time-varying tolling in the stochastic bottleneck model with price-sensitive demand and uncertain capacity. We find that price sensitivity and its interplay with uncertainty have important implications for the effects of tolling on travel costs, welfare and consumers. We evaluate three fully time-variant tolls and a step toll proposed in previous studies. We also consider a uniform toll, which affects overall demand but not trip timing decisions. The first fully time-variant toll is the 'first-best' toll, which varies non-linearly over time and results in a departure rate that also varies over time. It raises the generalised price (i.e. the sum of travel cost and toll), thus lowering demand. These outcomes differ fundamentally from those found for first-best pricing in the deterministic bottleneck model. We call the second toll 'second-best': it is simpler to design and implement as it maximises welfare under the constraint that the departure rate is constant over time. While a constant rate is optimal without uncertainty, it is not under uncertain capacity. Next, 'third-best' tolling adds the further constraint to the second-best that the generalised price should stay the same as without tolling. It attains a lower welfare and higher expected travel cost than the second-best scheme, but a lower generalised price. All our other tolls raise the price compared to the no-toll case.

In our numerical study, when there is less uncertainty: the second-best and third-best tolls achieve welfares closer to that of the first-best toll, and the three schemes become identical without uncertainty. As the degree of uncertainty falls, the uniform and single-step tolls attain higher welfare gains. Also, when demand becomes more price-sensitive, the uniform and single-step tolls attain relatively higher welfare gains. Our step toll would lower the generalised price without uncertainty but raises it in our stochastic setting.

Keywords: stochastic bottleneck model; price-sensitive demand; time-varying toll; step toll; uncertainty

JEL codes: R41; R48; D62; D80

Version of Feb. 2025

This is a pre-print version of the paper accepted in *Transportation Research Part B: Methodological*; article number: 103176; https://doi.org/10.1016/j.trb.2025.103176 . Please cite the published version.

1. Introduction

Congestion is one of the greatest challenges facing cities worldwide. Congestion causes different kinds of disutilities, including loss of time and inconveniences due to rescheduling, unpredictability and uncertainty. Travel conditions may vary due to a combination of exogenous shocks, including weather conditions and traffic incidents, and endogenous demand responses to those shocks. Especially when those responses take the form of rescheduling—for example, by departing earlier to create time "buffers"—the analysis of congestion and policies to address it requires models that can deal with the dynamics of departure times, the impact of that on dynamic patterns of travel delays and the feedback of that upon behaviour, to obtain a consistent representation of stochastic dynamic equilibria that can be used for policy evaluation. That is what our paper aims to offer, studying policies to address dynamic congestion in a stochastic setting where travellers can respond in terms of departure time choice, but also travel choices more generally in the sense that we will consider price-sensitive demand.

Our analysis employs the stochastic bottleneck model. We analyse different types of fully time-variant tolling, uniform tolling, and step tolling all under uncertain capacity and price-sensitive demand. As discussed below, there is extensive literature on the untolled equilibrium of the bottleneck model with uncertain capacity. Only a few papers consider time-variant tolling, and none consider price-sensitive demand. As we will see, the interplay between uncertainty and price sensitivity has important effects and complicates analysis. The effects of time-variant tolling also differ from the effects of the deterministic model. Some papers have looked at uniform tolling, where the toll is constant (Lam, 2000; Zhu et al., 2018; Jiang et al., 2021). In particular, Zhu et al. (2018) used a bottleneck model with uncertainty in the free-flow travel time that does not affect queuing. So, there is no interaction between uncertainty and the dynamics of congestion, and their model provides insights that are like those in the deterministic model, as the social optimum has no queuing. Conversely, in our model, the social optimum will have queuing when capacity is low. Our approach seems realistic but is difficult to analyse.

The three fully time-variant tolls that we consider follow from earlier studies with fixed demand. The first follows Lindsey (1994, 1999), and we call it 'first best' because it attains an overall social optimum. The first best leads to a departure rate that weakly increases over the morning, and it raises the generalised price (i.e. the sum of travel cost and toll) compared to the untolled case. The first best has a smaller reduction in travel cost and a smaller welfare increase than under certainty: the former leaves the price unchanged and halves the average travel cost. This outcome has important implications for the political feasibility of tolling and its desirability vis-à-vis alternative policies such as capacity expansion, travel credits, and flexible working hours. Our results deviate from those for a deterministic bottleneck in ways comparable to results under dynamic flow congestion (Chu, 1999; Mun, 1999, 2003).

Long et al. (2022) proposed what we call a 'second-best time-varying toll'. It maximises welfare under the conditions that the departure rate is constant over time and the toll starts and ends at zero. This greatly simplifies the scheme's design for the researcher and government, and it matches what is optimal in the deterministic bottleneck model; however, as we will show, it is 'second best' under uncertain capacity as it lowers welfare compared to first-best pricing. We also extend the analysis of Long et al. (2022) by adding price-sensitive demand and optimising total demand in the second step of the analysis.

Our third toll follows Xiao et al. (2015). We call it a 'third-best time-variant toll' because it adds an extra constraint to the second-best case: the generalised price should be at its untolled level. In fact, there is only one constant departure rate with a toll starting at zero and a price at its untolled level. Therefore, the third-best toll needs no maximisation of welfare. The third-best toll does not hurt users by raising the generalised price. In contrast, all our other tolls raise the price. This helps with the political and social feasibility of tolling. However, this advantage comes with the downside of a lower welfare. This makes it important to see how large this welfare loss is and how much better off the consumer is. This allows us to determine if it may be worth it. Again, we extend Xiao et al.'s (2015) model by adding price-sensitive demand. There is a large literature on Pareto-improving tolls under certainty. See, for instance, Lawphongpanich and Yin (2010), Tan et al. (2016) and Hall (2018). As in our setting, adding price-sensitive demand makes tolls that do not hurt users more challenging, and less beneficial for welfare.

The analysis of uniform (time-invariant) and step tolling (where the toll changes only in discrete steps) is important. In reality, tolls are not fully time-variant; they are uniform—as in London—or at most have a few steps in them—as in Singapore and on some US pay-lanes and bridges. We will see that the combination of uncertainty and price-sensitive demand changes how these coarse tolls perform compared to first-best tolling. A uniform toll has no effect if demand is fixed: it cannot alter departure rates directly but only affects total demand. We use the ADL step toll of Arnott et al. (1990).

Xiao et al. (2015) and Long et al. (2022) also considered single-step tolling, as did Jiang et al. (2022). However, they all used fixed demand. Under uncertain capacity, Yu et al. (2023) considered uniform tolling in conjunction with information provision. Jiang et al. (2021) and Zhu et al. (2018) analysed uniform tolling in a bottleneck model with an uncertain free-flow travel time.

Our core policy contribution is the comparison of various tolling policies when considering the interplay of uncertainty in congestion and price-sensitive demand. The literature has focused on the second-best toll as the best realistically feasible. But how much worse does it perform than the first best? Is its ease of use worth the lower welfare? All our tolls, except the third-best, hurt consumers by raising the generalised price. So, the third best may be easier politically to implement. But how much lower is its welfare, and how much does it help consumers? All these are important questions, but also questions that are absent without the uncertainty. Moreover, with fixed demand, one can always set the toll so that it does not hurt users. So, again, the interaction between uncertainty and price-sensitive demand is vital. Real-world tolls vary in steps or are uniform. How do these tolls compare with the time-variant ones? And how does this differ from under deterministic congestion? It is important for policymaking to have the answers to these questions, also if the required models become too complex to allow for intuitive closed-form solutions. As this is the situation in our setting, we will complement and extend the analytics using numerical analyses.

Our core methodological contributions are threefold: 1) We study time-variant tolling in the stochastic bottleneck model with uncertain capacity and price-sensitive demand; the interaction of these two will prove important and has been mostly ignored in the literature. Price sensitive demand lowers benefits from the second- and third-best cases because they equate the generalised price to the marginal social cost. 2) We present dynamic optimisation specifications to determine our settings using a Hamiltonian and optimisation in two steps. 3) The analysis of first-best tolling under price-sensitive demand and uncertain capacity is much more complex than it is for the second- and third-best tolls. But this complexity is needed to test if their relative ease of use for the regulator is worth it. We optimise using two steps. The first, comparable to the existing literature about fixed demand, involves optimising the departure rate and the toll pattern for a given number of travellers, using optimal control theory. This gives the optimal dynamic pattern *given* the number of users. The second step then optimises total demand and the starting level of the toll, whilst considering that for any total demand to be found, the optimal dynamic pattern from the first step will apply. This second step, which is vital when demand is price-sensitive, has not been considered before and will be shown to have important implications. Note that the optimality conditions from these two analytical steps simultaneously characterize the full optimum and are therefore not to be understood as phased actions by the toll authority as occurs in multi-stage games.

Table 1: A comparison of our paper with the literature

Citation	Distribution of uncertainty	Capacity is constant during the day	Bottleneck; uncertain capacity	Bottleneck; uncertain free-flow time	Price-sensitive demand	First-best toll	Second-best toll	Third-best toll	Step toll	Uniform toll	No-toll equilibrium
Arnott et al., 1991	Two-point distribution	Yes	√	-	-	-	-	-	-	-	√
Arnott et al., 1996, 1999	General distribution	Yes	√[2]	-	√	-	-	-	-	-	√
Lindsey, 1994, 1999	General & two-point distribution	Yes	√[2]	-	-	√	-	-	-	-	√
Long et al., 2022	General distribution	Yes	√	-	-	-	√	√	√	-	√
Xiao et al., 2015	Uniform distribution	Yes	√	-	-	-	-	√	√	-	√
Jiang et al., 2021	General distribution	Yes	-	√	√	-	-	-	-	√	√
Jiang et al., 2022	Uniform distribution	Yes	√	-	-	-	-	-	√	-	√
Zhu et al., 2018	Uniform distribution	Yes	-	√	√	√[1]	-	-	-	√	√
Yu et al., 2023	Two-point distribution	Yes	√	-	√	-	-	-	-	√	√
Chu, 1999	None	Yes	-	-	√	√	-	-	√	√	√
Mun, 1999, 2003	None	Yes	-	-	√	√	-	-	-	-	√
Fosgerau and Lindsey, 2013	General distribution	No	√	-	-	√	-	-	-	-	√
Hall and Savage, 2019	General distribution	No	√	-	-	√	-	-	-	-	√
Peer et al., 2010	Two-point distribution	No	√	-	-	-	-	-	-	-	√
This paper	Uniform distribution	Yes	√	-	√	√	√	√	√	√	-

Note: [1] For a more limited 'exogenous' distribution of free-flow travel time, Zhu et al. (2018) analyse a time-variant toll that works the same way as in the deterministic bottleneck model, and this is the first-best in their setting.

[2] These authors also consider uncertain demand.

Table 1 shows how our work relates to the literature and how we extend it. It shows that no previous studies considered time-varying tolling in the stochastic bottleneck model with uncertain capacity under price-sensitive demand. Section 2 reviews the literature, including works we have not discussed above. Section 3 explains the setup. Section 4 derives the socially optimal 'first-best' toll and compares it to that of the deterministic model and to the second- and third-best tolls. Sections 5 and 6 look at the uniform and single-step toll. Section 7 conducts a numerical study. Section 8 concludes.

2. Extended literature review

Early works on uncertainty in the bottleneck model include Arnott et al. (1991, 1996, 1999) and Daniel (1995). Arnott and co-authors were primarily interested in the effects of information provision and Daniel in competition among airlines. A very large literature on uncertainty in the bottleneck model has built on them. Small et al. (2024), Small (2015) and Li et al. (2020), among others, provide extensive overviews.[1] But few papers look at congestion pricing,[2] and fewer still include price-sensitive demand. Instead, most studies have investigated untolled equilibrium or information provision.

Zhu et al. (2018) analysed time-varying tolling under price-sensitive demand using a bottleneck model with uncertain free-flow travel time that does not affect queuing. This leads to an outcome similar to that in the deterministic setting, as queuing can be fully eliminated. In contrast, in our model, the social optimum has queuing in 'bad' states. Their model also misses the interaction between queuing and uncertainty. All this makes their model more tractable, but arguably less realistic, and yielding different policy implications as it makes tolling appear better for welfare and less harmful for consumers.

Yu et al. (2023) considered information provision and uniform tolling under uncertain bottleneck capacity and price-sensitive demand. Lam (2000), Jiang et al. (2021) and Zhu et al. (2018) analysed uniform tolling in a bottleneck model with an uncertain free-flow travel time. Xiao et al. (2015), Jiang et al. (2022) and Long et al. (2022) also considered single-step tolling under fixed demand. The literature that is most directly related to our study looks at time-variant tolling in the stochastic bottleneck model but under fixed demand. This literature yields the first-, second- and third-best tolls, as previously discussed (see Lindsey, 1994, 1996, 1999; Xiao et al., 2015 and Long et al., 2022). Finally, Zhang et al. (2018) studied a bottleneck model where the capacity drops by a random value when congestion reaches a certain level of severity, and they analysed time-varying and step tolling. Their model is similar to that of Zhu et al. (2018)—who used an uncertain free-flow travel time—in that optimal tolling removes all queuing. This is not true under our uncertain capacity, thereby complicating the analysis and making tolling less beneficial.

Uncertainty can take various forms in the bottleneck model. Previous studies have looked at: i) uncertain capacity (e.g. Arnott et al., 1991; Xiao et al., 2015; Long et al., 2022; Jiang et al., 2022); ii) uncertain demand (e.g. Fosgerau, 2010) iii) both uncertain capacity and demand (e.g. Arnott et al., 1996, 1999) iv) uncertain arrival times at the bottleneck (e.g. Daniel, 1995); v) uncertain free-flow travel time (e.g. Zhu et al., 2018; Lam, 2000; Siu and Lo, 2009; Jiang et al., 2021); and vi) uncertainty in the demand

[1] There is also a large body of literature that considers static congestion (see Zhang et al. (2022) for a detailed review). To examine the interaction between information and pricing instruments, Verhoef et al. (1996) used the static model. They found that information and tolling are nearly perfectly complementary in the face of stochastic congestion. This finding has then been extended by also considering networks (Yang, 1999; Maher et al., 2005; Meng and Liu, 2011; Lindsey et al., 2014; Klein et al., 2018).

[2] Many papers have considered alternative policies to reduce congestion and uncertainty. These policies include information provision (e.g. Arnott et al., 1991, 1996, 1999; Liu and Liu, 2018; Zhu et al., 2019; Yu et al., 2021; Han et al., 2021; Yu et al., 2023); ride-sharing (Long et al., 2018; Li et al., 2022; Liang et al., 2023); on-demand buses (Ma et al., 2023); tradable credits (Zhang et al., 2022); flexible working hours (Xiao et al., 2014b); and merging rules (e.g. Xiao et al., 2014a). Fosgerau (2010) evaluated the relationship between the mean and variance of delays. Lindsey (2009) studied self-financing under random capacity and demand.

function such as a random demand intercept (Fu et al., 2018). For uncertainty in capacity, most papers assume—as we do—that capacity is uncertain but its realised value is constant throughout the peak. A few studies have variable capacity within the day due to, for example, an accident that is cleared after an hour or a rain shower. Exceptions include Fosgerau and Lindsey (2013), Peer et al. (2010), Hall and Savage (2019), and Schrage (2006).[3]

We follow much of the literature in using a uniform distribution for the uncertainty, as this allows for more closed-form results. Xiao et al. (2014a, 2014b, 2015), Jiang et al. (2022) and Zhang et al. (2018) also used a uniform distribution. Arnott et al. (1991), Liu et al. (2020), and Yu et al. (2020, 2023) used two-point distributions, which seem more restrictive than a continuous distribution. Lindsey (1994, 1999), Long et al. (2022), Jiang et al. (2021) and Liu et al. (2023) used more general distributions.

Most papers—like ours—assume that drivers are rational and that they consider their expected price. So, we abstain from considering risk aversion or bounded rationality. Li et al. (2008) considered risk-aversion by adding the standard deviation of travel time to the user cost function, thus not only considering the expected cost. In Liu et al. (2020) and Jiang et al. (2022), users considered a linear combination of the mean cost and its variation. Siu and Lo (2009), Liu and Liu (2018), and de Palma and Fosgerau (2013) also considered risk aversion. Zhu et al. (2019) considered bounded rationality.

Fully time-variant tolls are practically impossible to implement in reality. More realistic toll schedules are uniform tolls that are constant over the day or step tolls with one or a few discrete steps in the toll. In the deterministic bottleneck model, Arnott et al. (1993), Laih (1994), Lindsey et al. (2012) and Ren et al. (2016) proposed four different equilibrium models to examine such schemes. They differ in how to ensure that the generalised price is the same before and after the toll is lowered at time t^-. The ADL model of Arnott and co-authors has a mass departure for arrivals after t^-. However, the Laih model has separate queues for arrival before and after t^- that do not interact. In the Braking model of Lindsey and co-authors, the first drivers who will arrive after t^- brake and temporarily completely block the road to prevent having to pay the (higher) toll or be overtaken. Finally, Ren et al. (2016) developed a model in between the Laih and braking model, where there are separate queues but the queue for arrivals after t^- hinders other drivers while not fully blocking the road. Van den Berg (2012) extended these models by adding price-sensitive demand and found that more steps can increase welfare gain and make consumers better off. Whilst considering uncertainty, Xiao et al. (2015), Long et al. (2022), Jiang et al. (2022), and Zhang et al. (2018) considered single-step tolling, but they used a fixed demand. The first three papers used the ADL equilibrium model, and the last one used the Laih model.

We study the three proposed time-varying tolls for the bottleneck model with uncertain capacity whilst adding price-sensitive demand. As we will see, this complicates analysis and has important effects. We also study uniform tolling and step tolling using the ADL equilibrium model. Further, we use a uniform distribution of the service time of the bottleneck—that is, the inverse of capacity—and this uncertain capacity varies over days but is constant throughout the peak.

[3] The first three papers used the bottleneck model; the last one followed Henderson (1974) and had dynamic flow congestion. Fosgerau and Lindsey (2013) included a random incident chance, where an incident temporarily lowers capacity by a fixed amount. There is at most one incident per day. In Peer et al. (2010), at each point in time, there is a fixed chance of an incident that lowers capacity, the capacity then remains low for the rest of the day. In Hall and Savage (2019), there is a random threshold—which varies over the days—and if the queue gets larger than the threshold, the capacity drops. In Schrage (2006), accidents follow a Poisson process. When one occurs, it temporarily lowers capacity, which thereafter gradually restores. The probability of an accident is independent of the time since the last accident but increases with traffic flow.

3. Model set-up

We assume that the bottleneck capacity is constant within a day but changes stochastically from day to day. We define the 'service time' of the bottleneck as $\phi = 1/s$, where s is the capacity. The service time follows a uniform distribution over an interval $[\phi_{min}, \phi_{max}]$, and $f(\phi) = 1/(\phi_{max} - \phi_{min})$ is the probability density function (PDF) of ϕ. This definition simplifies the mathematics but makes equations less intuitive. Commuters are unaware of capacity realisation on a given day before departure. From their day-to-day travel, they learn about capacity distribution and make their departure time choices by minimising their expected generalised price (Arnott et al., 1996; Lindsey, 1999; Xiao et al., 2015; Long et al., 2022). As discussed in the literature review, a uniform distribution of uncertainty is common in the literature. Extending our setting to a general distribution is an intriguing avenue for future work.

In the bottleneck model (Vickrey, 1969), commuters travel from home to work through the bottleneck. Without loss of generality, we assume that the free-flow travel time is zero. Thus, the travel time when departing at *t* equals the queuing time at the bottleneck $q(t,\phi)$, where $1/\phi$ is the realised capacity. Let $\omega(t)$ denote the maximum service time so that there is no queue at *t*. Following Lindsey (1994, 1996), the queuing time is

$$q(t,\phi) = \begin{cases} \phi \int_{\hat{t}}^{t} r(x)\, dx - (t - \hat{t}), \phi > \omega(t), \\ 0, \phi \leq \omega(t) \end{cases} \tag{1}$$

where $\hat{t}$ is the time when the queue begins to increase from zero and $r(t)$ is the departure rate that is the same for all capacity realisation as people depart without knowing what the capacity will be. For simplicity of notation, we assume that the departure rates are such that the queue starts to develop only once. This will prove true later on. The $\omega(t)$ is, for now, an unknown function, which we derive later using the explicit capacity distribution.

The travel time cost for commuters departing at *t* is

$$CTT(t,\phi) = \alpha\, q(t,\phi). \tag{2}$$

We define the schedule delay cost for commuters departing at *t* as:

$$SDC\big(t + q(t,\phi)\big) = \beta\, max\big\{t^* - \big(t + q(t,\phi)\big), 0\big\} + \gamma\, max\big\{\big(t + q(t,\phi)\big) - t^*, 0\big\}, \tag{3}$$

where α, β and γ are the values of time, schedule delay early and late, respectively. The t^* is the desired arrival time. Since capacity is stochastic, commuters departing at the same moment each day may experience different costs on different days. The expected travel cost at *t* is

$$E\big(C(t)\big) = \int_{\phi_{min}}^{\omega(t)} SDC(t) f(\phi) d\phi + \int_{\omega(t)}^{\phi_{max}} \big[CTT(t,\phi) + SDC\big(t + q(t,\phi)\big)\big] f(\phi) d\phi. \tag{4}$$

We remind the reader that $f(\phi)$ is the PDF of ϕ and $\omega(t)$ is the maximum service time for which no queue exists at *t*. The expected (generalised) price includes the travel cost and toll:

$$P_j(t) = E\left(C_j(t)\right) + \tau_j(t), \tag{5}$$

where *j* indicates the setting such as first best or untolled. In dynamic user equilibrium, this expected price should be constant over time during all used departure moments (and not lower at other times). In user equilibrium, this expected price should be equal to the marginal benefit as given by the inverse demand function. Users know the PDF of the service time (and thus the capacity and the resulting travel times), but they all depart before the capacity becomes known. The toll is independent of the state but can vary over time. Users know what the toll will be when they depart at time *t*.

When $j = NT$, we have no tolling and $\tau_{NT} = 0$. We also consider a uniform toll—indicated by *UT*—that is constant throughout the peak and a step toll—indicated by *ST*—that varies in two discrete steps. Finally, we have three fully time-varying tolls. Our first-best toll is indicated by *FB*. The second-best toll is represented by *SB* and maximises welfare whilst imposing that the departure rate should be constant and that the toll should start and end at zero. It is adapted from Long et al. (2022). Finally, the third-best toll (adapted from Xiao et al., 2015) is denoted by *TB*. It adds a further constraint to the second-best toll, namely, that the generalised price must stay at its untolled level. Only one unique constant departure rate satisfies the three constraints of the third-best scheme. So, it needs no maximisation of welfare. The second-best toll needs maximisation, but the constant departure rate makes this much simpler than for the first-best toll. We extend the models of Long et al. and Xiao et al. by adding price-sensitive demand. Price-sensitive demand lowers the benefit from the second- and third-best tolls because they do not ensure that the generalised price equals the marginal social cost. Price-sensitive demand also makes a Pareto-improving toll such as our third-best toll harder to be achieved.

The inverse demand function is $D(N)$, where N is the total demand. In user equilibrium, the expected price should equal the marginal willingness to pay as given by inverse demand: $P_j = D\left(N_j\right)$. The demand is independent of the capacity realisation, as no user knows capacity before departing. The regulator has the same information as the users and sets the toll only knowing the PDF and not knowing what the actual state will be;[4] hence maximising expected welfare or social surplus:

$$SS_j = \int_0^{N_j} D(n)\, dn - TC_j, \tag{6}$$

where TC_j is scenario j's total expected travel cost of $E\left(C_j(t)\right) \cdot N_j$. Hence, the regulator knows what demand will be if it adjusts the toll, what the departure rate will be, and what travel times will be in each state. But, like the users, for a given peak, the regulator does not know in advance what state will occur.

4. Social optimum under price-sensitive demand and capacity uncertainty

In the deterministic bottleneck model, the queue can be eliminated by a time-varying toll, and this achieves the social optimum in which the expected price remains unchanged vis-à-vis the no-toll equilibrium. As we will see, the toll in our stochastic bottleneck model will affect price and demand.

The untolled equilibrium for stochastic bottleneck capacity has been extensively studied, so we do not discuss it in detail. See Xiao et al. (2015) for a derivation under the same assumptions as ours (except that demand is fixed), or see Arnott et al. (1996), Lindsey (1999) and Long et al. (2022) for related settings.

The social surplus or welfare is

$$\max_{t_s, t_e, r_{FB}(t), N_{TV}} SS_{FB} = \int_0^{N_{FB}} D(n)\, dn - \int_{t_s}^{t_e} E\left(C_{FB}(t)\right) r_{FB}(t)\, dt, \tag{7}$$

where t_s, t_e, and $r_{FB}(t)$ denote the first departure time, the latest departure time, and the first-best departure rate at t, respectively. Then, the expected price follows Eq. (5) with $j = FB$. We solve for the first-best social optimum using two steps. In the first step, for a given demand, we optimise the departure rate using dynamic optimisation and minimisation of the expected total social cost. The optimal departure

[4] Future research could explore information provision and possible state-dependent tolls. Yu et al. (2023) did this for flat tolling. They found that state-dependent tolls did hardly better than state-independent tolls while being much harder to implement. Capacity that varies over the day or information about road condition becoming available during the day (from, say, news reports) would also be interesting extensions. However, these extensions are difficult to model. Hence, explaining the limited existing research on this.

rate, in turn, implies how the toll should change over time: the toll's pattern over time must be such that the expected price is constant over time. This step is similar to Lindsey (1994, 1999) in using optimal control theory,[5] except for our service time following a uniform distribution. The step is also akin to Fosgerau and Lindsey (2013). In the second step, we optimise total travel demand, which in turn implies the starting level of the toll. This step also considers the effect of a change in total demand on the outcome of the first stage. The two steps should not be interpreted as phased actions by the toll authority, as is common in multi-stage strategic games. Instead, the steps only concern the analytical sequence in our optimisation.

4.1. First step: analytics of optimising the departure rate and toll development over time

In the dynamic optimisation, the departure rate $r_{FB}(t)$ is the control variable. There are two state variables: 1) queuing time $q(t,\phi)$ from Eq. (1) and 2) cumulative departure $R(t)$, where $dR(t)/dt = r_{FB}(t)$. In the first step, the problem can be reformulated as the following minimisation:

$$\min_{t_s,t_e,r_{FB}(t)} \int_{t_s}^{t_e} E\big(C_{FB}(t)\big) r_{FB}(t)\, dt, \tag{8}$$

subject to

$$\frac{dq(t,\phi)}{dt} = \begin{cases} \phi\, r_{FB}(t) - 1, \phi > \omega(t) \\ 0, \phi \leq \omega(t) \end{cases}, \tag{9}$$

$$\frac{dR(t)}{dt} = r_{FB}(t), \tag{10}$$

where $R(t_s) = 0$ and $R(t_e) = N_{FB}$. Queue development depends on the capacity realisation, but the departure rate does not.

Let $\mu_1(t,\phi)$ and $\mu_2(t)$ denote the 'costate variables' of $q(t,\phi)$ and $R(t)$, respectively. We set up the equations so that $-\mu_2$ is the marginal social cost (MSC) of the total departures at *t*, that is, how much *higher* the total cost will be when the total departures at *t* are *higher*. Similarly, $\mu_1(t,\phi)$ is the shadow cost of queuing time when service time is ϕ (and thus capacity is $1/\phi$): it gives how much higher the total costs are when there is more queuing. The $\mu_1(t,\phi)$ not only depends on departure time *t* but also on the capacity realisation and thus ϕ. At any *t*, if ϕ is smaller (i.e. the capacity is larger), queuing will be shorter or even absent.

The Hamiltonian function for the minimisation is[6]

$$H(t) = r_{FB}(t)\left\{\int_{\omega(t)}^{\phi_{max}}\big[\alpha q(t,\phi) + SDC\big(t+q(t,\phi)\big)\big]f(\phi)d\phi + \int_{\phi_{min}}^{\omega(t)} SDC(t)f(\phi)d\phi\right\} +$$

$$\int_{\omega(t)}^{\phi_{max}} \mu_1(t,\phi)\,(\phi\, r_{FB}(t) - 1)\, f(\phi)d\phi + \mu_2(t) r_{FB}(t). \tag{11}$$

The optimality conditions are

$$\frac{dH(t)}{dr_{FB}(t)} = 0 = \int_{\omega(t)}^{\phi_{max}}\big[\alpha q(t,\phi) + SDC\big(t+q(t,\phi)\big)\big]f(\phi)d\phi + \int_{\phi_{min}}^{\omega(t)} SDC(t)f(\phi)d\phi +$$

$$\int_{\omega(t)}^{\phi_{max}} \phi\mu_1(t,\phi)f(\phi)d\phi + \mu_2(t) \tag{12}$$

[5] Yang and Huang (1997) used optimal control theory to analyse the deterministic bottleneck model, Mun (1999) to analyse his flow congestion model, Schrage (2006) and Yu et al. (2024) to analyse dynamic flow congestion.

[6] The Hamiltonian function may appear different than expected due to the integration over the uncertainty and the two possible outcomes: one with queuing and one without. Although (10) shows that there is a regime shift between these outcomes, there is no regime shift in the control variable r_{FB} because it depends on the expected outcome before the capacity is known. This is also why the first constraint for queuing is integrated over ϕ, while the second constraint for total departures is not. Thus, $\mu_1(t,\phi)$ depends on ϕ, whereas μ_2 does not.

$$\frac{d\mu_1(t,\phi)}{dt} = -\frac{\partial H}{\partial q} = \begin{cases} -r_{FB}(t)\left(\alpha + \frac{dSDC(t+q(t,\phi))}{dq(t,\phi)}\right), if\ \phi \geq \omega(t) \\ 0, if\ \phi \leq \omega(t) \end{cases} \tag{13}$$

$$\frac{d\mu_2(t)}{dt} = -\frac{\partial H}{\partial R} = 0. \tag{14}$$

$$H(t_s) = 0 \tag{15}$$

$$H(t_e) = 0 \tag{16}$$

Here, the last two equations (15) and (16) are the transversality conditions to set the freely chosen times for the start and end of the peak: t_s and t_e, respectively. The equations ensure that t_s and t_e are chosen to minimise total cost. Further, equations (13) and (14) govern the motion of the state variables of queue length and cumulative departures at *t*, respectively. Moreover, Eq. (12) determines the path of the control variable of the departure rate $r_{FB}(t)$.

Proposition 1:

The marginal social cost (MSC) is constant between t_s and t_e, and it equals $-\mu_2$:

$$MSC = -\mu_2$$
$$= \int_{\omega(t)}^{\phi_{max}} \left[\alpha q(t,\phi) + SDC\left(t + q(t,\phi)\right)\right] f(\phi) d\phi + \int_{\phi_{min}}^{\omega(t)} SDC(t) f(\phi) d\phi +$$
$$\int_{\omega(t)}^{\phi_{max}} \phi\, \mu_1(t,\phi)\, f(\phi) d\phi, \qquad t \in [t_s, t_e]. \tag{17}$$

Here, the first two terms give the expected private travel cost for a departure at t, and thus the expected marginal external cost (MEC) at t equals the last term:

$$MEC(t) = \int_{\omega(t)}^{\phi_{max}} \phi\, \mu_1(t,\phi)\, f(\phi) d\phi \geq 0. \tag{18}$$

The MEC is the expected cost due to queuing imposed on later departures at t. The MEC is zero at t_s and t_e, and non-negative in between.

Proposition 2:

Costate variable $\mu_1(t,\phi) \geq 0$ is the shadow cost of queuing time. The variable is non-negative, weakly decreases over departure time, and is zero for the last departures at t_e.

Proofs of Propositions 1 and 2:

Eq. (17) for μ_2 directly follows from the f.o.c. for the departure rate in (12). That μ_2 should be constant over time between t_s and t_e is clear from optimality condition (14). The first two terms in (17) are the expected private travel costs from (4). Given the problem set-up, $-\mu_2$ is the marginal social cost (*MSC*). By definition, the marginal external cost, *MEC*(t), is the difference between the private cost and MSC, and, thus, *MEC*(*t*) must equal the third term of eq.(17).

That $\mu_1(t)$ weakly decreases over time, follows from condition (13). For both f.o.c. (12) and transversality condition (16) to hold at t_e, it must be true that $\mu_1(t_e,\phi) = 0$ for any capacity realisation.

Finally, let us prove the non-negativity of *MEC*(*t*) and *MEC*(t_s) = *MEC*(t_e)=0. As ϕ, $\mu_1(t,\phi)$ and $f(\phi)$ are non-negative, the *MEC* must be as well. Lastly, at t_s, there will never be a queuing time as no queue has formed yet. This means that $\omega(t_s) = \phi_{max}$. Hence, at t_s, we integrate in (18) over a zero range, which makes *MEC*(t_s)=0. As we showed that $\mu_1(t_e,\phi) = 0$ for any ϕ, from (18), we can derive

$$MEC(t_e) = \int_{\omega(t)}^{\phi_{max}} 0\, d\phi = 0.$$

This completes the proof of Propositions 1 and 2. □

Remark 1:

That the shadow cost of queuing at the end of the peak is zero is logical. At t_e, all commuters have departed, and no later departures[7] *would be delayed by queuing regardless of the realised capacity. This outcome implies that no matter the capacity, the shadow cost of queuing at t_e is zero. Similarly, the earlier one departs, the more people one can harm by causing queuing. So, it is also logical that the shadow cost of queuing (weakly) decreases with the departure time and reaches zero at t_e.*

Now, we search for the pattern of the toll. For the expected price to be constant over time and thus for the users to be in user equilibrium, the toll must vary just like the expected *MEC*:

$$\tau(t) = \int_{\omega(t)}^{\phi_{max}} \phi\, \mu_1(t,\phi) f(\phi) d\phi + \tau_0, \tag{19}$$

where τ_0 is the starting level of the toll at t_s. Further, the toll must equal τ_0 at t_e.

We can use all this to derive the optimal departure rate, which must equal

$$r_{FB}(t) = 1/\omega(t),\ t \in [t_s, t_e], \tag{20}$$

where again $1/\omega(t)$ is the minimum of the stochastic capacity for which there is no queue at t. Intuitively, once a queue starts to develop, it does not collapse to zero until the last departure, as stated by Lindsey (1994, 1999). Suppose $t' \in [t_s, t_e]$ and $r(t') > 1/\omega(t')$. Then there will always be an interval where the expected delay cannot dissipate efficiently. Conversely, if $r(t') < 1/\omega(t')$, there will always be an interval in which capacity is not fully used.

Proposition 3:

The optimal departure rate $r_{FB}(t)$ is non-decreasing over departure time: $\frac{dr_{FB}(t)}{dt} \geq 0$.

Proof of Proposition 3: The proof is given in Appendix B.1 because the proof is rather long and technical.

Lemma 1:

For our uniform distribution, we get the following departure rates at the start and end of the peak:

$$r_{FB}(t_s) = 1/\phi_{max},$$
$$r_{FB}(t_e) = (\alpha + \gamma)/(\alpha\phi_{max} + \gamma\phi_{min}). \tag{21}$$

Proof of Lemma 1:

From the proofs of Propositions 1 and 2, we have $\omega(t_s) = \phi_{max}$. Then, $r_{FB}(t_s) = 1/\phi_{max}$. By differentiating Eq. (17), we have:

$$\omega(t)\mu_1\big(t,\omega(t)\big)\frac{d\omega(t)}{dt} = \int_{\phi_{min}}^{\omega(t)} \frac{dSDC(t)}{dt} d\phi - \int_{\omega(t)}^{\phi_{max}} \alpha d\phi\,.$$

As mentioned above, $\mu_1\big(t_e,\omega(t)\big) = 0$ at t_e. Then the right-hand side of the above equation equals zero at $t = t_e$. The last departure time cannot be earlier than the work start time, otherwise, the last traveller will shift to depart at t* with less schedule delay cost. Thus, we have $t_e \geq t^*$ and $\frac{dSDC(t)}{dt} = \gamma$ at t_e. Then, for our uniform distribution,

[7] Of course, one would delay the departures at t_e, but this effect has a zero size since you integrate from t_e to the same time t_e.

we have $\gamma(\omega(t_e) - \phi_{min}) - \alpha(\phi_{max} - \omega(t_e)) = 0$. The departure rate at t_e can be obtained as follows: $r_{FB}(t_e) = (\alpha + \gamma)/(\alpha\phi_{max} + \gamma\phi_{min})$. This completes the proof of Lemma 1. □

Hence, at the start of the peak, the departure rate is low and equals the minimum capacity of $1/\phi_{max}$. The departure rate ends high, being above the 'average' capacity of $1/\overline{\phi} = 2/(\phi_{max} + \phi_{min})$ but below the maximum capacity. The optimal departure rate continuously increases over departure times during at least a part of the peak and will never decrease. There may be departure windows when drivers always or never experience queuing, implying that the departure rate $r_{FB}(t)$ is constant over time in those ranges.

The departure rate starts low, as it is more costly if an early departure causes queuing. The optimal departure rate ends high because, for late departures, there are few or no users who can be affected. This outcome is also implied by the pattern of the shadow cost of queuing: $\mu_1(t,\phi)$.

4.2. Second step: setting total demand

Our first-step results with a uniform distribution are consistent with those of Lindsey (1994, 1999) for a general distribution, but our explicit distribution allows for more explicit results. Lindsey used a two-point distribution in the second part of the paper, and that leads to specific results. The uniform distribution we assumed leads to similar results as those for other continuous distributions. We now turn to the second step, where we set the total number of users to maximise the reduced-form social surplus (which is conditional on the departure rate following the socially optimal pattern):

$$SS_{FB} = \int_0^{N_{FB}} D(n)\,dn - N_{FB}E(C_{FB}^*(N_{FB})),$$

where the superscript * in C_{FB}^* indicates travel cost under the socially optimal departure rate. Maximising this objective with respect to N_{FB} gives:

$$D(N_{FB}) = E(C_{FB}^*(N_{FB})) + N_{FB}\frac{\partial E(C_{FB}^*(N_{FB}))}{\partial N_{FB}} = -\mu_2. \tag{22}$$

And thus, the inverse demand should equal the (expected) marginal social cost of $-\mu_2$.

Proposition 4:

The first-best toll at departure time t should equal the expected MEC(t), and it starts at t_s at zero and ends at t_e at zero:

$$\tau(t) = \int_{\omega(t)}^{\phi_{max}} \phi\,\mu_1(t,\phi)f(\phi)d\phi. \tag{23}$$

Proof of Proposition 4:

The above equation means that the expected price should equal the marginal social cost in the optimum. This occurs when the toll equals the *MEC*(t), which starts and ends at zero. So, in the first step's toll equation (19), the τ_0 must be zero, which gives us the above equation. For departure at t_s and t_e, the *MEC* is always zero, no matter the capacity, and the expected price equals the expected generalised cost while the toll is zero. □

This outcome is, of course, similar to the social optimum with a fixed capacity, where it is also optimal that the inverse demand equals the *MSC* and the toll equals *MEC*(t). However, in that case, the optimal departure rate equals the fixed capacity, whereas here the rate starts at the minimum capacity and then weakly increases to some final level that is in between the 'average' and the maximum. In the optimum, the peak is divided into different time windows: $[t_s, t_1]$; $[t_1, t_2]$; $[t_2, t^*]$; and $[t^*, t_e]$. They are denoted as '*Situations 1–4*', respectively. Commuters experience different levels of queuing, schedule delays, and prices in the various time windows. Because the details of the situations are rather technical, we discuss them in Appendix A.1. Using these details, we get the following results:

Lemma 2: *The cumulative departures are* $R(t_1) = \frac{t_1 - t_s}{\phi_{max}}$, $R(t_2) = \frac{t^* - t_s}{\phi_{max}}$ *and* $R(t^*) = \frac{t_1 - t_s}{\phi_{max}} + \frac{t^* - t_1}{\omega(t^*)}$.

Lemma 3: *In the social optimum, we have*:

$$t_s = \frac{2(\alpha+\gamma)(\beta+\gamma)\phi_{max}t^* - \gamma^2(\phi_{max}-\phi_{min})t_1}{2(\alpha+\gamma)(\beta+\gamma)\phi_{max} - \gamma^2(\phi_{max}-\phi_{min})} - \frac{\gamma\phi_{max}[\gamma(\phi_{max}+\phi_{min}) + 2\alpha\phi_{max}]N_{FB}}{2(\alpha+\gamma)(\beta+\gamma)\phi_{max} - \gamma^2(\phi_{max}-\phi_{min})} \tag{24}$$

$$t_e = \frac{2(\beta+\gamma)(\alpha\phi_{max}+\gamma\phi_{min})t^* + (2\beta+\gamma)\gamma(\phi_{max}-\phi_{min})t_1}{2(\alpha+\gamma)(\beta+\gamma)\phi_{max} - \gamma^2(\phi_{max}-\phi_{min})} + \frac{2\beta\phi_{max}(\alpha\phi_{max}+\gamma\phi_{min})N_{FB}}{2(\alpha+\gamma)(\beta+\gamma)\phi_{max} - \gamma^2(\phi_{max}-\phi_{min})} \tag{25}$$

Lemma 4: *The tolls at the first and the latest departure time are zero. The expected price of the first commuter is* $P_{FB} = \beta(t^* - t_s)$. *This is also the marginal social cost since an additional user departing before* t_s *does not impose any costs on others.*

Proofs: Appendixes B.2 and B.3 prove Lemmas 2 and 3. The last lemma directly follows from the earlier analysis.

To summarise, compared with the deterministic model, both the expected price and the total demand will change under tolling. Both departure rate and toll are now non-linear over time. The departure rate starts low at the start of the peak because it equals the minimum capacity; at the end of the peak, the rate is high but below the maximum capacity; in between these moments, the departure rate is non-decreasing, continuous and increasing for a range of departure times. Interestingly, this is the mirror image of the pattern without tolling: the untolled departure rate starts high, decreases in between, and ends low. All this differs from the deterministic bottleneck model, where first-best tolling leads to a constant departure rate and has the same price as the no-toll case.

4.3. Analytical comparison with the no-toll equilibrium and other proposed time-varying tolls

For the no-toll equilibrium, we use the results from Arnott et al. (1996, 1999), Linsdey (1994), Xiao et al. (2015), and Long et al. (2022) because our focus is on tolled equilibria. The analysis may be more complicated without tolling, as there are multiple cases of the no-toll equilibrium depending on the parameters, for example, with or without departure after t^*. In contrast, in the social optimum, there are always departures after t^* no matter what the parameters are.

By adopting the results of Long et al. (2022), the first and the last departure times under the no-toll equilibrium when ϕ follows the uniform distribution are given as follows:
(a) when there is departure after t^*,

$$t_s^{NT} = t^* - \frac{\alpha+\gamma}{\beta+\gamma} N\left[\frac{\phi_{min}+\phi_{max}}{2} - \frac{\alpha(\alpha\phi_{max}+\gamma\phi_{min}) + (\alpha+\gamma)\alpha\phi_{min}}{(\alpha+\gamma)^2(\phi_{min}+\phi_{max})}\right] \tag{26}$$

$$t_e^{NT} = t^* + N\left[\frac{\alpha\phi_{max}+\gamma\phi_{min}}{(\alpha+\gamma)} - \frac{(\alpha+\gamma)(\phi_{min}+\phi_{max})}{2(\beta+\gamma)} + \frac{\alpha(\alpha\phi_{max}+\gamma\phi_{min}) + (\alpha+\gamma)\alpha\phi_{min}}{(\alpha+\gamma)(\beta+\gamma)(\phi_{min}+\phi_{max})}\right]; \tag{27}$$

(b) when there is no departure after t*,

$$t_s^{NT} = t^* - N\hat{\phi} \tag{28}$$

$$t_e^{NT} = t^*, \tag{29}$$

where $\hat{\phi}$ is obtained by solving equation $\frac{1}{\hat{\phi}}\left(\frac{\phi_{min}+\phi_{max}}{2}-\frac{(\hat{\phi})^2-(\phi_{min})^2}{(\phi_{max})^2-(\phi_{min})^2}\right)+\frac{\hat{\phi}-\phi_{min}}{\phi_{max}-\phi_{min}}=\frac{\alpha+\beta+\gamma}{\alpha+\gamma}$. By comparing Eqs. (24)–(25) and (26)–(29), it becomes apparent that the departure timings under the social optimum differ from those under the no-toll equilibrium. Then, the price also changes under the social optimum, which differs from the results in the deterministic scenario.

The departure rate in the no-toll stochastic equilibrium is qualitatively the mirror image of the socially optimal rate. The untolled departure rate starts high, weakly decreases in between, and ends low. Numerical analysis by Xiao et al. (2015) shows that the rate may be flat in some periods, which we will also see in our numerical model for the first-best toll. We cannot analytically compare the prices in the no-toll and first-best equilibria. We will do this in the numerical model and find that no matter the parameters, the first-best toll has higher prices than the no-toll setting.

All this is also vastly different from the deterministic bottleneck model, where the no-toll rate is flat with a downward jump for arrivals at t^*. Moreover, the no-toll and first-best settings have the same prices in the deterministic model. The first-best toll in the stochastic model is concave over time, where it starts and ends at zero. Conversely, in the deterministic model, the toll is piecewise linear. Our stochastic bottleneck model thus has outcomes akin to dynamic flow congestion models (e.g. Agnew, 1977; Chu, 1999; Mun, 2003).

Let us compare the two other time-variant toll models in the literature. Long et al. (2022) proposed a time-varying toll that maximises welfare under the conditions that the departure rate is constant over time and the toll starts and ends at zero. This toll may have the advantage of being easier to design and implement, as the departure rate is constant. However, as the socially optimal rate varies (see also Lindsey, 1994, 1999), it must mean that their scheme has a lower welfare. Hence, we call it the 'second-best' time-variant toll, as the constraints lower welfare to some extent. For the second-best time-varying toll, we can further derive the first and the last departure time for our uniform distribution by following Long et al. (2022):

$$t_s^{SB}=t^*-\frac{N}{(\beta+\gamma)r_{SB}}\left[\frac{(\alpha+\gamma)}{2(\phi_{max}-\phi_{min})}\left(r_{SB}\phi_{max}^2+\frac{1}{r_{SB}}-2\phi_{min}\right)-\alpha\right], \tag{30}$$

$$t_e^{SB}=t_s^{SB}+\frac{N}{r_{SB}}, \tag{31}$$

where r_{SB} is the constant departure rate under the second-best toll. The rate is obtained by maximising the social surplus, given as follows:

$$SS_{SB}=\int_0^{N_{SB}}D(n)\,dn-TTC_{SB},$$

where the total costs in the second best are:

$$TTC_{SB}=\frac{(\beta+\gamma)\left(t^*-t_s^{SB}\right)^2}{2(\phi_{max}-\phi_{min})}\left(1-r_{SB}\phi_{min}+\ln\left(r_{SB}\phi_{max}\right)\right)-\frac{\gamma-\beta}{2}N\left(t^*-t_s^{SB}\right).$$

Note that r_{SB} cannot be obtained analytically, which will be illustrated in the numerical study.

Xiao et al.'s (2015) third-best toll adds an extra constraint to the second-best toll, namely, that the expected generalised price should be the same as in the untolled case. It thus adds another constraint; hence, we call it 'third best'. By definition, the first and the latest departure time follow Eqs. (26)–(29), and the third-best toll has the same length of peak period as the no-toll equilibrium. The constant departure rate r_{TB} can be given as follows:

(a) when there is departure after t*,

$$r_{TB} = \frac{\alpha+\gamma}{\alpha\phi_{max}+\gamma\phi_{min}} \tag{32}$$

(b) when there is no departure after t^*,

$$r_{TB} = \frac{1}{\hat{\phi}} \tag{33}$$

where $\hat{\phi}$ is obtained by solving $\frac{1}{\hat{\phi}}\left(\frac{\phi_{min}+\phi_{max}}{2} - \frac{(\hat{\phi})^2-(\phi_{min})^2}{(\phi_{max})^2-(\phi_{min})^2}\right) + \frac{\hat{\phi}-\phi_{min}}{\phi_{max}-\phi_{min}} = \frac{\alpha+\beta+\gamma}{\alpha+\gamma}$.

Interestingly, the constant departure rate under the third-best toll scheme when there is departure after t^* equals the optimal departure rate at t_e, that is, $r_{TB} = r_{FB}(t_e)$. And we have the below result.

Lemma 5:

The length of the peak period under the first-best toll is longer than that under the no-toll equilibrium or the third-best scheme when there are departures after t^*.

Proof of Lemma 5

From Eq. (26) and (27), it is evident that the length of the peak period under the no-toll equilibrium (and under the third-best scheme) is $t_e^{NT} - t_s^{NT} = \frac{N}{(\alpha+\gamma)/(\alpha\phi_{max}+\gamma\phi_{min})}$ when there is departure after t^*. From Proposition 3, the optimal departure rate is non-decreasing with t, and we have $r_{FB}(t_e) = (\alpha+\gamma)/(\alpha\phi_{max}+\gamma\phi_{min})$ at t_e. Thus, $r_{FB}(t) \le r_{FB}(t_e)$, when $t \le t_e$. Define r'_{FB} as the departure rate so that $r'_{FB}(t_e^{FB} - t_s^{FB}) = N$ under the first-best toll. This indicates $r'_{FB} < r_{FB}(t_e)$. The peak lengths compare as follows: $t_e^{FB} - t_s^{FB} = \frac{N}{r'_{FB}} > \frac{N}{r_{FB}(t_e)} = t_e^{NT} - t_s^{NT}$. □

As the third-best toll scheme is a constrained version of the second-best toll, it can, at most, do as well as the second-best scheme if the second-best would lead to the same price as no tolling. But, as we will see, this outcome only happens when there is no uncertainty in the capacity. The numerical model shows that the third-best toll always has lower prices than the second- and first-best cases, but it has a lower social surplus. Therefore, the third-best toll is preferable for users, making it more politically feasible to implement. All three time-variant tolls are non-linear and concave over time; hence, it can be difficult to implement all of them.

Beyond these results, we were unable to find further insightful analytical solutions. We therefore put much effort in developing numerical solution and optimization methods, and use them extensively in Section 7 to gain further insights. The said complexity also explains why, so far, little work has been done on tolling under uncertainty., As the numerical results also show, effects can vary strongly with parameters. Naturally, the same would be true for analytical solutions if these were obtained. Although it is impossible to exhaust all possible function forms, the numerical studies provided new insights beyond those from the analytics. The comparison with the no-toll benchmark is especially tedious, as various types of no-toll equilibria may exist depending on the parameter values.

5. Uniform toll

The uniform toll is constant during the morning peak, and it does not give an incentive to change the departure patterns other than indirectly via changing total demand. Thus, this scheme may affect the total number of travellers; but, for a given number of travellers, the equilibrium is the same as without tolling (Arnott, 1996; Long et al., 2022). However, these authors did not study uniform tolling. Yu et al. (2023) did look at step tolling.

The expected price follows from eq. (5) with $j = UT$. The social surplus is

$$SS_{UT} = \int_0^{N_{UT}} D(n)\, dn - N_{UT} E(C_{UT}). \quad (34)$$

Proposition 5:

The optimal uniform toll maximises objective (34) *and implies a toll that equals the marginal external cost (MEC) given unaltered equilibrium scheduling behaviour, which differs from the first-best toll and is now constant over time:*

$$\tau_{UT} = N_{UT} \frac{\partial E(C_{UT})}{\partial N_{UT}}. \quad (35)$$

The optimal demand N_{UT} is found by equating MSC and inverse demand: $P_{UT} = D(N_{UT})$. Users choose their departure rate in the same way as they do in the untolled equilibrium.

Proof of Proposition 5:

Appendix A.2 gives the derivations of this proposition. □

This tolling regime has many possible cases depending on the parameters, just as with the case without tolling. The uniform toll cannot alter departure rates. It can only lower the total number of users so that the (averaged-over-time) MSC equals the inverse demand. Conversely, without tolling, the inverse demand equals the travel cost, which is lower than the MSC. Hence, we can be certain that, compared to the untolled setting, the uniform toll raises the expected price, lowers the total number of users, and shortens the peak.

6. Single-step toll

The single-step toll consists of a 'time-variant' step-up component that is applicable during the step-tolling period in the centre of the peak and a time-invariant component that lasts the whole peak:

$$\tau_{ST}(t) = \rho_t + \mu. \quad (36)$$

Here, $\tau_{ST}(t)$ is the toll at departure time t, and ρ_t is the time-variant step part implemented from t^+ to t^-. The t^+ and t^- are the start time and end time of the step-tolling period. μ is the time-invariant part implemented throughout the peak.

Adapting Van den Berg's (2012) procedure, we again optimise in two steps. In the first step, given the number of users, the time-variant part ρ_t and the step-tolling period are obtained by minimising the total expected cost. In the second optimisation step, the total demand is set to maximise social welfare whilst considering the effect of the first step. This then implies the time-invariant part of the toll: μ.

Using fixed demand, Long et al. (2022) investigated the single-step toll in the stochastic bottleneck model, where the service time of the bottleneck follows a general distribution. If the step toll is implemented, commuters depart at a constant departure rate before t^+. When the toll is lifted at t^-, there will be a mass of commuters departing. Like Long et al. (2022), we use the ADL (Arnott et al., 1993) model with mass departures.

In the first optimisation step, the results are the same as those found by Long et al. (2022) for a uniform distribution. There are two cases under the single-step toll, depending on whether the peak ends at or after the preferred arrival time t^*. Here, under the single-step toll, the peak ends after t^* in Case I (Case 8.2 in Long et al. (2022)), and it ends at t^* in Case II (Case 9.2 in Long et al. (2022)). For the second optimisation step, Appendix A.3 gives detailed derivations, and the results are summarised below.

Following Long et al. (2022), the expected total social cost in Case I is:

$$TC_{ST}(N_{ST}) = -\frac{W(\overleftrightarrow{\phi})\left(M(\overleftrightarrow{\phi})\right)^2 (N_{ST})^2}{4(\beta+\gamma)^2} + \frac{(\alpha+\gamma)(\bar{\phi}-\tilde{\phi})\beta(N_{ST})^2}{\beta+\gamma}. \tag{37}$$

Here, $\overleftrightarrow{\phi}$ is the reciprocal of the average departure rate during the period from the departure time of the first commuter to the departure time of the first commuter who pays the toll. Further, $\overleftrightarrow{\phi}$ is obtained by solving the following equation:

$$M(\overleftrightarrow{\phi})\frac{dW(\overleftrightarrow{\phi})}{d\overleftrightarrow{\phi}} + 2W(\overleftrightarrow{\phi})\frac{dM(\overleftrightarrow{\phi})}{d\overleftrightarrow{\phi}} = 0,$$

where $W(\overleftrightarrow{\phi}) = \frac{1}{Y(\overleftrightarrow{\phi})+\frac{2}{(\alpha+\gamma)\bar{\phi}}}$, $Y(\overleftrightarrow{\phi}) = \frac{1}{\alpha\overleftrightarrow{\phi}-(\alpha-\beta)H(\overleftrightarrow{\phi})}$, $H(\overleftrightarrow{\phi}) = \bar{\phi} - G(\overleftrightarrow{\phi}) + \overleftrightarrow{\phi}F(\overleftrightarrow{\phi})$, $F(x) = \frac{x-\phi_{min}}{\phi_{max}-\phi_{min}}$, $M(\overleftrightarrow{\phi}) = \beta(\alpha+\gamma)(\bar{\phi}-\tilde{\phi})Y(\overleftrightarrow{\phi}) - \beta(\beta+\gamma)H(\overleftrightarrow{\phi})Y(\overleftrightarrow{\phi}) + (\beta+\gamma) + \frac{2\beta(\bar{\phi}-\tilde{\phi})}{\bar{\phi}}$, $G(x) = \frac{x^2-\phi_{min}^2}{2(\phi_{max}-\phi_{min})}$, $\tilde{\phi} = G\left(\frac{\alpha\phi_{max}+\gamma\phi_{min}}{\alpha+\gamma}\right)$, and $\bar{\phi} = \frac{\phi_{max}+\phi_{min}}{2}$. These definitions follow from those presented by Long et al. (2022). As proved in their Proposition 16, $\overleftrightarrow{\phi}$ is independent of N_{ST}.

Consequently, the total expected social cost in Eq. (37) is a function of travel demand. The total toll revenue is

$$TR_{ST} = \rho_t\, N_1 + \mu\, N_{ST}, \tag{38}$$

where $\rho_t = \frac{M(\overleftrightarrow{\phi})\, W(\overleftrightarrow{\phi}) N_{ST}}{2(\beta+\gamma)}$ is the optimal time-variant step part of the toll and $N_1 = N_{ST} - \frac{\rho_t}{W(\overleftrightarrow{\phi})}$ is the number of travellers departing during the step-tolling period. The step part of the toll, ρ_t, is a function of travel demand. The average marginal external cost (*MEC*) is the difference between the average marginal social cost and the average travel cost. From Eq. (37), the average *MEC* is

$$\overline{MEC}_{ST} = -\frac{W(\overleftrightarrow{\phi})\left(M(\overleftrightarrow{\phi})\right)^2 N_{ST}}{4(\beta+\gamma)^2} + \frac{(\alpha+\gamma)(\bar{\phi}-\tilde{\phi})\beta N_{ST}}{\beta+\gamma}. \tag{39}$$

We obtain the optimal time-invariant part of the toll by equalising the average MEC and the average toll, that is $\overline{MEC}_{ST} = \frac{TR_{ST}}{N_{ST}}$. Then, the optimal time-invariant part is

$$\mu = \overline{MEC}_{ST} - \frac{\rho_t N_1}{N_{ST}} = \frac{(\alpha+\gamma)(\bar{\phi}-\tilde{\phi})\beta N_{ST}}{\beta+\gamma} - \frac{W(\overleftrightarrow{\phi})M(\overleftrightarrow{\phi})N_{ST}}{2(\beta+\gamma)}. \tag{40}$$

The expected price in Case I is

$$P_{ST} = -\frac{W(\overleftrightarrow{\phi})\left(M(\overleftrightarrow{\phi})\right)^2 N_{ST}}{2(\beta+\gamma)^2} + \frac{2(\alpha+\gamma)(\bar{\phi}-\tilde{\phi})\beta N_{ST}}{\beta+\gamma}. \tag{41}$$

From Eq. (41), the optimal demand under a step toll can be found using $P_{ST} = D(N_{ST})$. The results for Case II will be similarly obtained and are given in Appendix A.4 to save space.

7. Numerical study

Because the analytical results are not easy to interpret, we conduct a numerical study to compare the different schemes, providing new insights beyond those from the analytics. In particular, we focus on how the various tolls compare in their effects on costs, consumers and overall welfare. Further, we perform extensive sensitivity tests to evaluate how robust these insights are. We use a uniformly distributed bottleneck service time as it is a common assumption, as discussed in the literature review.

The first subsection will look at fixed demand, while the second examines price-sensitive demand. Fixed demand allows for a clearer comparison of departure rates and costs: otherwise, the number of users will differ between regimes, which complicates comparisons.

Unless otherwise stated, we use the same values for unit cost parameters as those used by Long et al. (2022): $\alpha = 6.4\ \$/h$ and ratios $\beta/\alpha = 0.609$ and $\gamma/\beta = 3.9$. The desired arrival time $t^* = 9\ h$. These values and especially the ratios are highly common in the literature. Following Long et al. (2022), the mean bottleneck service time is $\overline{\phi} = 1s/veh$, which implies an 'average' capacity of 3600 veh/h. Let e be the spread of the service time range and $e = \phi_{max} - \phi_{min}$. Increasing the spread e under the fixed mean bottleneck service time (i.e. $\overline{\phi}$) makes the system more uncertain. When e approaches zero, the model approaches the deterministic scenario.

The inverse demand function is assumed to be linear: $D(n) = d_0 - d_1 n$. The average demand elasticity is -0.4 (Van den Berg, 2012). When the capacity is at the mean value (i.e. the capacity is $1/\overline{\phi}$), the number of commuters *N* is 5000 *veh.* under the no-toll scheme (Long et al., 2022). Calibrating this value implies $d_0 = 15.0893$ and $d_1 = 0.0022$ under this setting. Due to the uncertainty, the demand and price will change under the no-toll equilibrium when the elasticity changes.

We numerically solve the 'first-best' toll based on analytical results and by discretising the departure time into time windows. Discretisation is a common way to numerically solve differential equations. It does not rely on Monte Carlo simulation, random draws, or sample paths. Our approach for the first-best case is akin to that of Yang and Huang (1997; 1998). To solve for the departure rates under the first-best toll, we start from the time window at t_e and then go backward to the start time window at t_s. To achieve this solution, we use the results for situations I–IV from Appendix A. The rate is constant piecewise, and this approximation becomes ever more precise the more time windows there are.

7.1. Numerical evaluation of different scenarios under fixed demand

Long et al. (2022) proposed a 'second-best' time-varying toll that maximises welfare whilst having a constant departure rate and a toll starting and ending at zero. Xiao et al. (2015) introduced a 'third-best' time-varying toll that incorporates these aims and adds a further constraint that the expected generalised price (i.e. the sum of expected travel cost and toll) remains at its untolled level. In these previous works, the demand is fixed. Moreover, the time-varying toll schemes proposed in previous studies cannot achieve the system optimum. The 'first-best' toll of Lindsey (1999) achieves this optimum and leads to a departure rate that weakly increases over time.

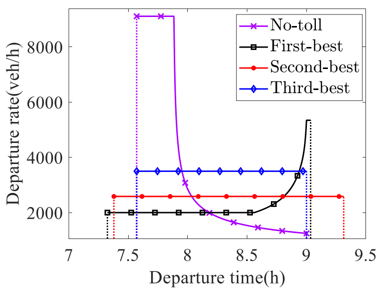
For a fair comparison with the literature, this section uses fixed demand. The next section adds price-sensitive demand to this. Fixed demand allows for a clearer comparison of departure rates and costs.

We denote the spread in the service time by e, and a larger e means more uncertainty in the capacity. For different spreads, Fig. 1 plots the equilibrium departure rates under the no-toll and the first-, second- and third-best time-varying toll schemes.

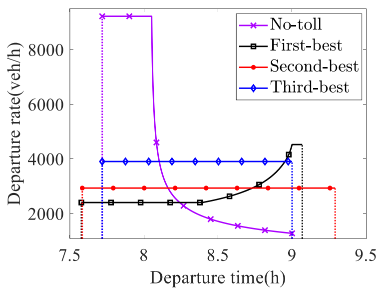
The departure rate in no-toll equilibrium is weakly **de**creasing over the departure time, while that under the first-best toll is weakly **in**creasing. So, the first-best rate is qualitatively the mirror image of the untolled outcome. The intuition is that as time progresses, more travellers will have departed and fewer travellers will be delayed by queuing, meaning that the cost of queuing decreases over time. Thus, in the earliest departure period, the departure rate is constant and equals the minimum capacity to avoid high queuing costs; subsequently, the rate starts to increase, before becoming constant again for the last departure period. During the early peak, capacity is underutilised unless the worst possible traffic conditions arise. These results are consistent with those of Lindsey (1994), who used a binary distribution, and Fosgerau and Lindsey (2013), who used a capacity that varies over the day. In line with our analytics, Fig. 1 shows that the optimal departure rate is constant after t^*, and is in between the 'average' capacity and the maximum capacity. Interestingly, when there is minimal uncertainty, the first-best departure rate after t^* is the same as that under the third-best toll.

The third-best time-varying toll keeps the price unchanged from the no-toll scheme, and the peak duration and timing are also the same. With our uniform distribution, the peak begins earlier and ends later in the social optimum than for the untolled case. Under the second-best time-varying toll, the peak is the longest due to a lower departure rate on average. Conversely, the third-best toll has the shortest departure window. Thus, compared with the no-toll case, both the first- and second-best tolls increase the price and peak duration under fixed demand, which differs from in the deterministic model.

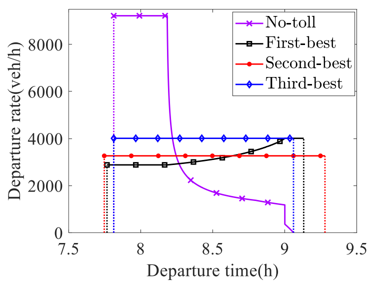
When spread e decreases, so that there is less uncertainty, the departure rate under the first-best toll becomes more concentrated, and the rates of all three time-varying schemes approach the mean capacity, as illustrated in Fig. 1(d). A constant departure rate is only optimal when there is no uncertainty: in the limit, as the uncertainty disappears, all three regimes become the optimal toll of the deterministic model.



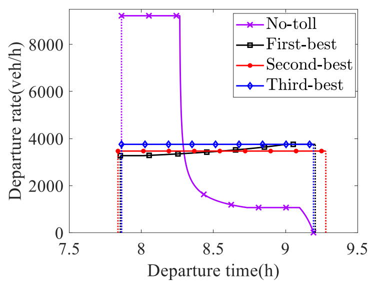
(a) High uncertainty with a large spread of e =1.6 *s/veh*.



(b) Medium uncertainty with a medium spread of e=1 *s/veh*.



(c) Low uncertainty with a small spread of e=0.5 *s/veh*.



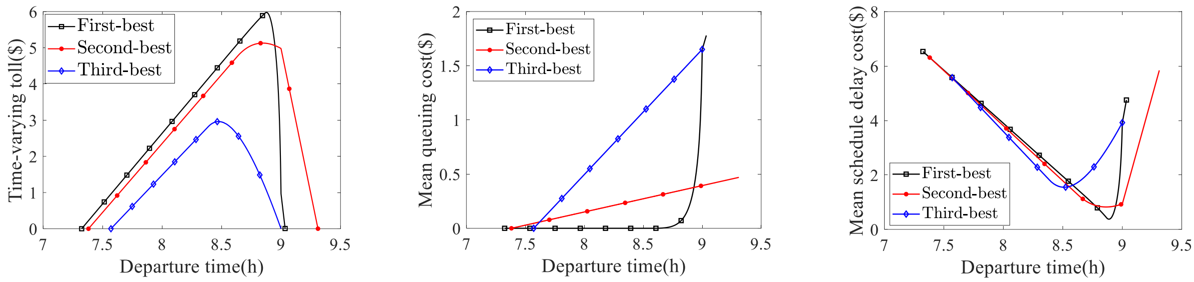
(d) Very low uncertainty with a spread of e=0.2 *s/veh*.

Fig. 1: Comparisons of equilibrium departure rates for four spreads of capacity.

Note: The service time varies from ϕ_{min} to $\phi_{max} = 2 - \phi_{min}$. Therefore, lowering the spread e decreases uncertainty without altering the mean.

Fig. 2 shows the toll, expected queuing cost, and expected schedule delay cost over departure time. All tolls start at zero. The third-best toll is much lower than the others and its peak is much shorter, which results in much higher travel times. In the graph, the first-best toll is higher than the second-best. But this result is not universal and depends on parameter levels: with low uncertainty, we obtain the opposite result (see Fig. C.1 in Appendix C). The first-best toll starts the earliest, regardless of the parameters.

Queuing is not eliminated under uncertainty. It occurs during the earliest departure period in the optimum, which is consistent with the analytics. However, subsequently, queuing is always possible. For the second- and third-best toll schemes, because of the constant departure rate, the expected queuing increases linearly over time. In the social optimum, expected travel time delays are low for much of the peak, but they sharply increase during the latter peak due to the sudden increase in the departure rate. This increase is more pronounced the more uncertain the capacity is. The mean schedule delay cost at the end of the peak is smaller than that at the beginning, as tolling will not eliminate all delays and the schedule delays need to be lower to compensate.

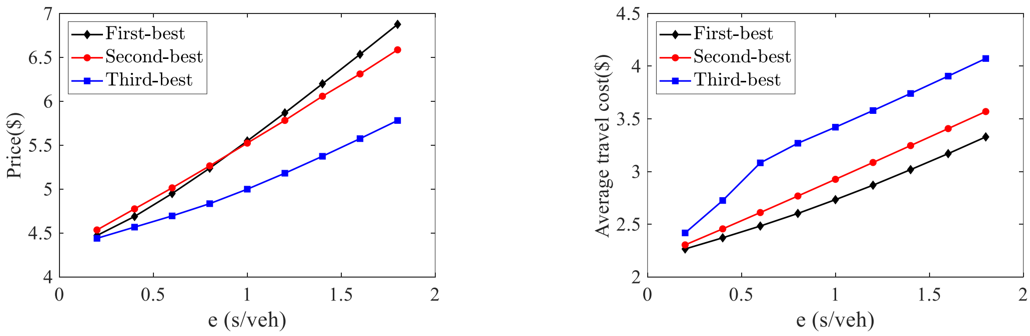


(a) Toll (b) Mean queuing cost (c) Mean schedule delay cost

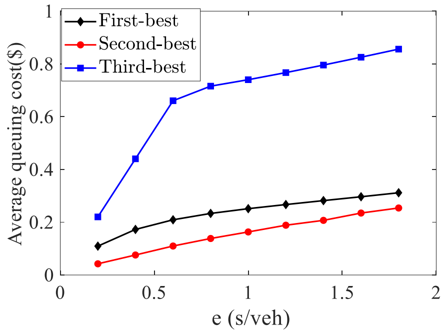
Fig. 2: Comparisons of toll, mean queuing cost and mean schedule delay when the spread is e = 1.6 *s/veh*.

Fig. 3 illustrates the effect of tolling over different degrees of uncertainty. When the spread e in the service time increases, there is more uncertainty, and the price and expected levels of travel cost, queuing cost and schedule delay cost increase. Unlike in the deterministic model, the first-best toll raises the expected price from the untolled setting, and it decreases travel cost less in percentage terms. These effects are stronger the more uncertainty there is.

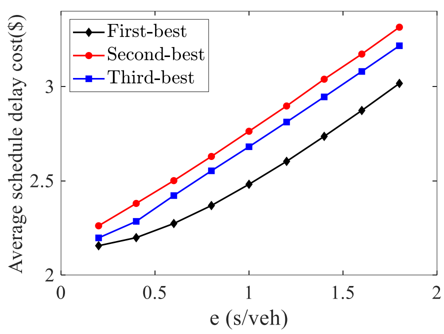
As mentioned above, the third-best toll keeps the expected price unchanged from the no-toll case. In Fig. 3(a), the first- and second-best tolls raise the price. When there is more uncertainty, the first-best price is higher than the second-best price. The expected average travel cost under the optimal toll is by design at the minimum, while this is not necessarily true for queuing costs. The expected queuing cost under second-best tolling is lower than that under first-best tolling, as illustrated in Fig. 3(c). Fig. 3(d) shows that the average schedule delay is the lowest under the social optimum, and it is the highest under the second-best toll.



(a) Expected price (b) Average travel cost



(c) Average queuing cost



(d) Average schedule delay cost

Fig. 3: The effect of the spread, e, that measures the degree of uncertainty on (a) expected price, (b) average travel cost, (c) average queuing cost, and (d) average schedule delay cost.

Along with the analysis above, we give some possible intuitive explanations concerning the results under the first- and second-best tolls. As shown in Fig. 1, the departure duration is the longest and the constant departure rate is the lowest under the second-best toll. The second-best toll spreads the departures greatly, so that queuing is lower while average schedule delays are larger. Compared to the first best, the second best is more effective at eliminating queues and less so at reducing schedule delays.

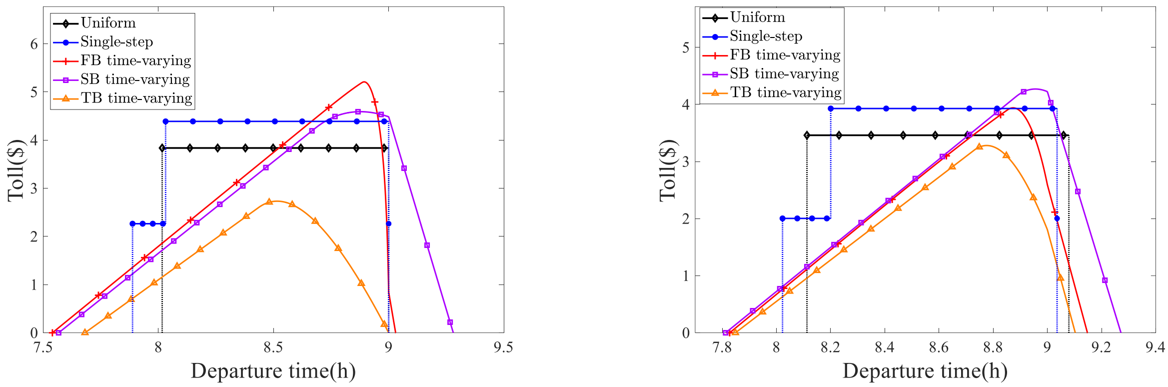
To summarise, the first-best departure rate weakly increases over time, which is the mirror image of the untolled rate that weakly decreases. By design, the second- and third-best cases have constant rates. The second-best may be easier to design than the first-best, but this comes at the downsides of a higher price and travel cost. The third-best tolling keeps the expected price at the no-toll level, leading to much higher costs and thus lower welfare. It is debatable if the higher social acceptability of the third-best tolling is worth these downsides.

7.2. Numerical evaluation of different scenarios under price-sensitive demand

Now, we extend our analysis to price-sensitive demand. As argued, keeping demand fixed always facilitates the comparison of departure rates and costs (otherwise, the tolls indirectly affect the rates and costs by changing demand). However, price-sensitive demand has important effects, in particular, due to its interaction with uncertainty. Consequently, we extend the analysis of the second- and third-best tolls by optimising demand as a separate step. The previous works on them used fixed demand.

7.2.1. Structures of different tolling schemes

For two levels of uncertainty, Fig. 4 plots the uniform toll, single-step toll, first-best (FB) time-varying toll, second-best (SB) toll and third-best (TB) toll with price-sensitive demand. The flat and step tolls are generally higher than the average toll of the time-varying schemes. As with the deterministic model, the marginal external cost is higher with coarser pricing, implying higher toll levels. For the step toll, the time-variant step part of the toll is lifted at some point; and at this moment, there is a mass departure of users. Finally, as with fixed demand, for high uncertainty, the first-best time-varying toll begins early and the average toll is higher than the corresponding level of the second-best toll; the reverse pattern of results hold with low uncertainty.



(a) High uncertainty due to a large spread: e = 1.6 *s/veh*. (b) Low uncertainty due to a small spread: e= 0.4 *s/veh*.

Fig. 4: Different tolling schemes for two levels of uncertainty as measured by spread, e.

Note: FB means first-best time-varying toll, SB means second-best time-varying toll, and TB means third-best time-varying toll.

7.2.2. The effect of uncertainty as measured by the spread, e, of the service time

To compare the efficiency of different tolling schemes with price-sensitive demand, we employ the index ω_i^{SS} that denotes the relative efficiency of scheme i, that is, its social surplus gain from the no-toll case relative to that of the first-best scheme:

$$\omega_i^{SS} = 100 \cdot \frac{SS_i - SS_{NT}}{SS_{FB} - SS_{NT}},$$

where SS_{NT}, SS_{FB} and SS_i denote the social surplus of the system under the no-toll case, first-best case and scheme i, respectively. The first-best toll has, by definition, a relative efficiency of 100, and the base-case untolled equilibrium of 0. All our other policies are in between these values, and the number reveals the relative performance of the policy.[8]

Fig. 5 shows how uncertainty affects demand, price, expected average travel cost, average toll, social surplus, and relative efficiency under different tolling schemes. There are kinks in the curves for the third-best toll around a spread of e = 0.6 when the equilibrium pattern changes. Specifically, the peak ends at (after) the desired arrival time when the spread is smaller (larger) than the value of e.

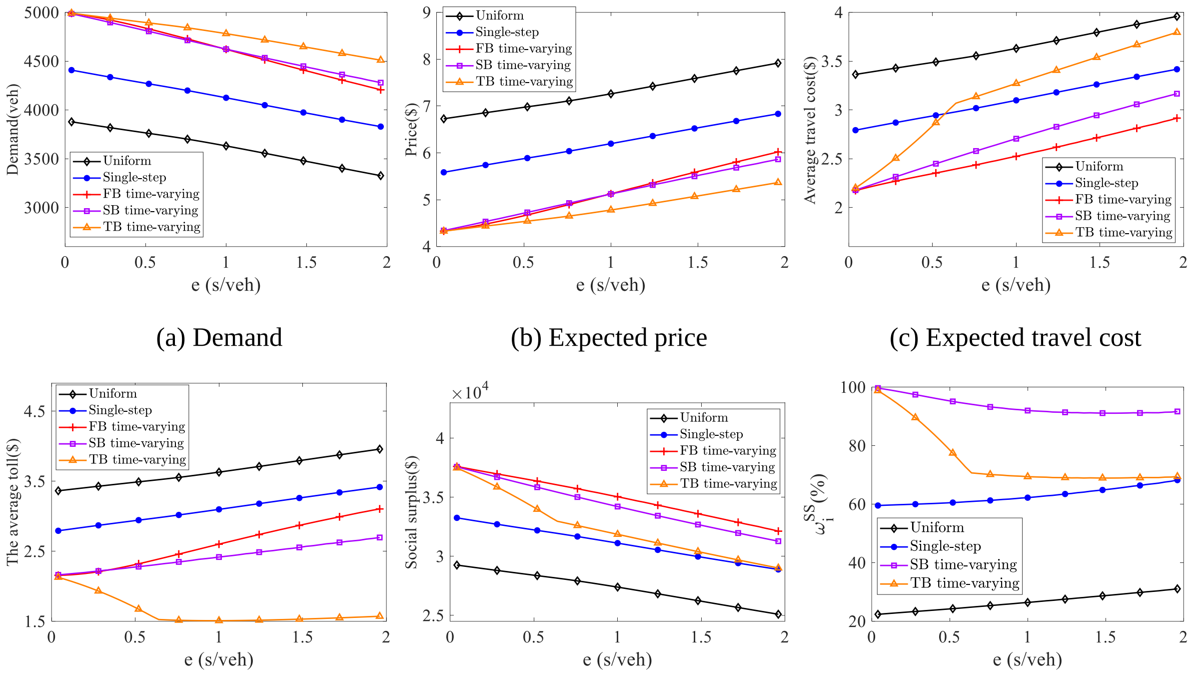
We make the following observations. First, travel demand decreases, and the expected price increases with e. The demand is highest and the price is lowest under the third-best toll, where they are the same as without tolling. With a larger spread and thus more uncertainty, the first- and second-best schemes raise the price more from the no-toll case, thus reducing demand more. There is only a moderate difference between the first- and second-best schemes, making the case that the relative ease of implementing the second-best tolling compared to the first-best tolling may be worth it. As with fixed demand, when spread e exceeds approximately 1, the price is higher and demand is lower under the first-best toll than under the second-best toll; whereas the reverse holds when the spread is below 1.

Second, from Figs. 5(c)–(d), it can be seen that the average travel cost and toll increase with e. When there is low uncertainty, the average travel cost and toll under the third- and second-best schemes approach those under the first-best toll. For low uncertainty, the third-best toll has similar costs as the first- and second-best, but costs quickly increase with the degree of uncertainty, and even for moderate uncertainty, third-best tolling has much higher costs. The single-step and uniform tolls have much higher costs than the first best, but they perform relatively better when there is more uncertainty (i.e. the higher e is). For moderate uncertainty, the step toll has lower costs than the third-best toll; and for high

[8] It is possible for a policy to have an efficiency below 0 if it has a welfare below the base case. But none of our regimes do.

uncertainty, the third-best toll even performs similarly to the uniform toll. Fig. 5(d) shows that the average toll under the third-best scheme is always the lowest.

Finally, Figs. 5(e–f) present welfare effects. More uncertainty (i.e. a higher e) reduces the relative efficiency of the second-best toll, but it always performs similarly to the first-best scheme. The third-best scheme has a lower welfare, and the difference is large for moderate and especially high uncertainty. Still, the third-best scheme always does better than the step toll, even though the average cost might be higher under the former. However, for a very high spread e and thus high uncertainty, the difference between the two is minimal. The three time-varying schemes all approach the toll in the deterministic model as the uncertainty becomes zero. The uniform and step tolls have lower welfare levels than the first-best toll, but higher uncertainty makes them perform relatively better. So, the relative loss from a simpler tolling scheme is smaller when we include the realism of uncertainty. This reflects two things: 1) Lowering demand is more important in a highly uncertain environment, whereas it becomes increasingly difficult to reduce average travel times. 2) The uniform and flat toll are better at changing demand levels than lowering delays.



(a) Demand (b) Expected price (c) Expected travel cost

(d) Average toll (e) Social surplus (f) Relative efficiency

Fig. 5: Effect of spread, e, and, thus, uncertainty on outcomes under price-sensitive demand.

Note: The relative efficiency of policy i is $\omega_i^{SS} = \frac{SS_i - SS_{NT}}{SS_{FB} - SS_{NT}}$, where SS_i is the social surplus or welfare.

Because tolling increases the expected price and has a lower welfare gain under uncertainty, it may be difficult to politically sell tolling. However, we also see that the third-best toll that keeps this price constant has a much lower welfare and higher travel costs. Thus, the latter option may not improve things. Uniform and step tolling are more realistic than time-variant tolls, and they perform relatively better when there is more uncertainty.

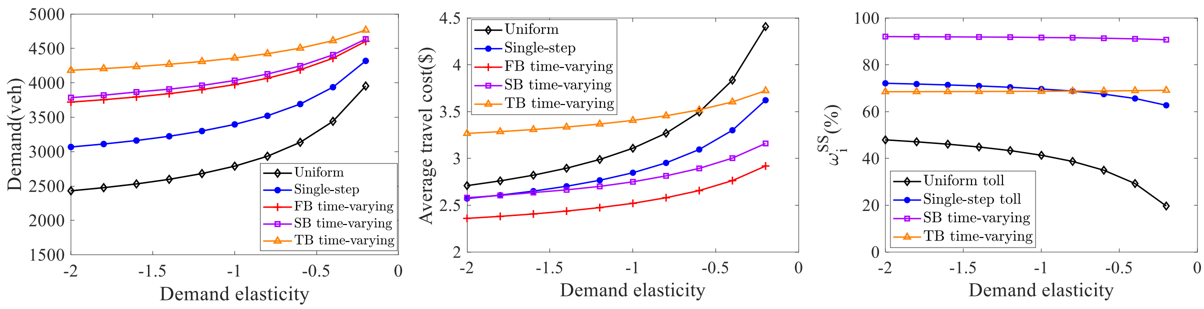
7.2.3. Effect of demand elasticity

Fig. 6 illustrates how demand elasticity affects our schemes' demand, average travel cost and relative efficiency when the spread e is 1.6. When the elasticity goes from -0.2 to -2, demand becomes more price sensitive. Compared to time-varying schemes, the performance of step-toll schemes is more sensitive to elasticity. The demand under the third-best toll is, by definition, the same as without tolling.

According to Fig. 6, more price-sensitive demand means lower demand and average travel costs. Furthermore, there is little difference in demand between the first and second-best schemes. Demand under the third-best toll is always the largest. This might also result in the highest travel costs, especially when demand is more price sensitive.

Finally, Fig. 6(c) illustrates that as demand becomes more price-sensitive, the relative efficiency of the step toll and uniform toll increases, while that of time-varying schemes remains similar. The efficiency of the single-step toll is higher than that of the third-best toll when demand is really price sensitive (i.e. the elasticity is below -0.8). Note again that the first-best scheme has, by definition, a relative efficiency of 100, and the base-case untolled equilibrium of 0.

In contrast, when there is less uncertainty, the uniform and single-step tolls become less efficient, and their performance becomes inferior to that of time-varying schemes (see Fig. C.2 in Appendix C). This outcome arises because these coarse schemes are better at changing the total demand and worse at changing departure times. When demand becomes more price-sensitive, the step and flat toll become more efficient since it is more important to change demand then. With less uncertainty, the time-varying tolls perform relatively better because then the shift in the departure rates matters more.



(a) Demand (b) Average travel cost (c) Relative efficiency

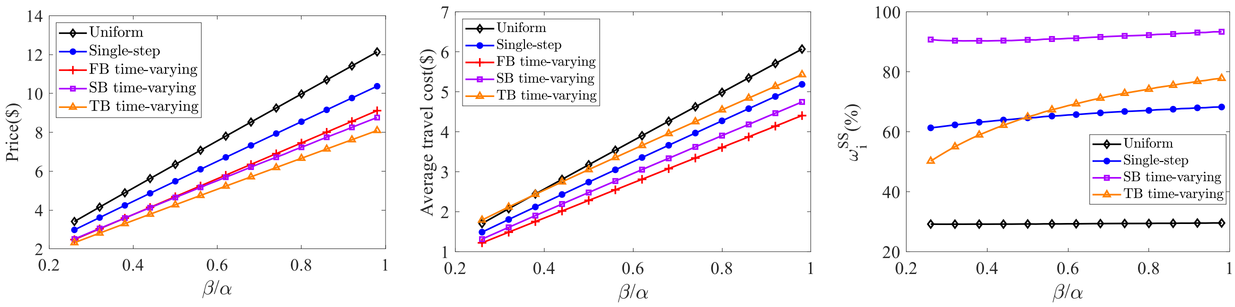
Fig. 6: Effect of elasticity on outcomes when the spread is $e = 1.6$ *s/veh*.

Note: The relative efficiency of policy *i* gives its welfare gain relative to first-best tolling and is $\omega_i^{SS} = 100\frac{SS_i - SS_{NT}}{SS_{FB} - SS_{NT}}$, where SS_i is the social surplus or welfare.

7.2.4. The effect of β/α

How the ratio β/α affects the price, average travel cost and relative efficiency is illustrated in Fig. 7. Here, the value of time $\alpha = 6.4$\$/h and the ratio of γ/β remain unchanged. The ratio β/α varies from 0.26 to 0.99 to ensure $\beta < \alpha$. With an increase of β/α, queuing becomes less costly while experiencing schedule delays becomes more costly. Further, with an increase of β/α, the price and average travel cost increase due to the increased value of schedule delay. Regardless of β/α, the performance of the second-best toll is close to that of the first-best. The price under the third-best toll is again always the lowest.

Furthermore, with a lower value of schedule delay, the expected travel cost might be higher under the third-best scheme than that under the uniform toll (see Fig. 7(b)). Under these circumstances, queuing is relatively costly, indicating that the third-best scheme is poor at reducing queue delays.



(a) Expected price (b) Expected travel cost (c) Relative efficiency

Fig. 7: Effect of β/α on outcomes when $e = 1.6$ *s/veh*.

Note: The relative efficiency of policy *i* is $\omega_i^{SS} = \frac{SS_i - SS_{NT}}{SS_{FB} - SS_{NT}}$, where SS_i is its social surplus or welfare.

7.3. Summarising the numerical analysis

This section compares the socially optimal time-varying toll to second-best and third-best tolls that, by assumption, have a departure rate that is constant over time and start and end at zero. The third-best toll adds a further constraint, namely that the expected price should be the same as without tolling. We also examined a uniform toll—which is constant over time—and a single-step toll. To analyse all this, we extended existing models to include price-sensitive demand by adding a second step in the optimisation that optimises the total number of users. Previous works had fixed demand, and therefore could not look at the uniform toll, as that toll can only have an effect under price-sensitive demand.

In the social optimum, the departure rate weakly increases over the morning. Further, in the early peak, the rate is constant over time and equals the minimum capacity. Subsequently, the rate increases over time, and, finally, in the latest part of the peak, it becomes constant again. This pattern differs greatly from that of the deterministic model, where the optimal departure rate is always constant. Conversely, the untolled outcome has a rate that is the mirror image of the optimal rate and weakly decreases over time. Unlike in the deterministic model, the first-best toll raises the expected price from the untolled setting, and it decreases travel costs and increases welfare less in percentage terms. These changes are stronger the more uncertainty there is.

The second-best toll attains a welfare level that is somewhat lower than the first-best one, whereas the third-best toll attains a much lower welfare. The relative efficiencies of the second-best and third-best schemes fall with the degree of uncertainty as measured by the spread of the service time, whereas the price sensitivity has little to no effect. The uniform toll has a welfare that is much lower than that of the single-step toll. For most parameter ranges, the step toll, in turn, has a lower welfare level than that of the third-best toll; exceptions include when demand is highly price-sensitive or when uncertainty is high.

As demand becomes increasingly price-sensitive, all schemes result in expected prices that are close to each other. The uniform toll has the highest price, followed by the single-step toll, then either the first-best or second-best toll, and, finally, the third-best toll that has the same expected price as the no-toll scheme. Whether the first- or second-best toll has a higher price depends on multiple parameters, but their prices will be close to each other. The uniform and step tolls attain welfare levels close to those of the first-best toll when the price sensitivity or uncertainty increases.

8. Conclusion

This study examines various tolling schemes in the stochastic bottleneck model with price-sensitive demand. We assume that the capacity is constant within a day but changes stochastically from day to day. We study three time-varying toll schemes, a uniform toll and a single-step toll. Our core contribution is doing so under uncertain capacity and price-sensitive demand. Previous works only looked at these aspects separately; however, we find that their interplay changes the results substantially. This approach is important for policymaking since the combined occurrence of them is likely in reality: travel time uncertainty is a fact of life in real transport systems. Solving the required models is, however, complex and does not result in a transparent analytical closed-form solution for the performance of alternative toll schedules, which is why we complemented our analytical work with extensive numerical modelling.

In the stochastic bottleneck model, the first-best social optimum is decentralised by a time-varying toll and has a continuous departure rate that weakly increases over time. This outcome differs greatly from the deterministic model, where the optimal rate is constant over time. Our core methodological analytical contribution is solving for time-varying tolls and the step toll in two steps. In the first step, under a given demand, the departure rate is optimised, which implies the toll development over time. In the second step, the total demand is optimised, which implies the toll's starting level.

Compared with the first-best toll, the second-best time-varying scheme is simpler to derive and implement because it imposes that the departure rate is constant over time and the toll starts and ends at zero. This outcome is optimal in the deterministic bottleneck model, but it is second-best in our case. This imposition leads to a lower welfare than under first-best tolling, but our numerical study reveals that the difference is not large. The third-best time-varying scheme adds a further constraint that the expected generalised price should be the same as in the no-toll case. This additional constraint is optimal in the deterministic model, but it substantially lowers welfare and raises costs under uncertainty. When tolling increases the expected price and offers low welfare benefits under uncertain conditions, it may become difficult to sell politically. However, the third-best toll performs much worse than the first-best toll; thus, this option may not be the most desirable.

The step toll has a welfare below that of the third-best toll for most parameter ranges. The uniform toll has a welfare that is well below that of the step toll. With more uncertainty or more price-sensitive demand, the uniform and step toll perform relatively better: they produce welfares and travel costs closer to those of the first best. This outcome increases the attractiveness of these realistic tolls and shows the importance of jointly considering uncertainty and price-sensitive demand.

Our study can be extended in several directions. The first extension could be a consideration of other step-toll models because we used the ADL step-toll. How would adding more steps to the step toll alter outcomes? The second- and third-best tolls have the advantage of being simpler to design due to the constant departure rate. However, they still have tolls that vary non-linearly over time; conversely, in the deterministic model, the toll has a linear slope. How would such a simple and easy-to-implement linear toll perform?[9] What toll would a profit-maximising road operator set? In the deterministic model, the operator uses the same toll pattern as the first-best toll but with a time-invariant markup added. Does the same hold with stochasticity? What happens if we consider large networks or different forms of congestion? What if demand is also uncertain or the uncertainty varies over the day (e.g. due to an accident that is cleared)? What about alternative policies to tolling, such as capacity expansion, working from home, flexible working hours, or travel credits? Finally, accounting for the effect of information is

[9] Chu (1999) analyses this for his dynamic flow congestion model without uncertainty.

important (Yu et al., 2021, 2023; Han et al., 2021; Verhoef et al., 1996). There are several interesting and important opportunities for further research on stochastic and congested transportation systems.

Finally, as a policy conclusion, we find that in the stochastic bottleneck model, unlike in the deterministic one, the first-best socially optimal toll cannot remove all queuing and hurts consumers by raising the expected price from that of the untolled case. Moreover, the percentage welfare gain is lower in our stochastic model than it is in the deterministic bottleneck model, which is likely to make tolling harder to implement politically. Adding a constraint requiring the expected price to remain the same as without tolling may seem beneficial to users (before considering revenue-recycling); however, this approach substantially raises travel costs and lowers toll revenue and welfare. This highlights that the interaction between uncertainty and price sensitivity complicates the design of transportation policies and alters their effects. Moreover, in reality, there is uncertainty, making its consideration important.

Acknowledgement

Financial support from the National Natural Science Foundation of China (Grant Nos. 72288101, 71931002) and China Scholarship Council (202007090099) are gratefully acknowledged.

Conflict of interest

None of the authors have financial or personal relationships with other people or organizations that could have inappropriately influenced or biases their work. No AI has been used by the authors.

References

Agnew, C.E., 1977. The theory of congestion tolls. Journal of Regional Science, 17(3), 381-393.

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. 1991. Does providing information to drivers reduce traffic congestion? Transportation Research Part A, 25(5), 309-318.

Arnott, R., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: a traffic bottleneck with elastic demand. Am. Econ. Rev., 83 (1), 161-179.

Arnott, R., de Palma, A., Lindsey, R., 1996. Information and usage of free-access congestible facilities with stochastic capacity and demand. Inter. Econ. Rev., 37, 181-203.

Arnott, R., de Palma, A., Lindsey, R., 1999. Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand. Eur. Econ. Rev., 43, 525-548.

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

de Palma, A. Fosgerau, M., 2013. Random queues and risk averse users. European Journal of Operational Research, 230(2), 313-320.

Chu, X., 1999. Alternative congestion technologies. Regional Science and Urban Economics, 29 (6), 697–722.

Fosgerau, M., 2010. On the relation between the mean and variance of delay in dynamic queues with random capacity and demand. J. Econ. Dyn. Control 34(4), 598–603

Fosgerau, M., Lindsey R., 2013. Trip-timing decisions with traffic incidents. Regional Science and Urban Economics, 43, 764-782.

Fu, X., van den Berg, V.A.C., Verhoef, E. T., 2018. Private road networks with uncertain demand. Research in Transportation Economics, 70, 57-68.

Hall, J.D. 2018. Pareto improvements from Lexus Lanes: The effects of pricing a portion of the lanes on

congested highways. Journal of Public Economics, 158, 113-125.
Hall, J.D., Savage, I., 2019. Tolling roads to improve reliability. Journal of Urban Economics 113, 103187.
Han, X., Yu, Y., Gao, Z.Y., Zhang, H.M., 2021. The value of pre-trip information on departure time and route choice in the morning commute under stochastic traffic conditions. Transportation Research Part B: Methodological, 152, 205-226.
Henderson, J.V., 1974. Road Congestion. A Reconsideration of Pricing Theory. Journal of Urban Economics 1, 346–365.
Jiang, G.G., Lo, H.K., Tang, Q.R., Liang Z., Wang, S.L., 2021. The impact of road pricing on travel time variability under elastic demand, Transportmetrica B: Transport Dynamics, 9(1), 595-621.
Jiang, G.G., Wang, S.L., Lo, H.K., Liang, Z., 2022. Modeling cost variability in a bottleneck model with degradable capacity. Transportmetrica B: Transport Dynamics, 10(1), 84-110.
Klein, I., Levy, N., Ben-Elia, E., 2018. An agent-based model of the emergence of cooperation and a fair and stable system optimum using ATIS on a simple road network. Transportation research part C: emerging technologies, 86, 183-201.
Laih, C.H., 1994. Queuing at a bottleneck with single and multi-step tolls. Transportation Research Part A, 28 (3), 197-208.
Lam, T.C., 2000. The effect of variability of travel time on route and time-of-day choice. PhD dissertation, Department of Economics, University of California at Irvine.
Lawphongpanich, S., Yin, Y., 2010. Solving the Pareto-improving toll problem via manifold suboptimization. Transportation Research Part C: Emerging Technologies, 18(2), 234-246.
Li, H., Bovy, P.H. and Bliemer, M.C., 2008. Departure time distribution in the stochastic bottleneck model. International Journal of ITS Research, 6(2) 79–86.
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: Methodological, 139, 311-342.
Li, M., Jiang, G., Lo, H.K., 2022. Pricing strategy of ride-sourcing services under travel time variability. Transportation Research Part E: Logistics and Transportation Review, 159, 102631.
Liang, Z., Jiang, G., Lo, H. K., 2023. Dynamic equilibrium analyses in a ride-sourcing market under travel time uncertainty. Transportation Research Part C: Emerging Technologies, 153, 104222.
Lindsey, R., 1994. Optimal departure scheduling in the morning rush hour when capacity is uncertain. Presented at the 41st North Amer. Meeting of the Regional Sci. Assoc., Niagara Falls, Ontario.
Lindsey, R., 1996. Optimal departure scheduling for the morning rush hour when capacity is uncertain. Volume 2: Modelling Transport Systems. Proceedings of the 7th World Conference on Transport Research World Conference on Transport Research Society.
Lindsey, R., 1999. Optimal departure scheduling for the morning rush hour when capacity is uncertain. In: Emmerink, R., Nijkamp, P. (Eds.), Behavioural and Network Impacts of Driver Information Systems (Vol. 12). Astigate: Aldershot, England.
Lindsey, R., 2009. Cost recovery from congestion tolls with random capacity and demand. J. Urban. Econ. 66, 16-24.
Lindsey, R., Daniel, T., Gisches, E., Rapoport, A., 2014. Pre-trip information and route-choice decisions with stochastic travel conditions: Theory. Transportation Research Part B: Methodological, 67, 187-207.
Lindsey, R., van den Berg, V.A.C., Verhoef, E.T., 2012. Step tolling with bottleneck queuing congestion. J. Urban Economics, 72 (1), 46-59.
Liu, Q., Jiang, R., Liu, R., Zhao, H., Gao, Z., 2020. Travel cost budget based user equilibrium in a

bottleneck model with stochastic capacity. Transportation Research Part B: Methodological, 139, 1-37.

Liu, Q., Jiang, R., Liu, W., Gao, Z., 2023. Departure time choices in the morning commute with a mixed distribution of capacity. Transportation Research Part C: Emerging Technologies, 147, 104011.

Liu, P., Liu, Y., 2018. Optimal information provision at bottleneck equilibrium with risk-averse travelers. Transportation Research Record, 2672(48), 69-78.

Long, J., Tan, W., Szeto, W.Y., Li, Y., 2018. Ride-sharing with travel time uncertainty. Transportation Research Part B: Methodological, 118, 143-171.

Long, J., Yang, H., Szeto, W.Y., 2022. Departure time equilibrium and tolling strategies of a bottleneck with stochastic capacity, Transportation Science, 56(1), 79-102.

Ma, W., Zeng, L., An, K., 2023. Dynamic vehicle routing problem for flexible buses considering stochastic requests. Transportation Research Part C: Emerging Technologies, 148, 104030.

Maher, M., Stewart, K., Rosa, A., 2005. Stochastic social optimum traffic assignment. Transportation Research Part B: Methodological, 39(8), 753-767.

Meng, Q., Liu, Z., 2011. Trial-and-error method for congestion pricing scheme under side-constrained probit-based stochastic user equilibrium conditions. Transportation, 38(5), 819-843.

Mun, S.I., 1999. Peak-load pricing of a bottleneck with traffic jam. Journal of Urban Economics, 46(3), 323-349.

Mun, S.L., 2003. Bottleneck Congestion with Traffic Jam: A reformulation and correction of earlier result. Kyoto University working paper (version of February), accessed from https://www.econ.kyoto-u.ac.jp/~mun/papers/Bottleneck0920.pdf on 17 November 2023.

Peer, S., Koster, P.R., Verhoef, E.T., Rouwendal, J. 2010. Traffic incidents and the bottleneck model. Working paper (version of 13 January 2014). Accessed from www.researchgate.net/profile/Stefanie-Peer/publication/229054198_Traffic_incidents_and_the_bottleneck_model/links/0c96052d3a1d55358b000000/Traffic-incidents-and-the-bottleneck-model.pdf on 21 November 2023.

Ren, H., Xue, Y., Long, J., Gao, Z. 2016. A single-step-toll equilibrium for the bottleneck model with dropped capacity. Transportmetrica B: Transport Dynamics, 4(2, 92-110.

Siu, B., Lo, H.K., 2009. Equilibrium trip scheduling in congested traffic under uncertainty. In: Lam, W.H.K., Wong, S.C., Lo, H.K. (Eds.), Transportation and Traffic Theory 2009: Golden Jubilee. Springer US, pp. 19-38.

Small, K.A., 2015. The bottleneck model: An assessment and interpretation. Economics of Transportation, 4(1-2), 110-117.

Small, K.A., Verhoef, E.T. and Lindsey, R., 2024. The Economics of Urban Transportation, 3rd Edition. Taylor & Francis.

Schrage, A., 2006. Traffic Congestion and Accidents. University of Regensburg Working Papers in Business, Economics and Management Information Systems, 419 (version of 9 November 2006). Accessed from https://epub.uni-regensburg.de/4535/1/Congestion_and_Accidents_WP.pdf on 21 November 2023.

Tan, Z., Yang, H., Tan, W., Li, Z., 2016. Pareto-improving transportation network design and ownership regimes. Transportation Research Part B: Methodological, 91, 292-309.

Van den Berg, V.A.C., 2012. Step-tolling with price-sensitive demand: why more steps in the toll make the consumer better off. Transportation Research Part A, 46(10), 1608-1622.

Verhoef, E.T., Emmerink, R.H.M., Nijkamp, P., Rietveld, P., 1996. Information provision, flat- and fine congestion tolling and the efficiency of road usage. Reg. Sci. Urban Econ., 26(5): 505-530.

Vickrey, W.S., 1969. Congestion theory and transport investment. Am. Econ. Rev (Papers and Proceedings), 59 (2), 251-261.

Xiao, L.L., Liu, R., Huang, H.J., 2014a. Stochastic bottleneck capacity, merging traffic and morning commute. Transportation Research Part E: Logistics and Transportation Review, 64, 48-70.

Xiao, L.L, Liu, R., Huang, H.J., 2014b. Congestion behavior under uncertainty on morning commute with preferred arrival time interval. Discrete Dynamics in Nature and Society, 2014, 767851.

Xiao, L.L., Huang, H.J., Liu, R.H., 2015. Congestion behavior and tolls in a bottleneck model with stochastic capacity. Transportation Science, 49(1), 46-65.

Yang, H., 1999. System optimum, stochastic user equilibrium and optimal link tolls. Transportation Science, 33, 354-360.

Yang, H., Huang, H.J., 1997. Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory. Transportation Research Part B: Methodological, 31(6), 425-440.

Yang, H., Meng, Q., 1998. Departure time, route choice and congestion toll in a queuing network with elastic demand. Transportation Research Part B: Methodological, 32(4), 247-260.

Yu, X., van den Berg, V.A.C., Li, Z.C., 2023. Congestion pricing and information provision under uncertainty: Responsive versus habitual pricing. Transportation Research Part E: Logistics and Transportation Review, 175, 103119.

Yu, X., van den Berg, V.A.C., Verhoef, E.T., 2024. Congestion pricing under dynamic flow congestion and heterogeneous preferences. Tinbergen Institute Discussion Paper, 2024-025. Accessed from https://papers.tinbergen.nl/24025.pdf on 7 February 2025.

Yu, Y., Han, X., Jiang, R., Darr, J., Jia, B., 2020. Departure time and route choices with accurate information under binary stochastic bottleneck capacity in the morning commute. IEEE Access, 8, 225551-225565.

Yu, Y., Han, X., Jia, B., Jiang, R., Gao, Z.Y., Zhang, H.M., 2021. Is providing inaccurate pre-trip information better than providing no information in the morning commute under stochastic bottleneck capacity? Transportation Research Part C, 126, 103085.

Zhang, F., Lu, J., Hu, X., Fan, R., Chen, J., 2022. Managing bottleneck congestion with tradable credit scheme under demand uncertainty. Research in Transportation Economics, 95, 101232.

Zhu, S., Jiang, G.G., Lo, H.K., 2018. Capturing value of reliability through road pricing in congested traffic under uncertainty. Transportation Research Part C, 94, 236-249.

Zhu, Z., Li, X., Liu, W., Yang, H., 2019. Day-to-day evolution of departure time choice in stochastic capacity bottleneck models with bounded rationality and various information perceptions. Transportation Research Part E, 131, 168-192.

Zhang, W.W., Zhao, H., Jiang, R., 2018. Impact of capacity drop on commuting systems under uncertainty. Journal of Advanced Transportation, 6809304.

Appendix A: Analytics for different toll schemes

A.1 The expected generalised price in each time window in the social optimum

***Situation 1*:** $[t_s, t_1]$. During the period, commuters always experience schedule delay early and never experience queuing regardless of the realised capacity. The departure rate $r_{FB}(t) = 1/\phi_{max}$, where $t \in [t_s, t_1]$. The generalised price during $[t_s, t_1]$ is:

$$P_{FB}(t) = \beta(t^* - t) + \tau_{FB}(t) \tag{A.1}$$

***Situation 2*:** $[t_1, t_2]$. During the second period, commuters always experience schedule delay early. They experience queuing for some realisations of capacity. They experience queuing if $q(t, \phi) > 0$, i.e., $\phi > \omega(t)$; otherwise, they do not. The generalised price during $[t_1, t_2]$ is:

$$P_{FB}(t) = \beta(t^* - t) + (\alpha - \beta)\int_{\omega(t)}^{\phi_{max}} q(t, \phi) f(\phi)\, d\phi + \tau_{FB}(t) \tag{A.2}$$

***Situation 3*:** $[t_2, t^*]$. Commuters possibly experience schedule delay either early or late, and possibly experience queuing depending on the realised capacity. They experience schedule delay late with queuing if $t + q(t, \phi) > t^*$, i.e., $\phi > \frac{t^* - t_1}{R(t) - (t_1 - t_s)/\phi_{max}}$; they experience schedule delay early with queuing if $t + q(t, \phi) < t^*$, i.e., $\omega(t) < \phi < \frac{t^* - t_1}{R(t) - (t_1 - t_s)/\phi_{max}}$; they experience schedule delay early without queuing if $\phi < \omega(t)$. The generalised price during $[t_2, t^*]$ is:

$$P_{FB}(t) = \int_{\phi_{min}}^{\omega(t)} \beta(t^* - t) f(\phi) d\phi + \int_{\omega(t)}^{\frac{t^* - t_1}{R(t) - (t_1 - t_s)/\phi_{max}}} [\beta(t^* - (t + q(t, \phi))) + \alpha q(t, \phi)] f(\phi) d\phi + \int_{\frac{t^* - t_1}{R(t) - (t_1 - t_s)/\phi_{max}}}^{\phi_{max}} [\gamma(t + q(t, \phi) - t^*) + \alpha q(t, \phi)] f(\phi)\, d\phi + \tau_{FB}(t) \tag{A.3}$$

***Situation 4*:** $[t^*, t_e]$. Commuters always experience schedule delay late regardless of the capacity, and experience queuing for some realisations of capacity. They experience queuing if $q(t, \phi) > 0$, i.e., $\phi > \omega(t)$, otherwise, they do not. The generalised price during $[t^*, t_e]$ is:

$$P_{FB}(t) = \int_{\phi_{min}}^{\omega(t)} \gamma(t - t^*) f(\phi) d\phi + \int_{\omega(t)}^{\phi_{max}} [\gamma(t + q(t, \phi) - t^*) + \alpha q(t, \phi)] f(\phi)\, d\phi + \tau_{FB}(t) \tag{A.4}$$

A.2 Proof of Proposition 5 on the uniform toll

The social surplus equals the total benefit minus total expected social cost. The total expected cost equals the expected travel cost multiplied by the number of users. Under the uniform toll, the social welfare needs to be maximised subject to the constraint that price equals the sum of mean travel cost and the toll. Therefore, the problem can be formulated as:

$$\max SS_{UT} = \int_0^{N_{UT}} D(n)\, dn - N_{UT} E\big(C_{UT}(N_{UT})\big)$$

$$\text{s.t. } D(N_{UT}) = E\big(C_{UT}(N_{UT})\big) + \tau_{UT}$$

where the subscript "UT" denotes the uniform toll scheme. N_{UT}, $E\big(C_{UT}(N_{UT})\big)$ and τ_{UT} denotes the demand, expected travel cost of user and toll under the uniform toll, respectively. To find the optimal constant toll τ_{UT}, the following Lagrangian is maximised,

$$L(N_{UT}, \tau_{UT}, \lambda) = \int_0^{N_{UT}} D(n)\, dn - N_{UT} E\big(C_{UT}(N_{UT})\big) + \lambda(D(N_{UT}) - E(C_{UT}) - \tau_{UT})$$

where λ is the multiplier. The first-order conditions are

$$\frac{\partial L}{\partial N_{UT}} = D(N_{UT}) - N_{UT}\frac{\partial E(C_{UT})}{\partial N_{UT}} - E(C_{UT}) + \lambda\left(\frac{\partial D(N_{UT})}{\partial N_{UT}} - \frac{\partial E(C_{UT})}{\partial N_{UT}}\right) = 0$$

$$\frac{\partial L}{\partial \tau_{UT}} = -\lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = D(N_{UT}) - E(C_{UT}) - \tau_{UT} = 0$$

These conditions imply that the toll is

$$\tau_{UT} = D(N_{UT}) - E(C_{UT}) = N_{UT}\frac{\partial E(C_{UT})}{\partial N_{UT}} \tag{A.5}$$

Therefore, the uniform toll is set such that the average marginal social cost equals the price, and thus the toll equals the average marginal external cost.

A.3 Second step of the optimization of the single-step toll scheme

In the first step of the optimisation, the step part of the toll is obtained by minimising the total social cost under the given demand. In the second step, the social welfare is maximised to find the optimal flat part of toll implementing during the whole peak. The problem can be formulated as follows,

$$\max SS_{ST} = \int_0^{N_{ST}} D(n)\, dn - TC_{ST}(N_{ST})$$

$$\text{s.t. } D(N_{ST}) = AC_{ST}(N_{ST}) + \rho_{ST}$$

where the subscript "ST" denotes the single-step toll scheme. TC_{ST} and AC_{ST} denote the total expected social cost and the average expected travel cost under the step part of toll, respectively, and $AC_{ST} = \frac{TC_{ST}}{N_{ST}}$. ρ_{ST} are defined as $\rho_{ST} = \frac{\rho_t N_1}{N_{ST}} + \mu$, where ρ_t, μ, and N_1 are the time-variant step part of the toll, time-invariant part of toll and the number of users departing from t^+ to t^-, respectively.

Here, similar to the procedure in Van den Berg (2012), in the first step, given the number of users, the time-variant part ρ_t and the step tolling period are obtained by minimising the total expected social cost. The results of TC_{ST}, ρ_t and N_1 follow those in Long et al. (2022), and TC_{ST} and ρ_t are both the function of travel demand, as given in their study. In the second step, we use a Lagrangian to maximising the social welfare, since total expected travel cost and the step part of toll can be expressed as a function of travel demand. To find the optimal flat part of the toll μ, the following Lagrangian is maximised,

$$L(N_{ST}, \mu, \lambda) = \int_0^{N_{ST}} D(n)\, dn - TC_{ST}(N_{ST}) + \lambda(D(N_{ST}) - AC_{ST}(N_{ST}) - \rho_{ST})$$

The first-order conditions are

$$\frac{\partial L}{\partial N_{ST}} = D(N_{ST}) - \frac{\partial TC_{ST}}{\partial N_{ST}} + \lambda\left(\frac{\partial D(N_{ST})}{\partial N_{ST}} - \frac{\partial AC_{ST}}{\partial N_{ST}} - \frac{\partial \rho_{ST}}{\partial N_{ST}}\right) = 0$$

$$\frac{\partial L}{\partial \mu} = -\lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = D(N_{ST}) - AC_{ST} - \rho_{ST} = 0$$

These conditions imply that the optimal flat part of the toll can be given as follows,

$$\mu = \rho_{ST} - \frac{\rho_t N_1}{N_{ST}} = \frac{\partial TC_{ST}}{\partial N_{ST}} - AC_{ST} - \frac{\rho_t N_1}{N_{ST}} \tag{A.6}$$

The average marginal social cost is $MSC_{ST} = \frac{\partial TC_{ST}}{\partial N_{ST}}$. Therefore, Eq. (A.6) can also be expressed as: $\mu = MSC_{ST} - AC_{ST} - \frac{\rho_t N_1}{N_{ST}} = MEC_{ST} - \frac{\rho_t N_1}{N_{ST}}$, where MEC_{ST} is the average marginal external cost. The optimal flat part of toll μ is set such that price equals the average marginal social cost. The optimal demand can be obtained from $D(N_{ST}) = MSC_{ST}$.

A.4 The results in Case II for the single-step toll

As given in Long et al. (2022), the expected social cost in Case II is:

$$TC_{ST}(N_{ST}) = (N_{ST})^2 \varsigma(\overleftrightarrow{\phi}, \vec{\phi})$$

where $\varsigma(\overleftrightarrow{\phi}, \vec{\phi}) = \beta\vec{\phi} - X(\overleftrightarrow{\phi}, \vec{\phi})\left[1 - \beta(\overleftrightarrow{\phi} - \vec{\phi})Y(\overleftrightarrow{\phi}) + \frac{2\left(\beta\vec{\phi} - X(\overleftrightarrow{\phi},\vec{\phi})\right)}{(\alpha+\gamma)\bar{\phi}} - X(\overleftrightarrow{\phi}, \vec{\phi})Y(\overleftrightarrow{\phi})\right]$, $X(\overleftrightarrow{\phi}, \vec{\phi}) = \frac{\beta\vec{\phi} - (\alpha+\gamma)(H(\vec{\phi}) - \vec{\phi})}{(\alpha+\gamma)[H(\overleftrightarrow{\phi})Y(\overleftrightarrow{\phi}) - H(\vec{\phi})/W(\overleftrightarrow{\phi})] + (\alpha+\beta+\gamma)[\vec{\phi}/W(\overleftrightarrow{\phi}) - \overleftrightarrow{\phi}Y(\overleftrightarrow{\phi})] + 1}$. As defined in Long et al. (2022), $\overleftrightarrow{\phi}$ and $\vec{\phi}$ are the reciprocals of average departure rates during the period from the departure time of the first commuter to the departure time of the first commuter who pays the toll and during the period from the departure time of the first commuter who pays the toll to the end of the tolling period, respectively. $\overleftrightarrow{\phi}$ and $\vec{\phi}$ are obtained by $\frac{\partial\varsigma(\overleftrightarrow{\phi},\vec{\phi})}{\partial\overleftrightarrow{\phi}} = 0$ and $\frac{\partial\varsigma(\overleftrightarrow{\phi},\vec{\phi})}{\partial\vec{\phi}} = 0$. The total toll revenue follows Eq. (38), where $\rho_t = N_{ST}X(\overleftrightarrow{\phi}, \vec{\phi})$ and $N_1 = N_{ST} - \frac{\rho_t}{W(\overleftrightarrow{\phi})}$ in Case II. As proved in Proposition 20 by Long et al. (2022), $\overleftrightarrow{\phi}$ and $\vec{\phi}$ are independent of the number of users, and TC_{ST} and ρ_t depend on the demand. Hence, the average MEC in Case II is:

$$MEC_{ST} = N_{ST}\varsigma(\overleftrightarrow{\phi}, \vec{\phi}) \tag{A.7}$$

The optimal time-invariant part of the toll in Case II is:

$$\mu = N_{ST}\left[\varsigma(\overleftrightarrow{\phi}, \vec{\phi}) - X(\overleftrightarrow{\phi}, \vec{\phi})\left(1 - \frac{X(\overleftrightarrow{\phi},\vec{\phi})}{W(\overleftrightarrow{\phi})}\right)\right] \tag{A.8}$$

where $\mu > 0$ in Case II. The generalised price in Case II is:

$$P_{ST} = 2N_{ST}\varsigma(\overleftrightarrow{\phi}, \vec{\phi}) \tag{A.9}$$

From Eq. (A.9), the optimal demand under a step toll can be found by solving $P_{ST} = D(N_{ST})$.

Appendix B: Proofs

B.1 The proof of Proposition 3

Proof. Differentiating Eq. (17), for our uniform distribution, we have

$$\omega(t)\mu_1(t, \omega(t))\frac{d\omega(t)}{dt} = \frac{dSDC(t)}{dt}(\omega(t) - \phi_{min}) - \alpha(\phi_{max} - \omega(t))$$

For early departures, $\frac{dSDC(t)}{dt} = -\beta < 0$ and the right-hand side of the equation is negative. Since $\omega(t) \geq 0$ and $\mu_1(t, \omega(t)) \geq 0$, condition $\frac{d\omega(t)}{dt} \leq 0$ should be met. By $r_{FB}(t) = \frac{1}{\omega(t)}$, we have $\frac{dr_{FB}(t)}{dt} \geq 0$.

For late departures, $\frac{dSDC(t)}{dt} = \gamma > 0$. Let $\Psi(t) = \frac{dSDC(t)}{dt}(\omega(t) - \phi_{min}) - \alpha(\phi_{max} - \omega(t))$. Differentiating $\Psi(t)$ with respect to *t*, we have $\Psi'(t) = (\alpha + \gamma)\frac{d\omega(t)}{dt}$.

Suppose $\frac{d\omega(t)}{dt} > 0$, then we have $\omega(t)\mu_1(t, \omega(t))\frac{d\omega(t)}{dt} \geq 0$. Therefore, $\Psi(t) \geq 0$, $\Psi'(t) = (\alpha + \gamma)\frac{d\omega(t)}{dt} > 0$, and $\Psi(t)$ monotonously increases with *t*. This also indicates that $\Psi(t)$ cannot approach to zero at t_e. Since $\mu_1(t_e, \omega(t)) = 0$ at t_e as, we have $\Psi(t)|_{t=t_e} = 0$. The two results contradict each other. This means the assumption

$\frac{d\omega(t)}{dt} > 0$ is not valid. Therefore, $\frac{d\omega(t)}{dt} \leq 0$ should hold, and we have $\frac{dr_{FB}(t)}{dt} \geq 0$.

This completes the proof. □

B.2 The proof of Lemma 2

Proof. The departure rate $r_{FB}(t) = 1/\phi_{max}$, where $t \in [t_s, t_1]$. Then, the cumulative departures at t_1 is $R(t_1) = (t_1 - t_s)/\phi_{max}$. From the schedule delay experience during *Situation 2* and *Situation 3*, the boundary condition for t_2 is that commuters departing at t_2 arrive exactly at t^* when the realised capacity is the minimum, i.e., $\phi = \phi_{max}$. Then, we have $R(t_2) = (t^* - t_s)/\phi_{max}$. The cumulative departures at t^* is obtained by substituting $t = t^*$ into Eq. (A.3) and (A.4), and then equalising the two equations.

This completes the proof. □

B.3 The proof of Lemma 3

Proof. By the definition of $\omega(t)$, we have $q\big(t_e, \omega(t_e)\big) = 0$. i.e.,

$$\omega(t_e)\left[N_{FB} - \frac{(t_1 - t_s)}{\phi_{max}}\right] - (t_e - t_1) = 0$$

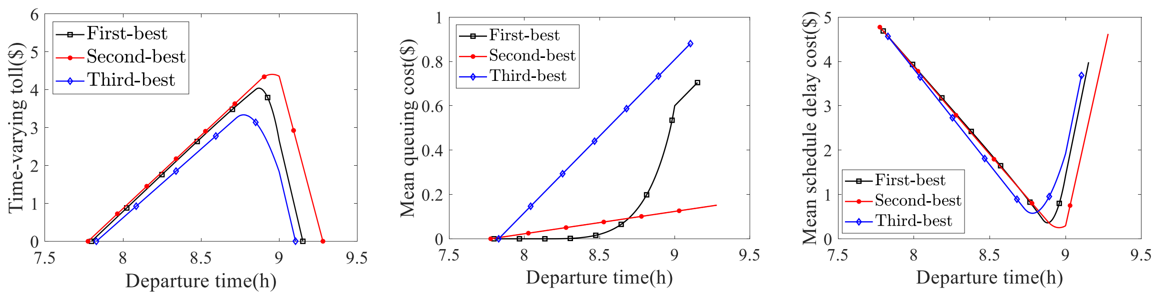
In the optimum, $P_{FB}(t_s) = P_{FB}(t_e)$, then we have

$$\frac{(\alpha+\gamma)}{2} f(\phi)\left(N_{FB} - \frac{t_1 - t_s}{\phi_{max}}\right)\left((\phi_{max})^2 - \omega^2(t_e)\right) - \gamma(t^* - t_1) = \beta(t^* - t_s)$$

Substituting $f(\phi) = 1/(\phi_{max} - \phi_{min})$ and $\omega(t_e) = (\alpha\phi_{max} + \gamma\phi_{min})/(\alpha + \gamma)$ into the two equations and rearrange them, the results can be obtained.

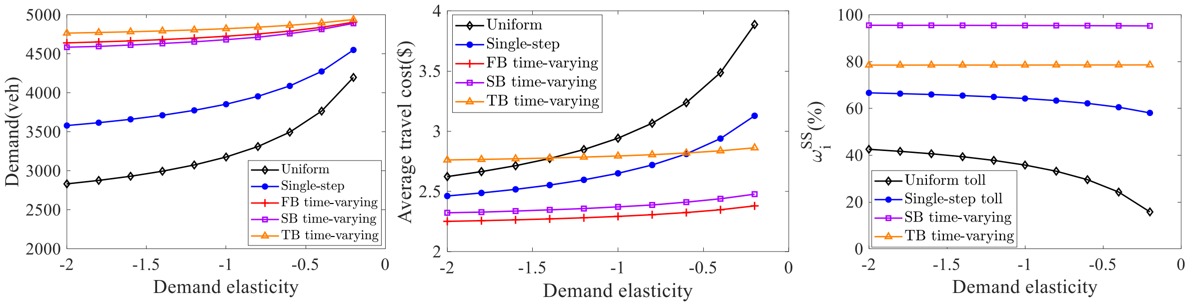
This completes the proof. □

Appendix C: Numerical results under low uncertainty



(a) Toll (b) Mean queuing cost (c) Mean schedule delay cost

Fig. C.1: Comparisons of toll, queuing cost and schedule delay cost with a small spread e=0.4 *s/veh*.



(a) Demand (b) Average travel cost (c) Relative efficiency

Fig. C.2: The effect of elasticity on the outcomes with a small spread e=0.4 *s/veh*.