TI 2022-045/VII
Tinbergen Institute Discussion Paper

Self-financing roads under coarse tolling and preference heterogeneity

Revision: January 2024

Vincent A.C. van den Berg[1]

[1] Vrije Universiteit Amsterdam and Tinbergen Institute

Self-financing roads under coarse tolling and preference heterogeneity

Version of 12 January 2024

Vincent A.C. van den Berg[a,b,*]

a: Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands

b: Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS Amsterdam, The Netherlands

*: email: v.a.c.vanden.berg@vu.nl , tel: +31 20 598 6049 , ORCID ID: 0000-0001-8337-7986

Abstract

We consider whether a road is self-financing under flat or step tolling and optimized bottleneck capacity while incorporating preference heterogeneity and price-sensitive demand. Previous work has shown that a *sufficient condition* for the toll revenue to equal the capacity cost is that the toll equals the marginal external costs (MECs) of all types of users at all their travel moments. However, under 'ratio' heterogeneity between the value of time (VOT) and values of schedule delay, a coarse toll must differ from the heterogeneous MECs. We expand the literature by showing that the capacity rule also has a second-best correction: the capacity is set higher than following the first-best rule to reduce the distortion from overpricing High-VOT users. This issue has been ignored in previous work and makes self-financing less likely than previously thought, but it can still occur if Low-VOT users are much more price sensitive than High-VOT users. In our numerical model, the Low-VOT type must be almost twice as price sensitive as the High-VOT type for there not to be a loss, and, typically, there is a 0% to 10% loss. Nevertheless, imposing self-financing only causes a tiny welfare loss. We also discuss other forms of heterogeneity. Under proportional heterogeneity, the self-financing result holds as the coarse toll equals the homogeneous MEC. Under heterogeneity in the preferred arrival time, the self-financing typically also holds.

Keywords: self-financing; second-best; capacity setting; flat toll; step toll; heterogeneity

JEL codes: D62, H23, R41, R48

1. Introduction

The self-financing result of Mohring and Harwitz (1962) states that—if the toll equals the marginal external congestion cost (MEC) and three other technical conditions hold—the toll revenue will exactly equal the cost of the optimized road capacity. This ‘first-best’ capacity minimizes the total cost, which is the sum of capacity and total travel cost. Self-financing can help with the social acceptability of congestion pricing. It means that people have to pay to use the roads, but this contribution is used for road capacity. It means no cross-subsidization between modes, and no distortive taxes are needed. It can ensure adequate funds for new roads when government budgets are tight. Finally, road capacity is mainly set to deal with peak demand, so it can be seen as fair that users of the center peak pay more. All this is why self-financing has been extensively studied (see the literature review in Section 2).

Our main methodological contribution is deriving explicit formulas for the second-best bottleneck capacity under coarse tolling, preference heterogeneity and price-sensitive demand. Previous work has not considered that the capacity setting will be second-best if the toll cannot equal the MEC throughout the peak. Both the heterogeneity and the price sensitivity substantially complicate the capacity setting. The coarse tolls are weighted averages of the MECs. The weights depend on the derivatives of the demand and cost functions and are *independent* of the numbers of users of the types.[1] Using these results, we derive the resulting profit or loss, allowing us to analyze when a system is self-financing.

Our policy contribution is adding to the discussion on self-financing roads by considering coarse tolls under heterogeneity and dynamic congestion. We find that it is most likely that the system will make a small loss. Imposing that the toll should cover the capacity cost typically results in a minuscule welfare loss.[2]

Mohring and Harwitz (1962) studied homogeneous users and static congestion. Arnott et al. (1993a) showed that the self-financing result carries over to bottleneck congestion, both with a fine toll that can vary every second and with a coarse toll. In reality, congestion varies over the day, but tolls are coarse, being either a constant flat toll—as in London—or at most having a few steps—as in Singapore, Stockholm, and on many roads and bridges in the USA. Thus, we consider two types of coarse tolls: 1) a flat toll that is constant throughout the peak and 2) a step toll that varies in discrete steps.

Under preference heterogeneity, Arnott and Kraus (1995) found that a *sufficient condition* for self-financing in any dynamic congestion model (in addition to the conditions from Mohring and Harwitz (1962)) is that the coarse toll equals the MEC throughout the peak. They then consider what we call ‘ratio heterogeneity’ in the ratio of the value of time (VOT) and the value of schedule delay, showing that the sufficient condition cannot hold.

[1] Although, as a reviewer pointed out, the demand derivative for a type may vary with the number of users of this type, only for a linear demand is the derivative constant.

[2] This last result is in line with Verhoef and Mohring (2009) who considered static congestion.

This ratio heterogeneity is also called ‘flexibility heterogeneity’ since the ratio determines how flexible people are regarding when to arrive. It means people differ in their preferences for trading off travel time and schedule delay (Hall, 2018). This heterogeneity could, for example, stem from differences in job type, trip purpose or family status. Income differences may have little to do with this heterogeneity. For example, a low-income shift worker or a parent picking up their child is very inflexible in when to travel, while a pensioner or shopper with the same income may be very flexible; a surgeon may be inflexible in when to start work, while a work-from-home professional with the same income may be more flexible (Van den Berg and Verhoef, 2011a).[3] Empirical research has shown that, for the same income, such differences result in substantial preference heterogeneity (Kouwenhoven et al., 2014) and, even for the same person, there are changes in preferences over days and trip purposes (Börjesson et al., 2013).

The marginal external congestion cost (MEC) is higher for inflexible users—who have a VOT closer to their value of schedule delay early—than for flexible ones.[4] Hence, an anonymous coarse toll cannot always equal the heterogeneous MECs. Accordingly, we derive that a coarse toll must be second-best: the inflexible users will see a toll below their MEC and the flexible users a toll above it. Only a few papers have looked at user heterogeneity and self-financing in a dynamic congestion model, which is the present paper’s aim: how large a deficit or profit will there be, and what is the welfare loss of imposing self-financing?

What Arnott and Kraus (1995) and all others did not consider is that, under ratio heterogeneity, the capacity rule will also be second-best. The second-best capacity is higher than following the first-best rule, which minimizes total cost.[5] Even with the extra second-best capacity, the scheme may have a zero or even positive profit if the average second-best toll well exceeds the expected MEC. Our toll rules show that, for this to happen, the inflexible users must be much more price sensitive than the flexible ones. The single-step toll tends to have a lower loss or profit than the flat toll and is less likely to have a loss. This is because the step toll has a smaller second-best capacity adjustment.

Ratio heterogeneity seems to be the most interesting form of heterogeneity as it affects self-financing. We find that separate ‘proportional heterogeneity’—which varies all values of time and schedule delay in a fixed proportion—and heterogeneity in the preferred arrival time do not lead to a heterogeneous MEC; and hence the system is self-financing. With heterogeneity in preferred arrival times, the heterogeneity needs to be such that the peak is single-peaked in the morning with a coarse toll and single-peaked within a tolling period with step tolling. This is a pattern in travel times that we typically observe in reality. If this does not hold, the MEC will be heterogeneous, and self-financing is not ensured. Heterogeneity between values of schedule

[3] The value of time (VOT) is the cost of one hour of travel time, the value of schedule delay early (late) is the cost of arriving one hour earlier (later) than most preferred. The VOT is the ratio of the marginal utility of time and (the minus of) the marginal utility of income, and similar for the values of schedule delay. The primary effect of an income increase on the values of time and schedule delay is probably via the decrease in the marginal utility of income it causes. The marginal utilities of time and schedule delay may be more affected by such things as job type, gender, family situation, trip purpose and age than by income.

[4] This is because more inflexible users need a steeper travel time development over time to be in user equilibrium than more flexible ones, thus causing larger congestion effects, and because the travel closer to the center of the peak when delays and externalities are larger.

[5] This second-best correction raises welfare by increasing the number of flexible users who face a toll that exceeds their MEC.

delay early and late means that there must also be ratio heterogeneity, and thus it thus has similar effects. The equations for toll and capacity setting are already very complex; adding more realism with multiple dimensions of heterogeneity would make analytical analysis even more difficult, if not impossible. Considering both ratio and proportional heterogeneity yields qualitatively the same results as our model with only ratio heterogeneity but substantially complicates the exposition.

The following section will give an extended literature review, showing how our paper fits in and extends the literature. Section 3 presents the basic model. Section 4 considers flat and single-step tolling under ratio heterogeneity. Section 5 turns to the numerical model and conducts extensive sensitivity analyses. Sections 6 and 7 look at other forms of heterogeneity. Section 8 concludes. The nomenclature box in Appendix D summarizes the notation.

2. Extended literature review

The self-financing result was first derived by Mohring and Harwitz (1962). It states that toll revenue from congestion externality pricing will exactly cover the cost of optimally set road capacity when: i) capacity and number of users are continuous; ii) capacity cost is homogeneous to the degree one in capacity (i.e., doubling capacity doubles capacity costs); and iii) that per car travel cost only depends on the ratio of the number of cars and capacity (i.e., doubling both usage and capacity leaves travel cost unchanged). We assume the conditions hold throughout this paper. The congestion charge only covers the marginal external congestion costs. The pricing of other externalities would come on top of this.

This analysis has been extended to include that roads last many years and the amount of travel changes over time (e.g., Arnott and Kraus, 1998b), that there is uncertainty in what demand and costs will be (e.g., D'Ouville and McDonald, 1990; Fu et al., 2018, Kraus, 1982; Lindsey and de Palma, 2014; Lu and Meng, 2017) , that input prices may vary with the amounts used (e.g., Small, 1999), and that realistic road networks consist of many roads (e.g., Yang and Meng, 2002).[6] Verhoef and Mohring (2009) considered imposing self-financing under static congestion when some of the assumptions of Mohring and Harwitz (1962) do not hold. They found that imposing self-financing when the assumptions are relaxed tends to lead to a small welfare loss at the most.

These papers focused on static congestion; the extension to dynamic congestion was introduced by Arnott et al. (1990, 1993a) and Arnott and Kraus (1993, 1995, 1998a). The bottleneck model is the workhorse model for dynamic congestion, where travel times vary over the peak, and it has been very heavily used in the literature. See Small (2015) and Li et al. (2020) for detailed overviews.

Heterogeneity in preferences is an essential component of our setting as it can cause the self-financing result to break down, and, of course, people are heterogeneous in reality. It was first

[6] See Lindsey (2012) for a more extensive overview.

introduced to the bottleneck model by Vickrey (1973) and extended by Newell (1987), Arnott et al. (1988, 1993a), Lindsey (2004) and many others. Van den Berg and Verhoef (2011b) introduced the distinction between ratio/flexibility heterogeneity and proportional heterogeneity. Ratio heterogeneity means that there is heterogeneity in the ratio of the value of time to the value of schedule delay. This means that user types separate over time if there is congestion. Proportional heterogeneity varies all values of time and schedule delay in fixed proportions. This heterogeneity affects the outcome under fully time-variant and step tolling but not when there is flat or no tolling (Van den Berg, 2014). Later papers—such as Wu and Huang (2015), Liu et al. (2015a), Chen et al. (2015a), Hall (2018, 2021, 2023), Akamatsu et al. (2021) and Guo et al. (2023)—have looked at multiple dimensions of heterogeneity.[7]

With flat tolling, the toll is a constant amount throughout the peak. Examples would be the London congestion charge and the various schemes in Norway. For the bottleneck model under homogeneous users, Arnott et al. (1990, 1993) showed that the optimal flat toll equals the marginal external cost (MEC), where this MEC is constant over time under flat tolling. The MEC gives the difference between the marginal social cost and private travel cost.

With a step toll, the toll has one or more discrete steps over time, but it is constant otherwise. Examples are the schemes in Singapore and Stockholm and various toll roads and lanes in the USA. The various proposed step-toll models differ in how they ensure that the generalized price is constant over time. These include the ADL model (Arnott et al.,1990), the Laih model (Laih, 1994, 2004) and the Braking model (Lindsey et al., 2012; Xiao et al., 2012). Again, under homogeneity, the step toll will equal the MEC that now has steps in it (Van den Berg, 2012). Various extensions have been made by, for instance, Knockaert et al. (2016), Ren et al. (2016Li et al. (2017) and Xu et al. (2019).

Let us now turn to the papers on coarse tolling under preference heterogeneity. For the bottleneck model, Van den Berg and Verhoef (2011b) studied flat tolling under ratio heterogeneity, and Xiao et al. (2011) added proportional heterogeneity to the ADL step-toll model. Under homogeneous users and bottleneck congestion, a fully time-variant congestion toll leaves the generalized prices the same as without tolling. Xiao et al. (2011) find that their ADL step toll lowers generalized prices. This also occurs with homogeneity due to the 'mass departures' (Lindsey et al., 2012), but this is strengthened by the proportional heterogeneity making tolling more beneficial for users. Xu et al. (2019) studied the ADL, Laih and Braking models under proportional heterogeneity, while Van den Berg (2014) studied separate ratio heterogeneity, proportional heterogeneity and heterogeneity between values of schedule delay early and late. Finally, Chen et al. (2015b) and Li et al. (2017) looked at coarse tolling under more general heterogeneity. Liu et al. (2015b) looked at reserving travel windows as an alternative to (step) tolling.

[7] Hall (2021, 2023) shows that adding heterogeneity in the preferred arrival time to ratio and proportional heterogeneity has large effects. Conversely, Arnott et al. (1988) found that if there is only heterogeneity in the preferred arrival time, and not also other heterogeneity, this does very little; that is, as long as the queue is single peaked, if there are multiple peaks then travel times can be much smaller.

Finally, a small body of literature investigates coarse tolling under other dynamic congestion models. Bichsel (2001) used a model with peak and off-peak travel where each period has separate static congestion. He assumed that off-peak demand is more price elastic than peak demand and found that the degree of self-financing is lower when the ratio of peak demand to off-peak demand is high. Chu (1999) used his dynamic flow congestion model. Börjesson and Kristoffersson (2012) used their model for the Stockholm step-toll system. Zheng et al. (2012) used a macroscopic fundamental diagram to model a flat cordon change for Zurich. Ge et al. (2016) used cell-transmission models to study step tolling.

3. Model setup

3.1. General costs functions and welfare

This section focuses on arbitrary discrete heterogeneity in values of time and schedule delay, with a homogeneous preferred arrival time. Therefore, for now, we consider any number of discrete types of users; Sections 4 to 6 will consider only two types for ease of presentation. A type is defined by its users having the same travel time and schedule delay preferences. There are differences in willingness to pay for the trip within a type, as there is an independent price-sensitive demand for each type: $D_i[N_i]$. The N_i is the number of users of type i. Throughout this paper, we assume that the three conditions of Mohring and Harwitz (1962) hold. We will not go into the details of how the bottleneck model works. See Small and Verhoef (2007), Small (2015) and Li et al. (2020) for overviews.

We remind the reader that throughout this paper, we assume that the three technical assumptions from Mohring and Harwitz (1962) hold. Otherwise, even with homogeneity, self-financing is not ensured.

The travel cost per trip for type i user as a function of the arrival time t equals

$$c_i[t] = Max(-\beta_i \cdot t, \gamma_i \cdot t) + \alpha_i \cdot TT\,[t]. \qquad (1)$$

It is the sum of the schedule delay and travel time cost. The preferred arrival time, t^*, is normalized to zero and is assumed to be the same for all. So, t=0 means an arrival at the most preferred moment. The β_i is the value of schedule delay early for type i: it is the value of an hour earlier arrival than most preferred. The γ_i is the corresponding value for an hour-late arrival. The $TT[t]$ is the travel time when arriving at t. Throughout the paper, we will use square brackets to indicate that something is a function of what is listed within the brackets. The α_i is the value of time (VOT) for type i. We assume $\alpha_i > \beta_i > 0$ and $\gamma_i > 0$ to ensure the standard equilibrium without mass departures. We normalize the free-flow travel time to zero.[8] Travel time equals the number of cars in the queue before reaching the bottleneck divided by the capacity s. The queue is assumed to be at a single point. Appendix D summarizes the meaning of the notation.

[8] This normalization does not affect results. The numerical model will be more realistic and includes a free-flow travel time of 30 minutes and fuel costs.

In the user equilibrium of a flat toll, the travel cost of a type i is constant over time. We also consider a single-step toll, which is higher in the center of the peak, lower outside it, and is constant otherwise. In its user equilibrium, the cost is lower in the center peak period than in the shoulder periods, and we thus may have two travel cost functions for a type.

Total cost is the sum of the capacity cost, $k \cdot s$, and the travel costs of the different types:

$$TC = k \cdot s + \sum N_i \cdot E[c_i]. \tag{2}$$

Here, $E[c_i]$ gives the average of the travel cost per person for i, averaged over arrival time and weighted by the arrival rate of i. We assume that the capacity cost is linear in capacity s. Hence, k is the marginal capacity cost. N_i is the total number of users of type i.

There are separate inverse demands for all types. The generalized price—henceforth price for brevity—is the sum of the travel cost, c_i, and the possible toll, τ. In user equilibrium, the price for type i, P_i, equals its inverse demand, $D_i[N_i]$, for all moments that a type i user arrives and is thus constant over time; the price is no lower at all moments that there are no arrivals of type i:

$$D_i[N_i] = P_i \equiv c_i[t] + \tau[t] \tag{3}$$

Consumer benefit, B_i, for type i is the integral of its inverse demand, and welfare equals the sum of the consumer benefits minus the total cost:

$$B_i[N_i] = \int_0^{N_i} D_i[n_i]\,dn_i, \tag{4}$$

$$W = \sum B_i - TC. \tag{5}$$

3.2 Capacity setting and self-financing when the toll equals the marginal external costs throughout the peak

This section will discuss the general outcome when the coarse toll equals the (potentially heterogeneous) marginal external costs (MECs) throughout the peak. The results follow from Arnott and Kraus (1995) and Van den Berg and Verhoef (2011b), so we can be brief. But this will result in two Lemmas that form the start of our analysis. Arnott and Kraus (1995) assumed heterogeneity in the value of schedule delay for a fixed value of time. In contrast, we use a heterogeneous value of time and fixed values of schedule delay. Their setup implies that, effectively, there is both ratio and proportional heterogeneity,[9] but, as we will show, only the presence of the ratio heterogeneity affects self-financing. Considering both simultaneously would leave results qualitatively the same, but complicates the exposition.

[9] There is ratio heterogeneity as the ratio α/β_i (and by extension α/γ_i) varies over persons and so, with queueing, users self-separate over arrival times. But, there is also, in effect, proportional heterogeneity as the values of schedule delay vary, and so $\delta_i \equiv \beta_i \cdot \gamma_i / (\beta_i + \gamma_i)$ varies and scales the travel cost, and it also leads to self-separation without queueing (which does not occur under pure ratio heterogeneity).

Lemma 1: First-best capacity[10]

Consider M discrete types of users.[11] Let us assume that, in user equilibrium, the average travel cost,[12] $E[c_i]$, of type i increases with the number of users of any type and decreases with the bottleneck capacity, s.[13] Suppose the toll equals the marginal external cost (MEC) of all types at all moments. Then, the first-best optimal capacity minimizes the total cost from (2) and follows the ***first-best*** *condition:* $-\sum N_i \cdot \partial E[c_i]/\partial s = k$.

Proof: The socially optimal toll optimizes the number of users of each type, even if the toll is flat or has a single step. Hence, maximizing welfare with respect to capacity, s, is equivalent to minimizing total cost as the numbers of users solely determine consumer benefit. The first order condition is $k+\sum N_i \cdot \partial E[c_i]/\partial s = 0$, which implies the first-best capacity rule. At this optimum, the second-order conditions also hold. □

Remark 1: The first-best capacity rule *only holds* when the toll optimizes the number of users of *each* type. When the toll does not equal the (potentially heterogeneous) MECs at all moments, the capacity rule will have a second-best adjustment. We will show this for ratio heterogeneity later on.

Remark 2: The condition $-\sum N_i \cdot \partial E[c_i]/\partial s = k$ implies that, at the optimal capacity, the extra cost of a marginal expansion of capacity, which is k, equals the reduction in total travel cost it causes, which is $-\sum N_i \cdot \partial E[c_i]/\partial s$. The minus sign ensures the number is positive as travel costs fall with s. So, at the optimal capacity, the marginal cost of capacity expansion equals the marginal capacity cost.

Lemma 2: Self-financing

With a total-cost minimizing capacity, a ***sufficient condition*** *for exact self-financing is that the (coarse) toll equals the marginal external cost (MEC_i) of type i at each moment a type i user travels. Self-financing means that the toll revenue equals road capacity cost.*

Proof: As discussed, this Lemma is the starting point of this paper and is based on Arnott and Kraus (1995, pp. 279–280). They prove this for any dynamic congestion model with total costs that are homogeneous to the degree zero in the ratio of the number of users of each type and s, which is true with bottleneck congestion.[14] □

[10] We use the term 'Lemma' here instead of proposition as it is a known result from the literature that forms the start of our analysis.

[11] As noted, there is arbitrary discrete heterogeneity in values of time and schedule delay, with a homogeneous preferred arrival time.

[12] Averaged over arrival time t and weighted by the arrival rates of type i at t.

[13] These assumptions hold in our single bottleneck setting when there is a continuous peak, without a period without departures inside it (see, e.g., Arnott et al., 1988; Arnott and Kraus, 1995; Van den Berg and Verhoef, 2011b). They also obviously hold for the travel cost functions we will derive below. In fact, we could even be more general and allow $\partial E[c_i]/\partial N_i>0$ but $\partial E[c_i]/\partial N_j \geq 0$ for type $j \neq i$. In networks with pure bottlenecks, it is possible to have a paradox—akin to Breass' paradox—where an (average) travel cost can increase with a capacity (e.g., Arnott et al. 1993b). We focus on a single bottleneck.

[14] To see this in their proof, remember that a toll equal to MEC_i means that a type i's inverse demand, D_i, equals their marginal social cost.

Even if a second-best optimal toll differs from the marginal external cost at some times, a scheme with optimized capacity could still have a zero profit if the profits at some moments happen to cancel out the losses at others. However, this occurs only for unique combinations of parameters, and, as we will see, losses are likely to occur.

In conclusion, Section 3 presented our model setup used throughout this paper. It also presented previous results in our notation, which will prove helpful for comparison and understanding.

4. Two-type ratio heterogeneity

4.1 Flat toll and ratio heterogeneity

This section considers 'ratio heterogeneity' where the value of time (α_i) varies over two types of users, while the values of schedule delay are fixed. This ensures that the ratio α_i/β and α_i/γ vary without values of schedule delay also varying. The type with the lower value of time (VOT) is less flexible as the VOT is then closer to the values of schedule delay early. The equations we derive for toll and capacity will be very complex. Therefore, including more realism with many types or multiple dimensions of heterogeneity would make analytical analysis difficult, if not impossible. Our results would hold qualitatively unchanged when considering both ratio and proportional heterogeneity.

With only ratio heterogeneity, with flat tolling and without any tolling, each type's travel costs, MECs and MSCs are constant over time. With a flat toll, the toll always equals μ:

$$\tau[t]=\mu.$$

For **given numbers of users of each type**, a flat toll leads to the same user equilibrium as no tolling. So, we can use the results of Arnott et al. (1988) and Van den Berg and Verhoef (2011b):

$$c_L^F = \delta \frac{N_L + \frac{\alpha_L}{\alpha_H} \cdot N_H}{s}, \tag{6a}$$

$$c_H^F = \delta \cdot \frac{N_L + N_H}{s}. \tag{6b}$$

The travel cost for a type is constant over time (for all moments that it travels in equilibrium), so we can omit the t as a variable. Superscript ^{F} indicates the flat toll equilibrium and $\delta=(\beta\cdot\gamma)/(\beta+\gamma)$ is a compound preference parameter. A subscript L indicates the Low-VOT type that has the lower value of time and is thus less flexible. The H is for the High-VOT type, which cares more about travel time than schedule delay. The High-VOT users choose to travel at the edges of the peak, where travel times are short and schedule delays are high. The Low-VOT users choose to travel in the center of the peak. Appendix D summarizes the meaning of the notation.

The above travel costs lead to a total cost of

$$TC^F = c_L^F \cdot N_L + c_H^F \cdot N_H + k \cdot s = \delta \cdot \frac{\left(N_L + N_H\right)^2}{s} - \delta \cdot \frac{N_L \cdot N_H}{s}\left(1 - \frac{\alpha_L}{\alpha_H}\right) + k \cdot s. \tag{7}$$

Marginal external cost, MEC_i, of type i equals its marginal social cost, $MSC_i^F = \partial TC^F / \partial N_i$, minus its travel cost c_i^F. Under ratio heterogeneity, the marginal external cost of the Low-VOT users (MEC_L) exceeds that of the High-VOT users:

$$\begin{aligned} MEC_L^F &= MSC_L^F - c_L^F = \frac{\delta}{s}\left(N_L + N_H\right), \\ MEC_H^F &= MSC_H^F - c_H^F = \frac{\delta}{s}\left(\frac{\alpha_L}{\alpha_H} N_L + N_H\right). \end{aligned} \tag{8}$$

The MECs are constant over time. The difference in MECs depends on the ratio of values of time, and the difference increases with the degree of heterogeneity.

Maximizing welfare under user-equilibrium constraint (3) and a flat toll μ is equivalent to maximizing the following Lagrangian:

$$L^F = B_L + B_H - \left(c_H^F \cdot N_H + c_L^F \cdot N_L + k \cdot s\right) + \lambda_L^F \cdot \left(c_L^F + \mu - D_L\right) + \lambda_H^F \cdot \left(c_H^F + \mu - D_H\right). \tag{9}$$

The λ_i^F is the user-equilibrium multiplier for type i. It can be interpreted as the welfare change if we were to add a marginal toll for only type i. As we will see, type L is underpriced and type H is overpriced, so $\lambda_L^F > 0$ and $\lambda_H^F < 0$.

Lemma 3: Ratio heterogeneity and the flat toll

With ratio heterogeneity, the second-best flat toll does ***not*** *equal the average marginal external cost, as it is:*

$$\mu^F = MEC_L^F \cdot w_L^F + MEC_H^F \cdot (1 - w_L^F) = \frac{\delta}{s} N_H + \frac{\delta}{s} N_L \left(1 - \frac{-\frac{\partial D_L}{\partial N_L}\left(1 - \frac{\alpha_L}{\alpha_H}\right)}{-\frac{\partial D_L}{\partial N_L} - \frac{\partial D_H}{\partial N_H} + \frac{\delta}{s}\left(1 - \frac{\alpha_L}{\alpha_H}\right)}\right), \tag{10}$$

with the Low-VOT type's weight, w_L^F, depending on the derivatives of the demand and cost functions:

$$w_L^F = \frac{-\frac{\partial D_H}{\partial N_H} - \frac{\partial c_L}{\partial N_H} + \frac{\partial c_H}{\partial N_H}}{\left(-\frac{\partial D_H}{\partial N_H} - \frac{\partial c_L}{\partial N_H} + \frac{\partial c_H}{\partial N_H}\right) + \left(-\frac{\partial D_L}{\partial N_L} - \frac{\partial c_H}{\partial N_L} + \frac{\partial c_L}{\partial N_L}\right)} = \frac{-\frac{\partial D_H}{\partial N_H} + \frac{\delta}{s}\left(1 - \frac{\alpha_L}{\alpha_H}\right)}{-\frac{\partial D_L}{\partial N_L} - \frac{\partial D_H}{\partial N_H} + \frac{\delta}{s}\left(1 - \frac{\alpha_L}{\alpha_H}\right)}. \tag{11}$$

with the unique exception when the weights, w_i^F, happen to equal frequencies or shares, $f_i = N_i / \sum N_j$, of the types.

Proof of Lemma 3. See Appendix A.□

When α_L/α_H decreases, we have more diverse values of time. This does not affect the MEC_L but lowers the MEC_H and, thus, the average MEC. The optimal flat toll decreases less than the average MEC when the VOTs become more diverse, as this also raises the weight of the Low-VOT type in the toll rule and this type has the higher MEC. As shown in Appendix A, in optimum, $\lambda_L^F = -\lambda_H^F$ must hold. This implies that the optimal toll is set to balance the underpricing of the Low-VOT type with the overpricing of the High-VOT type. When $\partial D_i/\partial N_i$ is smaller in the absolute sense, type *i* is more price sensitive, and its weight in the coarse toll setting is larger, and the toll is closer to its MEC_i. If one type has a fixed demand, the coarse toll equals the MEC_j of the other type.

Proposition 1: Optimal capacity with a flat toll and two-type ratio heterogeneity
With a flat toll, the capacity rule has a second-best correction and follows:

$$-\sum N_i \cdot \partial c_i^F/\partial s = k - \lambda_L^F\left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right) \tag{12}$$

where $\lambda_L^F > 0$ *is the multiplier for the Low-VOT user equilibrium in (9). Since* $\lambda_L^F > 0$ *and* $\left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right) > 0$*, for a given number of users, the capacity with a flat toll is set higher than following the first-best capacity rule (which would be* $-\sum N_i \cdot \partial c_i^F/\partial s = k$*).*

Proposition 2: Self-financing with a flat toll and two-type ratio heterogeneity
With a second-best flat toll and capacity, the profit is

$$\begin{aligned}\Pi^F &= \left(\sum_i \left(\mu^F - MEC_i\right)N_i\right) - s\cdot\lambda_L^F\cdot\left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right)\\ &= \delta\frac{N_L}{s}\left(1-\frac{\alpha_L}{\alpha_H}\right)\left(N_L+N_H\right)\left\{\left(w_L^F - f_L\right) - \frac{\delta}{s}\left(1-\frac{\alpha_L}{\alpha_H}\right)\frac{1}{-\partial D_L/\partial N_L}\left(1-w_L^F\right)\left(1-f_L\right)\right\};\end{aligned} \tag{13}$$

where w_L^F *is the weight of the Low-VOT users in the toll rule of Proposition 1 and* $f_L=N_L/(N_L+N_H)$ *is their frequency. For there not to be a loss, the toll must exceed the average externality (weighted by frequency). For this to happen, the weight,* w_L^F *of Low-VOT type—with the high externality—must be well above its frequency,* f_L*.*

Proofs of Proposition 1 and 2. See Appendix A.□

Proposition 1 implies that the volume–capacity ratio with flat tolling is lower than with first-best fully time-variant tolling due to the addition of the second-best correction $\lambda_L^F\left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right)$ to the capacity rule. This lowers the welfare distortion from overpricing the High-VOT users—as they see a toll above their MEC_i—but raises the distortion from underpricing Low-VOT users. The second-best addition increases welfare as the first effect is stronger than the second. This second-best correction from Proposition 1 and the similar one for step tolling later in Proposition

3 are the core methodological contribution of our paper, as the existing literature has ignored this.

With ratio heterogeneity, it is impossible to analytically derive when there will be a profit and when a loss. In equation (13), $\sum_i \left(\mu^F - MEC_i\right) N_i$ gives how much the toll revenue deviates from the total external cost of $\sum_i MEC_i \cdot N_i$. The deviation is proportional to the term $\left(w_L^F - f_L\right)$ between the curly brackets in the second line of (13). So, if $w_L^F = f_L$, the toll would equal the average MEC and $\sum_i \left(\mu^F - MEC_i\right) N_i$ would be zero. However, then there would be a loss, as $s \cdot \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right)$ is positive.[15] Hence, with the second-best capacity, the flat toll must exceed the average MEC for there not to be a loss. So, for there not to be a loss, the Low-VOT type must be weighted more strongly in the toll setting than its frequency, $w_L^F > f_L$, and the flat toll must exceed the average MEC of $f_L \cdot MEC_L + (1 - f_L) \cdot MEC_H$. If Low-VOT users are more price sensitive, $\partial D_L / N_L$ is closer to zero and the profit increases (or loss decreases). When the High-VOT users are more price sensitive, the profit falls. The degree of ratio heterogeneity has an uncertain effect on the profit. It increases the second-best capacity, lowering profit. But it also increases the weight of the Low-VOT users, which raises the toll and thus the profit.

To conclude, under flat tolling and ratio heterogeneity, the MECs differ over types, and an anonymous coarse toll cannot equal the MECs. Moreover, the second-best capacity is set higher than following the first-best rule. The flat toll system will make a loss unless the Low-VOT type is much more price sensitive than the High-VOT type, and thus the toll well exceeds the average MEC. Accordingly, a loss is most likely to occur.

4.2. Single-step Laih toll

Now, we turn to step tolling, where the toll varies in steps. Van den Berg (2014) considers step tolling under continuous heterogeneity; we consider step tolling and capacity setting under discrete heterogeneity.

We only consider a single step in the toll and use the Laih (1994) equilibrium model, as this keeps the exposition clearer. Obvious extensions would be multiple steps and alternative equilibrium models. See Lindsey et al. (2012) for an overview of such models. The models differ in how they equate the generalized price before and after the toll is lowered. The Laih model attains this by having the users waiting beside the road just before the bottleneck without impeding other drivers.[16] The analysis would be more complicated in the other models as the

[15] The volume–capacity ratio is the total number of users, N_L+N_H, divided by the capacity s. For a given number of users, the second-best capacity with flat tolling exceeds the first-best capacity. Accordingly, the volume–capacity ratio is lower with the flat toll, and the difference increases with the degree of heterogeneity in the values of time.

[16] The ADL model has a mass departure instead. In the Braking model, people waiting for the toll to be lowered fully block the road. When α≥γ, the Laih and ADL models lead to the same equilibrium costs; under the more usual α<γ assumption, the ADL model has lower costs. The braking always leads to higher costs, as the blockage of the road means that the bottleneck capacity goes unused for some time. Ren et al. (2016) developed a model that is in between the Laih and Braking models: some users stop driving before the toll is lowered, but instead of blocking the road, other drivers can pass them but are hindered.

exact distribution of user types over time is uncertain. Nevertheless, in these models, the results would probably be similar since costs would be similar and depend similarly on the preferences, capacity and numbers of users of each type.

The toll is higher in the center of the peak; outside this period, the toll is lower. In the center period, of each type i, $V_i \geq 0$ users travel and the toll equals $\mu+\rho$. In the shoulder periods, there are $N_i - V_i \geq 0$ users and the toll is μ. Hence, μ is the flat part of the toll, and ρ is the step part:

$$\tau[t] = \begin{cases} \mu+\rho & \text{if } t^{+} \leq t \leq t^{-} \\ \mu & otherwise \end{cases}.$$

With single-step tolling, travel costs and MECs are constant within a period but differ between periods. Travel costs are lower in the center peak period, whereas MECs are higher. The generalized price for a type of $c_i[t]+\tau[t]$ is the same in the center and shoulder periods.

The peak starts at t_s when the first arrival occurs and ends at t_e when the last arrival occurs. The early shoulder period lasts from t_s until t^+ when the toll is increased. The late shoulder period lasts from t^- to t_e.

We will summarize the results in text as they are similar to those in Van den Berg (2014) for continuous heterogeneity. Appendix B gives a detailed derivation. To easily check the meaning of the notation, please see the nomenclature box in Appendix D.

High-VOT users are less willing to queue since they value travel time more than schedule delay. Hence, both in the shoulder periods and in the center period, the High-VOT users will arrive further from t^* when travel times are lower, but schedule delays are higher. So, self-separation over time occurs. Of each type, a fraction $\gamma/(\gamma+\beta)$ arrives early and the remainder late. This is true both in the center and in the shoulder periods. The optimal step part of the toll minimizes the total travel cost for a given N_L and N_H. To do so, its level, ρ, equates the generalized prices in the center and shoulder periods, whilst its timings are such that the queue reaches zero size at t^+ and t^-.

Total cost is minimized when, of each type, half the users travel in the center period: $V_i=N_i/2$. This allows us to write the total cost as

$$TC^{SS} = \frac{3}{4}\frac{\delta}{s}(N_L+N_H)(N_L+N_H) - \frac{\delta N_L N_H}{2\,s}\left(1-\frac{\alpha_L}{\alpha_H}\right) + k\,s, \tag{14}$$

where superscript ^{SS} indicates the single-step toll equilibrium. For given numbers of users, total cost is lower now than with flat tolling or no tolling, but the decrease is smaller than with homogeneity. The costs per trip simplify to

$$c_L^{SS} = \begin{cases} c_L^{cp} = \frac{1}{2}\delta\frac{N_L + \frac{\alpha_L}{\alpha_H}\cdot N_H}{s} & \text{if } t^{+} \leq t \leq t^{-} \\ c_L^{sh} = \frac{1}{2}\delta\frac{N_L + \frac{\alpha_L}{\alpha_H}\cdot N_H}{s} + \frac{1}{2}\delta\frac{N_L+N_H}{s} & \text{if } t_s \leq t < t^{+} \text{ or } t^{-} < t \leq t_e \end{cases} \tag{15a}$$

$$c_H^{SS} = \begin{cases} c_H^{cp} = \frac{1}{2}\ \delta\frac{N_H + N_L}{s} & \text{if } t^+ \le t \le t^- \\ c_H^{sh} = \delta\frac{N_H + N_L}{s} & \text{if } t_s \le t < t^+ \text{ or } t^- < t \le t_e. \end{cases} \tag{15b}$$

Here, superscript ^{cp} indicates the center peak and ^{sh} the shoulder periods.

The MEC_L of the Low-VOT users exceeds the MEC_H of the High-VOT users in each period. The differences in MECs between the center and shoulder periods are the same for both types:

$$MEC_L^{sh} = MSC_L \ - c_L^{sh} = \frac{\delta}{s}\frac{N_L+N_H}{2}, \tag{16a}$$

$$MEC_H^{sh} = MSC_H \ - c_H^{sh} = \frac{\delta}{s}\frac{\frac{\alpha_L}{\alpha_H}N_L+N_H}{2}, \tag{16b}$$

$$MEC_i^{cp} = MSC_i \ - c_i^{cp} = MEC_i^{sh} + \frac{\delta}{s}\frac{N_L+N_H}{2}. \tag{16c}$$

In user equilibrium, the ρ equals the difference in cost between the center peak and shoulder period, and this turns out to also be the difference in MEC between these periods:

$$\rho = \frac{1}{2}\ \delta\frac{N_L + N_H}{s} = MEC_i^{cp} - MEC_i^{sh}. \tag{17}$$

We can show that maximizing welfare is equivalent to maximizing the below Lagrangian:

$$L^{SS} = B_L + B_H - TC^{SS} + \lambda_L^{sh} \cdot \left(c_L^{sh} + \mu - D_L\right) + \lambda_H^{sh} \cdot \left(c_H^{sh} + \mu - D_H\right). \tag{18}$$

The user-equilibrium multiplier, λ_i^{sh}, ensures that type i's generalized price equals its inverse demand, D_i. Problem (18) is akin to the earlier problem (9) for the flat toll, as we already optimized V_L, V_H and ρ. The difference is that the step toll, compared to the flat toll, leads to lower costs by halving the queuing times for the same N_L and N_H.[17]

Lemma 4: Step toll and ratio heterogeneity

Under a Laih single-step toll and two-type ratio heterogeneity, within each period, the marginal external cost (MEC_L) of the Low-VOT users exceeds that of the High-VOT users. The step part of the toll, ρ, equals the difference in MECs between the center peak and shoulder periods. This difference is the same for both types. The flat part of the toll, μ, balances the underpricing of Low-VOT users with the overpricing of High-VOT users:

$$\mu^{SS} = MEC_L^{sh} \cdot w_L^{SS} + MEC_H^{sh} \cdot \left(1 - w_H^{SS}\right) = \frac{\delta}{2\,s}N_H + \frac{\delta}{2\,s}N_L\left(\frac{\alpha_L}{\alpha_H} + \left(1 - \frac{\alpha_L}{\alpha_H}\right)\frac{-\frac{\partial D_L}{\partial N_L}+\frac{\delta}{2\,s}\left(1-\frac{\alpha_L}{\alpha_H}\right)}{-\frac{\partial D_L}{\partial N_L}-\frac{\partial D_H}{\partial N_H}+\frac{\delta}{2\,s}\left(1-\frac{\alpha_L}{\alpha_H}\right)}\right) \tag{19}$$

where the weight of type L is:

$$w_L^{SS} = \frac{\left(-\frac{\partial D_H}{\partial N_H} - \frac{\partial c_L^{cp}}{\partial N_H} + \frac{\partial c_H^{cp}}{\partial N_H}\right)}{\left(-\frac{\partial D_H}{\partial N_H} - \frac{\partial c_L^{cp}}{\partial N_H} + \frac{\partial c_H^{cp}}{\partial N_H}\right) + \left(-\frac{\partial D_L}{\partial N_L} - \frac{\partial c_H^{cp}}{\partial N_L} + \frac{\partial c_L^{cp}}{\partial N_L}\right)} = \frac{-\frac{\partial D_H}{\partial N_H} + \frac{\delta}{2 \cdot s}\left(1 - \frac{\alpha_L}{\alpha_H}\right)}{-\frac{\partial D_L}{\partial N_L} - \frac{\partial D_H}{\partial N_H} + \frac{\delta}{2 \cdot s}\left(1 - \frac{\alpha_L}{\alpha_H}\right)}. \tag{20}$$

[17] We get the same analytical results if we solve the full problem in one go. Moreover, for the numerical model, we directly maximized welfare to s, N_L, N_H, V_L, V_H, ρ, and μ with user-equilibrium constraints for both the shoulder and peak periods. However, the presented two-step optimization method is easier to follow.

As with the flat toll, the step toll balances the overpricing of the High-VOT users and the Low-VOT users' underpricing. The toll generally differs from the average *MEC* (weighted by the number of users of each type). The deviation from the average MEC will tend to be smaller than with the flat toll because travel costs are lower. The more price sensitive a type *i* is, the closer the toll is to its *MEC*$_i$.

Proposition 3: Optimal capacity with a step toll

With two-type ratio heterogeneity and a single-step Laih toll, the capacity rule has a second-best correction and follows:

$$-\sum N_i \cdot \partial E[c_i]/\partial s = k - \lambda_L^{sh} \left(\frac{\partial c_L^{sh}}{\partial s} - \frac{\partial c_H^{sh}}{\partial s} \right), \qquad (21)$$

where E[c$_i$*] is the average travel cost for type i*[18] *in user equilibrium and* $\lambda_L^{sh} > 0$ *is the Low-VOT type's multiplier in (18). For given numbers of users, the capacity is set higher than following the first-best rule but lower than with a flat toll.*

Proposition 4: Self-financing with a step toll

Under two-type ratio heterogeneity, the profit of a single-step Laih toll is:

$$\begin{aligned}\Pi^{SS} &= \sum_i \left(\mu^{SS} - MEC_i^{sh} \right) N_i - s \cdot \lambda_L^{sh} \cdot \left(\frac{\partial c_L^{sh}}{\partial s} - \frac{\partial c_H^{sh}}{\partial s} \right) \\ &= \delta \frac{N_L}{2 \cdot s} \left(1 - \frac{\alpha_L}{\alpha_H} \right) \left(N_L + N_H \right) \left\{ \left(w_L^{SS} - f_L \right) - \frac{\delta}{2 \cdot s} \left(1 - \frac{\alpha_L}{\alpha_H} \right) \frac{1}{-\partial D_L / \partial N_L} \left(1 - w_L^{SS} \right) \left(1 - f_L \right) \right\}\end{aligned}, \qquad (22)$$

where λ_L^{sh}*>0 is the Low-VOT type's multiplier from (18),* w_L^{SS} *is its weight in the toll rule from (20) and f*$_L$*=N*$_L$*/(N*$_L$*+N*$_H$*).* [19]

Proofs of Lemma 4 and Propositions 3 and 4. Since, after we have already optimized the step part of the toll (i.e., V_L, V_H and ρ), problem (18) is so similar to problem (13) for the flat toll, we will skip these proofs as they are almost identical to those seen before. The only difference is that we now have slightly different cost functions. □

4.3. Comparing the profit under flat and single-step tolling

The ratio of flat-toll to step-toll profit or loss is

$$\frac{\Pi^F}{\Pi^{SS}} = 2 \cdot \frac{\dfrac{N_L^F + N_H^F}{s^F}}{\dfrac{N_L^{SS} + N_H^{SS}}{s^{SS}}} \cdot \frac{N_L^F \cdot \left(w_L^F - f_L^F \right) - \lambda_L^F \cdot \left(1 - f_L^F \right)}{N_L^{SS} \cdot \left(w_L^{SS} - f_L^{SS} \right) - \lambda_L^{SS} \cdot \left(1 - f_L^{SS} \right)}, \qquad (23)$$

where we use superscripts ^{F} and ^{SS} to indicate the schemes for clarity. Accordingly—if the weights, capacities, multipliers and number of users were the same in both cases—the profit or

[18] E[c_i]=$(c_i^{sh} \cdot (N_i - V_i) + c_i^{cp} \cdot V_i)/(N_i)$, such that total cost is $TC^{SS} = \mathrm{E}[c_L] \cdot N_L + \mathrm{E}[c_H] \cdot N_H + k \cdot s$.

[19] Again, the volume–capacity ratio, (N_L+N_H)/*s*, is smaller than with the first-best rule, but it will be larger than with the flat toll as the second-best capacity correction is smaller.

loss would be twice as large with the flat toll (F) as with the step toll (SS). However, we will see that the difference tends to be smaller. The second term in (23) will be somewhat smaller than 1 because the volume–capacity ratio, $\left(N_L^j + N_H^j\right)/s^j$, will be higher with the step toll due to the smaller second-best capacity correction with step tolling compared to flat tolling. The third term will tend to be close to 1.[20] Hence, the profit or loss with a step toll will tend to be somewhat more than half that of the flat toll.

Equation (23) also indicates that the two tolling schemes attain a zero profit at similar combinations of parameters but not identical ones. For both, the profit is larger (or loss smaller) if the Low-VOT users are more price sensitive—meaning that the derivative $\partial D_L/\partial N_L$ is closer to zero—and when the High-VOT users are less price sensitive. The profit or loss goes towards zero as the degree of heterogeneity goes to zero (i.e., as α_L approaches α_H). As the step toll has the smaller second-best capacity expansion, it is more likely not to have a loss.

Finally, for flat and step tolling, a deficit seems most likely. The extra capacity due to the second-best capacity rule is expensive, so the second-best toll would need to be much higher than the average externality for there not to be a loss.

4.4. Conclusions on the analytical ratio heterogeneity model

This section studied second-best flat or step tolling and capacity setting under ratio heterogeneity in the ratio of the value of time and schedule delay. This heterogeneity has also been called ‘flexibility heterogeneity.’ With a flat toll, the toll is constant throughout the peak. With a step toll, the toll is μ in the early and late parts of the peak; in the center peak, it is higher and is $\mu + \rho$.

Both the flat toll and the step toll are a weighted average of the marginal external costs (MECs) of the types, with weights depending on demand and travel cost derivatives and not directly on the number of users of a type.

For given numbers of users, the capacity is set higher with the flat toll than with the step toll, which in turn has a higher capacity than under the first-best fully time-variant toll. This limits the welfare reduction due to overpricing of High-VOT users, where this issue is more severe with the flat toll.

The step and flat tolling attain a zero profit at similar combinations of parameters but not identical ones. For both, the profit is larger (or loss smaller) if the Low-VOT users are more price sensitive or the High-VOT users are less price sensitive. The profit or loss goes towards zero as

[20] This follows from three points.

Point 1, N_L^F / s^F *vs* N_L^{SS} / s^{SS} is uncertain, since $\left(N_L^F + N_H^F\right)/s^F < \left(N_L^{SS} + N_H^{SS}\right)/s^{SS}$ and $f_L^F > f_L^{SS}$ as step tolling is relatively more detrimental for Low-VOT users. But on the whole, the two ratios will be similar.

Point 2. $N_H^F / s^F < N_H^{SS} / s^{SS}$, as the step toll will tend to have the higher volume–capacity ratio, $\left(N_L^F + N_H^F\right)/s^F < \left(N_L^{SS} + N_H^{SS}\right)/s^{SS}$, and $f_L^F > f_L^{SS}$. However, $\lambda_L^F > \lambda_L^{SS}$ as the flat toll underprices low-VOT users more. Therefore, these two effects work in opposite directions and will mostly cancel each other out.

Point 3. Finally, $w_L^F > w_L^{SS}$ and $f_L^F > f_L^{SS}$, so again these two effects on the relative profit work in opposite directions and will mostly cancel each other out.

the degree of heterogeneity goes to zero. As the step toll has a smaller second-best capacity expansion than the flat toll, it is less likely to make a loss. The flat toll's profit or loss tends to be almost twice that of the step toll.

5. Numerical model for ratio heterogeneity

Now, we turn to our numerical model. This will illustrate our analytical model. It also studies the effects of imposing self-financing, which we cannot do with pure analytics. The sensitivity analyses study four things. (i) How likely is it that there will be a loss? (ii) How significant is the potential lack of self-financing? (iii) How sensitive is the outcome to parameter values? (iv) What are the effect of imposing self-financing by adding an extra flat charge to the toll?

For comparison and calibration, we also look at the outcomes without tolling and with fully time-variant tolling. This allows us to put the effects of flat and step tolling into perspective.

5.1 Base case calibration

We aim to keep the setup comparable with Van den Berg and Verhoef (2011a, 2011b) and Van den Berg (2014). The numerical model thus also considers fuel costs of €7.30 and a free-flow travel time of 30 minutes, but these other travel costs are not included in the results in Table 1. We use the following linear inverse demand that is type-specific:

$$D_i = d0_i - d1_i \; N_i$$

The demand parameters $d0_i$ and $d1_i$ are such that the no-toll equilibrium has 6000 Low-VOT users, 3000 High-VOT users, and an average fuel-cost elasticity of 0.4. This elasticity is close to the average in Brons et al. (2008).[21] The marginal capacity cost parameter, k, is set such that the first-best capacity is s=3600, which is the fixed capacity in the earlier works. In the no-toll equilibrium, the capacity is also assumed to be 3600. The VOTs are α_L=€7.50/h and α_H=€15.00/h. This ensures that, in the no-toll equilibrium, the average value of time is €10.00/hour. This figure is close to the official Dutch average (Kouwenhoven et al., 2014). The values of schedule delay are β=€6.09/h and γ=€23.76/h, and they follow from the ratios of the value of time to values of schedule delay in Small (1982) and Arnott et al. (1993). Most of the bottleneck literature has used these ratios.

Finally, we assume that the slope of the inverse demand of the Low-VOT type is 0.7 that of the High-VOT type, implying that the Low-VOT type is more price sensitive. The sensitivity analysis will examine the effects of changing this ratio of demand slopes.[22]

[21] We use fuel cost for the elasticity as there is a large body of empirical literature on this, while little is known for toll payments. We assume that people are equally sensitive to all cost components.

[22] The remaining parameters are k=14.103, $d0_L$=34.4378, $d0_H$=29.5063, $d1_L$=0.00405556, $d1_H$=0.00579365 and $\delta=\beta\cdot\gamma/(\beta+\gamma)$= 4.85013. That $d0_L>d0_H$ is a result of our desire that the Low-VOT type is much more numerous and the continuous linear demand. If one were concerned that this means that the Low-VOT type has the higher maximum willingness-to-pay for the trip, one could cut off their demand function at, say, 15 without changing anything except lowering their consumer surplus in all settings by the same amount.

5.2 Results for the base calibration

Table 1 gives the numerical results. The most important result is that both coarse toll schemes have a loss. The flat toll has a loss of 127 or 3.2% of capacity cost; the step toll has a loss of 66 or 1.7%. This is consistent with the analytical section, which discussed that the ratio of profit or loss with flat tolling tends to be almost twice that with step tolling. With the first-best toll, the toll equals the time-variant MEC, and there is zero profit. With the flat toll, the average MEC is €7.78, and the toll is €7.58. So, the flat toll is below the average MEC, while it would need to exceed it for a zero profit. The flat toll is €1.15 lower than the Low-VOT type's MEC_L and €1.81 higher than the MEC_H. With the single-step toll, the toll is below the average *MEC* and the MEC_L, and the toll exceeds the MEC_H. However, the differences are smaller as travel costs and MECs are lower with the step toll.

Imposing self-financing does little harm to welfare. When adding the constraint that the flat toll revenue has to equal capacity cost, the welfare is only 0.02% lower, while the generalized prices increase by about 1.5%. With a step toll, the effect of self-financing is even smaller: it causes a 0.004% fall in welfare.

The flat toll is a blunter instrument that can limit the number of users without affecting queuing delays. With fixed demand, it would have a welfare gain of zero. Now, it attains 24.9% of the first-best welfare gain relative to the no-toll case, but both types face substantial price increases.[23] For given numbers of users, the single-step toll removes half the queuing compared to the no-toll case. It is a more precise instrument and is less harmful to consumers. Nevertheless, both types see an increase in generalized price, but the High-VOT type sees a smaller one. The single-step toll attains 58.8% of the first-best gain in our numerical base case. With fixed demand, this percentage would be 50%, both with homogeneity and with ratio heterogeneity (Laih, 2004; Van den Berg, 2014).

[23] For a given numbers of users, the price with a flat toll would be twice that of the no-toll case as the MSC is double the private travel cost; but, of course, the flat toll also reduces the number of users.

Table 1: Results for the different policies under the base case calibration

	No-toll	**First-best**	**Flat toll**	**Self-financing flat toll**	**Step toll**			**Self-financing step toll**		
					Center period	Shoulder period	Combined	Center period	Shoulder period	Combined
Number of Low-VOT users, N_L	6000	5607.5	4814.9	4751.1	2596.1	2596.1	5192.2	2581.3	2581.3	5162.6
Number of High-VOT users, N_H	3000	3074.1	2277.3	2233.7	1319.6	1319.6	2639.2	1309.3	1309.3	2618.6
Travel cost for the Low-VOT type [d]	10.10	5.85[a]	7.33	7.33	4.12	9.08	6.60	4.12	9.08	6.60
Travel cost for the High-VOT type [d]	12.13	5.85[a]	8.73	8.72	4.96	9.91	7.44	4.96	9.91	7.43
(Average) MEC_{L} [f]	**12.13**	**5.85**[a]	**8.73**	**8.72**	**9.91**	**4.95**	**7.44**[a]	**9.91**	**4.96**	**7.44**[a]
(Average) MEC_H [f]	**8.08**	**5.85**[a]	**5.77**	**5.76**	**8.27**	**3.31**	**5.79**[a]	**8.31**	**3.27**	**5.79**[a]
(Average) toll	**x**	**5.85**[a]	**7.58**	**7.84**	**9.26**	**4.30**	**6.78**	**9.38**	**4.42**	**6.90**
Step part of the toll	x	x	x	x	4.96			4.96		
Flat part of the toll	x	0	7.58	7.83	4.2			4.42		
Toll revenue	x	50771	53773	54741	53098			53687		
Capacity cost of $k{\cdot}s$	50771	50771	55566	54741	54030			53687		
Profit	**x**	**0**	**-1794**	**0**	**-932**			**0**		
Profit as a percentage of capacity cost	**x**	**0%**	**-3.23%**	**0%**	**-1.73%**			**0%**		
Capacity, s	3600[c]	3600	3940	3881	3831			3807		
$(N_L+N_H)/s$: volume-capacity ratio	2.50	2.41	1.80	1.80	2.04			2.04		
Total travel cost [e]	97003	50771	55169	54332	68969			68528		
Welfare, W	48301	91136	60240	60226	73910			73907		
Relative efficiency [b]	**0**[b]	**1**[b]	**0.249**[b]	**0.248**[b]	**0.5979**[b]			**0.5978**[b]		

Notes: [a] This is an average over time.

[b] The relative efficiency of a policy is its welfare gain from the no-toll equilibrium divided by the corresponding welfare gain of the first-best social optimum.

[c] The NT capacity is assumed to be 3600 in the no-toll equilibrium, whereas, for the other cases, the capacity is set at the optimized level. For the no-toll case, one could also optimize the capacity. The second-best capacity would be much higher than in the first-best case, but using this capacity would complicate the comparison with previous papers that used a fixed capacity. The second-best capacity is higher because the no-toll case has more users and mostly because, for a given N_L and N_H, the no-tolling user costs are almost twice the user cost in the first-best case, making capacity building more attractive. Finally, the no-toll case would have a second-best capacity reduction that corrects for latent demand due to the unpriced congestion (see also Small and Verhoef (2007) and de Palma and Lindsey (2007)). The flat and step-toll settings have a second-best capacity increase compared to the first-best rule, and thus they have a lower volume–capacity ratio.

[d] This excludes the costs only added in the numerical model from fuel, free-flow travel time and operation costs.

[e] Total travel cost=$\sum N_i \cdot c_i$.

[f] $\mathrm{MEC_L}$: Marginal external cost of the Low-VOT type, $\mathrm{MEC_H}$: Marginal external cost of the High-VOT type,

5.3 Sensitivity analyses

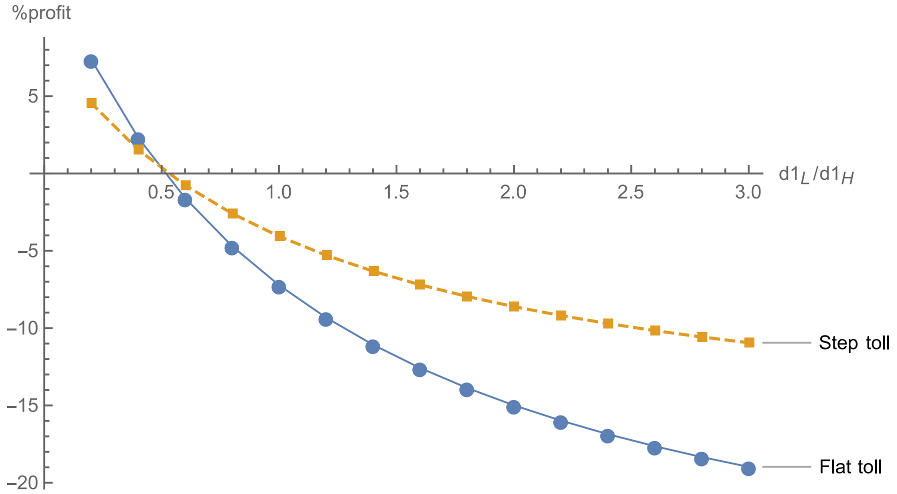
We now turn to our sensitivity analyses. The most important one is the sensitivity to the relative demand sensitivities of the two types, $d1_L$ vs $d1_H$, which allows us to see when the system is self-financing and how likely a non-negative profit is. We also look at the degree of ratio heterogeneity and the average elasticity to the fuel cost.

5.3.1. Difference in the slopes of the demand functions of the two types: $d1_L$ vs $d1_H$

We start by looking at the ratio of demands $d1_L/d1_H$. In the base calibration, this ratio was 0.7, and so the Low-VOT type was more price sensitive than the High-VOT type. Fig. 1 shows that the profit increases for both schemes as the Low-VOT users become relatively more price sensitive. For both of them, the Low-VOT users would need to be about twice as price sensitive as the High-VOT users for there not to be a loss. As we argued using equation (23), the flat toll's profit or loss is almost twice that of the step toll. Although this is not clearly visible, the step toll is more likely not to make a loss: its curve intersects the zero-profit x-axis a tiny bit more to the left. Most importantly, when $d1_L/d1_H$ is between 0.5 and 1—which seems the most reasonable range—the deficit for the flat toll is between 0% and 7% and for the step toll between 0% and 4%.

Now, we turn to the welfare effects in the left panel of Fig. 2. The step toll has a relative efficiency of around 0.6. It thus attains about 60% of the welfare gain from the no-toll case that the first-best policy gives. The relative efficiency for the flat toll is between 0.25 and 0.30. As the Low-VOT type becomes relatively less price sensitive, the relative efficiency of both policies decreases slightly. This probably occurs because neither policy removes all queueing while both partly reduce congestion by reducing consumption, which is more difficult when the Low-VOT type—which is more numerous—is less price sensitive.

Fig. 1. Profit as a percentage of capacity cost over the ratio, $d1_L/d1_H$, of the demand slopes

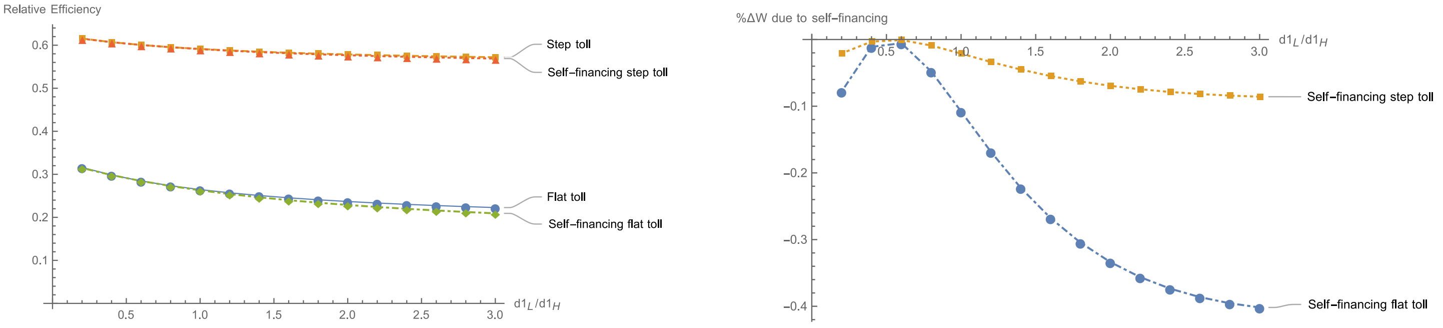


Note: The %profit is the profit as a percentage of the capacity cost. The $d1_L/d1_H$ is the ratio of the Low-VOT type's demand slope and the corresponding slope for the High-VOT type.

Fig. 2(right) looks at the effects on welfare of imposing self-financing, assuming that a positive profit is also impossible (for instance, because toll revenue is earmarked to be used on roads). So, if there would be a loss, the flat part of the toll is raised to obtain self-financing; conversely, if there would be a profit, it is lowered. The welfare loss of imposing self-financing is a minute 0% to 0.4% for the flat toll and 0% to 0.1% for the step toll. This suggests that we can attain self-financing with little harm to society. Self-financing has a more noticeable effect on prices, the number of users, and consumer surplus. For instance, the number of users can fall by up to 10%.

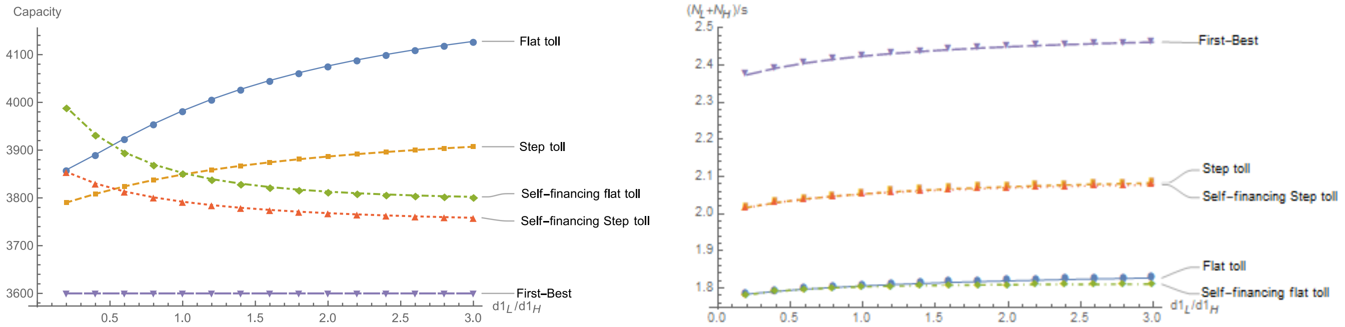
Finally, Fig. 3 looks at the capacity and the volume–capacity ratio. As the analytics showed, the flat toll has a lower-volume capacity ratio, $(N_H+N_L)/s$, than the step toll, which in turn has a lower ratio than the first-best case. The first-best toll removes all queuing, and step tolling halves the queueing for given usage numbers. So, the travel costs are much higher with the flat toll than with the step toll, which in turn has higher travel costs than the first-best toll. Higher travel costs are a second reason for more capacity and, thus, a lower volume-capacity ratio.

Fig. 2. Welfare effects as the ratio $d1_L/d1_H$ of demand slopes changes: Relative efficiency (left panel), welfare loss due to imposing self-financing (right panel)



Note: The relative efficiency of a policy is its welfare gain from the no-toll equilibrium divided by the corresponding welfare gain of the first-best social optimum. The $d1_L/d1_H$ is the ratio of the Low-VOT type's demand slope and the corresponding slope for the High-VOT type.

Fig. 3. The effect of the ratio of demand slopes on capacity (left panel) and volume-capacity ratio (right panel)

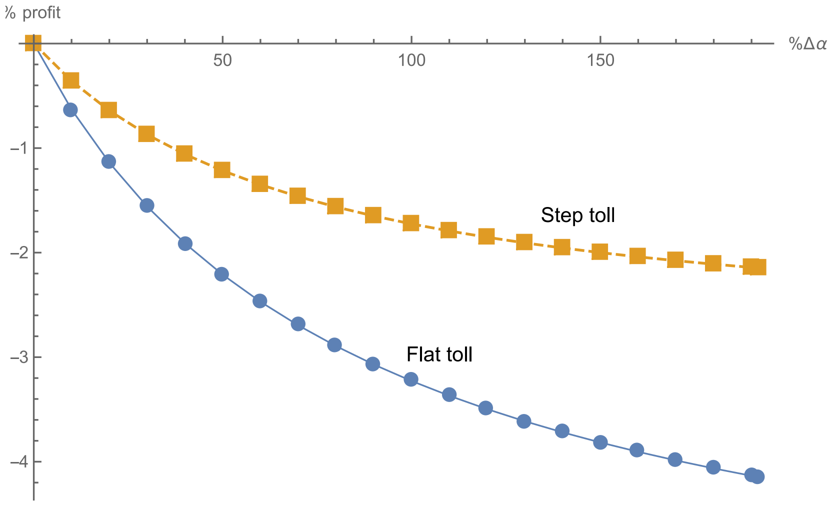


Note: The $(N_L+N_H)/s$ is the volume-capacity ratio as it is the ratio of the total number of users (i.e. the volume) to the capacity s. The $d1_L/d1_H$ is the ratio of the Low-VOT type's demand slope and the corresponding slope for the High-VOT type.

5.3.2. Degree of heterogeneity in the value of time

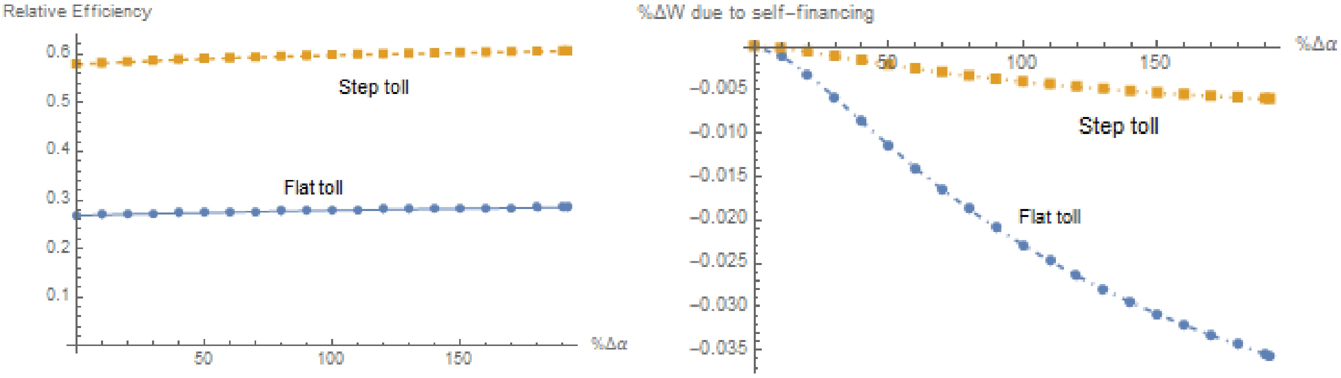
Now we turn to the degree of ratio heterogeneity, which we measure by the percentage difference in values of time: $\%\Delta\alpha = 100\% \cdot (\alpha_H / \alpha_L - 1)$. With homogeneity, the coarse toll equals the homogeneous *MEC* and, as Fig. 4 shows, the profit is exactly zero. As the degree of heterogeneity rises, the *MEC* becomes more heterogeneous and the loss convexly increases. The loss ranges between 0% for homogeneous users and 4% for the largest degree of heterogeneity.

Fig. 4. The effect of the degree of heterogeneity (%Δα) on profit (as a percentage of the capacity cost).



Note: The %profit is the profit as a percentage of the capacity cost. The $\%\Delta\alpha = 100\% \cdot (\alpha_H / \alpha_L - 1)$ is the percentage difference between the two values of time.

Fig. 5. Welfare effects and the degree of heterogeneity (%Δα).



Note: The relative efficiency of a policy is its welfare gain from the no-toll equilibrium divided by the corresponding welfare gain of the first-best social optimum. The $\%\Delta\alpha = 100\% \cdot (\alpha_H / \alpha_L - 1)$ is the percentage difference between the two values of time.

As Fig. 5 illustrates, the degree of heterogeneity has little to no effect on the welfare effect of the step toll. With fixed demand, the single-step toll always attains 50% of the first-best gain. Here, the step toll's gain is slightly higher as, with price-sensitive demand, it also reduces overconsumption caused by the externality. For the flat toll, the relative efficiency is much lower and falls with the degree of heterogeneity. This is primarily a scale effect as no-toll welfare increases with this degree by lowering the no-toll travel cost for the Low-VOT type.

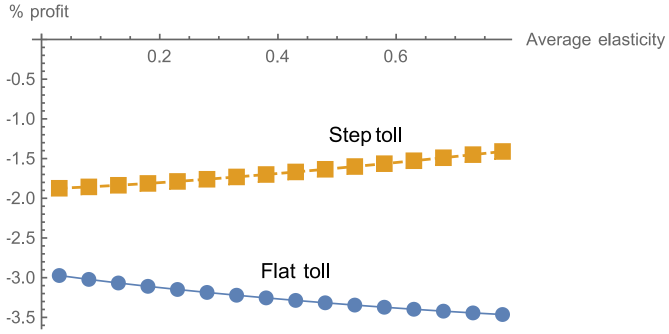
Since the loss with a coarse toll increases with the %Δα, the welfare loss from imposing self-financing increases with %Δ*α*. Still, for the maximum degree of heterogeneity at which the α_L is only just above *β*,[24] the welfare loss is only 0.035%.

[24] Which must be so for the regular equilibria to hold in the bottleneck model or any other dynamic congestion model (Arnott et al., 1993a).

5.3.3. Average elasticity

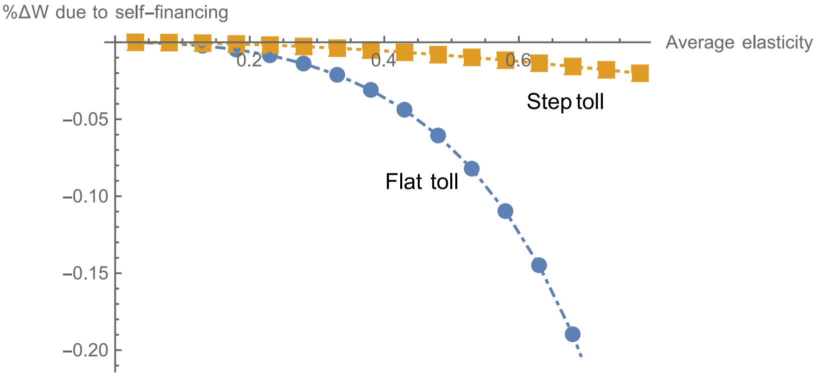
Finally, we look at the average elasticity to the fuel cost. We use the fuel cost elasticity, as there is a large body of empirical literature on this, while less is known for toll payments.[25] Fig. 6 indicates that as demand becomes more elastic, the percentage loss becomes a bit larger for the flat toll and a bit lower for the step toll. Fig. 7 shows that imposing self-financing does little harm to welfare. A more elastic overall demand means this percentage welfare loss becomes slightly larger.

Fig. 6. The effect of the average fuel cost elasticity on profit (as a percentage of the capacity cost)



Note: We use the absolute value of the fuel-price elasticity so that a larger number implies a more elastic demand.

Fig. 7. Welfare effects of imposing self-financing over the average fuel cost elasticity



This concludes our sensitivity analysis. Appendix C gives some further analyses, such as the effects of changing the marginal capacity cost. To summarize, it is most likely that a coarse toll will make a loss. Low-VOT users need to be much more price sensitive than the High-VOT users for the coarse toll to be far enough above the average MEC and, thus, to be able to cover the capacity cost. As overall demand becomes more price sensitive or the degree of ratio heterogeneity increases, the deficit with a coarse toll rises, and the welfare decrease due to imposing self-financing also rises.

6. Proportional heterogeneity

Ratio heterogeneity seems to be the most interesting heterogeneity, as it means that the self-financing theorem does not hold, and the toll and capacity rules need second-best adjustments. With proportional heterogeneity, the values of time and schedule delay vary over types in fixed

[25] To do so, we adjust the two demand slopes and the two demand intercepts so as to keep the total numbers of users of the two types and the relative size of the demand slopes, $d1_L/d1_H$, constant in the no-toll equilibrium. Just altering the demand slopes would also directly change the numbers of users, making for an unfair comparison of the effect of the price sensitivity.

proportions. This means $\alpha_i=\mu\cdot\beta_i$ and $\gamma_i=\eta\cdot\beta_i$, where scalars μ and η are the same for all. The preferred arrival time is assumed to be homogeneous in this section. As Van den Berg (2011a) argued, this heterogeneity could stem from income differences. This section studies proportional heterogeneity with discrete user types, where all values are the same percentage higher for a type with higher values. Again, there are separate demands for each type. To check the meaning of notation, see the nomenclature box in Appendix D.

Solely proportional heterogeneity implies that the marginal external cost (MEC) is independent of the type. But, for step tolling, the MEC does vary over arrival time windows: the MEC is lower in the center peak than in the shoulders. Conversely, with flat tolling or no tolling, the MEC is also constant over arrival times.

6.1 Flat toll

Again, we first turn to the flat toll but now consider the more general case with *M* discrete types. The flat toll cannot remove queueing; it can only remove the persons from the road who have a value of the trip below the marginal social cost.

Proposition 5: Proportional heterogeneity and flat tolling

With proportional heterogeneity, the marginal external costs (MECs) are the same for both user types. So, the optimal toll equals the MEC throughout the peak, μ=MEC, and the system is self-financing.

Proof: We consider *M* discrete type. As all types have the same ratios of the value of time to values of schedule delay, all types travel jointly, and each type can travel at any moment between t_s and t_e and be in user equilibrium. Conversely, with ratio heterogeneity, as we saw before, types travel separated over time: the higher the ratio of a type, the closer it travels to the preferred arrival time. With proportional heterogeneity, the total cost with *M* types of users is:

$$TC=\sum_{i=1}^{M}c_i\cdot N_i=\sum_{i=1}^{M}\delta_i\frac{N}{s}\cdot N_i,\ \text{with}\ c_i=\delta_i\frac{\sum_{j=1}^{M}N_j}{s}\ \text{and}\ \delta_i=(\beta_i\cdot\gamma_i)/(\beta_i+\gamma_i).$$

This total cost is very similar to under homogeneity (e.g., Arnott et al., 1993), only now δ_i varies over the types *i* and we sum over all types.

Clearly, $\frac{\partial c_i}{\partial N_j}=\frac{\delta_i}{s}$, for any type *j*. So, the MEC_j must be the same for all types as the terms in the summation are independent of what type *j* is:

$$MEC_j=\sum_{i=1}^{M}\frac{\partial c_i}{\partial N_j}\cdot N_i=\sum_{i=1}^{M}\frac{\delta_i}{s}\cdot N_i.\square$$

As the toll equals the homogeneous MEC with proportional heterogeneity, Lemma 1 tells us that the capacity will follow the first-best rule and minimize total cost. Lemma 2 implies that the flat toll will have toll revenue equal to capacity costs.

6.2. Laih single-step tolling under two-type proportional heterogeneity

The Laih (1994, 2004) single-step toll under proportional heterogeneity was first studied by Xiao et al. (2011) and Van den Berg (2014). We now consider two types of users that differ in their δ_i. The analysis is more difficult as there are three possible equilibria depending on the relative sizes of the high- and low-values types. However, these equilibria only differ in whether they are fully or partially separated. We assume N_H>0 and N_L>0, as otherwise there would be no heterogeneity.

The toll again is μ in the shoulders of the peak and is $\mu+\rho$ in the center peak. Of each type, $V_i \geq 0$ users travel in the center peak and the remaining $N_i - V_i$ in the shoulder periods. In the fully separated equilibrium, all users with high values of time and schedule delay travel in the center peak and all those with low values in the shoulder periods. In the two partially separated equilibria, one type travels in both the center and shoulder periods, while the other type uses only one. The three possible user equilibria can be summarized as:

i. Partially separated equilibrium with $N_L>V_L>0$ and $V_H=N_H$. This occurs if $\delta_H N_H < \delta_L N_L$.

ii. Fully separated equilibrium with $V_L=0$ and $V_H=N_H$. This occurs if $\delta_H N_H \geq \delta_L N_L$ &$N_L \geq N_H$.

iii. Partially separated with $V_L=0$ *and* $N_H>V_H>0$. This occurs if $\delta_H N_H \geq \delta_L H_L$ & $N_H > N_L$.

Proposition 6: Self-financing of step tolling under proportional heterogeneity

With two-type proportional heterogeneity, for all three possible user equilibria, the toll $\tau[t]$ equals the $MEC_i[t]$ when type i travels as the MEC_i is independent of the type. Consequently, the system is self-financing with the step-toll revenue equal to the cost of the optimal capacity.

Proof of Proposition 6: The MEC_i is constant within a travel period and the same for all types, but now it does differ between the shoulder periods and the center peak. As with ratio heterogeneity, the difference in toll between the toll in the shoulders and the center equals the difference in MEC_i. So, as the MEC is the same for all, the toll equals the MEC throughout the peak, and thus, following Lemmas 1–2, the scheme is self-financing. □

To conclude, with proportional heterogeneity, flat and coarse tolling lead to marginal external cost pricing and, therefore, self-financing of a capacity that follows the first-best rule. When there are many types, even more equilibria are possible. Nevertheless, the outcome will be qualitatively the same: (i) the types with the higher values use the center period; (ii) there will be at most one type that uses the center and the shoulders; (iii) the MEC in a period will be the same for all types of that period so that self-financing can hold.

7. Other forms of heterogeneity

Having looked at ratio and proportional heterogeneity, we now briefly discuss the two other possible dimensions of heterogeneity in preferences in a dynamic congestion model with linear scheduling costs. These dimensions are in the preferred arrival time (t_i^*) and between the value of schedule delay early and late (β/γ_i), as used in Arnott et al. (1988).

Heterogeneity in β/γ_i means that α/γ_i also varies, so there will effectively be ratio heterogeneity for late arrivals if the high-γ type also arrives late in user equilibrium. So, using our earlier results, self-financing is not ensured.

Following Arnott et al. (1988), heterogeneity in t_i^* means that users separate over arrival time into periods around their preferred arrival time. However, as long as the value of time and schedule delay are homogeneous, the MEC will be the same for all types if, in equilibrium, the queue has a single peak with flat tolling, and with step tolling a single peak in the center period and in each shoulder period. If this is not the case, then the MEC will differ over types and self-financing may not hold.

We do not look at general heterogeneity in multiple dimensions and with many types. Such heterogeneity is, of course, present in reality, but it tremendously complicates the analytical analysis of step tolling due to the explosion of the number of possible equilibria. Yet, using the results of van den Berg (2011a) and Hall (2018), it seems plausible that, as long as t^* is homogeneous, the effects of the different dimensions of heterogeneity are qualitatively the same under more general heterogeneity. However, following Hall (2018, 2023), if t^* is also heterogeneous, we add the extra complication of marginal vs inframarginal users, where inframarginal users can only arrive at their preferred arrival time, but their preferences do not affect the equilibrium. This also substantially changes the equilibrium travel time developments. The inframarginal users at arrival time t would have a ratio of value of time to value of schedule delay below that of the marginal users. Analysis of this is beyond the scope of this paper, but we speculate that the marginal users would probably determine the toll, whereas the optimal capacity depends on both the marginal and inframarginal users.[26] This would imply a second reason why self-financing may not hold. This issue is akin to the effect of marginal vs inframarginal users on quality setting in Spence (1975).

8. Conclusion

We studied whether a road with bottleneck congestion is self-financing under flat or step tolling: Does the toll revenue cover the bottleneck capacity costs? Self-financing will hold if the toll can equal the marginal external cost (MEC) throughout the peak. However, with ratio heterogeneity between the value of time (VOT) and values of schedule delay, the MEC is heterogeneous, and, hence, the toll cannot equal the MEC at all moments. Accordingly, the system is only exactly self-financing for very specific parameter combinations. Other dimensions of preference heterogeneity may not in themselves lead to a violation of the self-financing result.

We derived explicit formulas for the capacity setting under ratio heterogeneity. The capacity rule has a second-best correction: the capacity is set higher than following the first-best rule. The second-best correction is larger with flat tolling than with step tolling, so the deviation from self-financing

[26] Hall (2023) argued that not including both ratio and t^* heterogeneity can lead to incorrect travel time developments. To test for the effect of this, we change the ratio of VOT to values of schedule delay in the sensitivity check in Fig. C.6. Changing the mean VOT (while keeping the ratio αL/αH constant) scales the queuing delays, but it has no effect on tolls, costs, and welfare. This also becomes clear by looking at the travel cost and toll equations. Hence, we vary β from 0.25 to 7.25 per hour (so that it remains below α_L=7.5/h) whilst keeping the ratio β/γ constant. The base-case level of β is €6.09/h. Lowering the value of schedule delay leads to lower travel times and makes schedule delays less costly. It also lowers the levels of the tolls. Interestingly, the effect of changing β on the percent profits is opposite for the step and flat toll. However, for both tolls, the relative efficiency and welfare loss of imposing self-financing go up with β.

will be larger with flat tolling. The second-best coarse toll is the weighted average of the MECs, where weights depend on cost and demand derivatives and not on the frequency of the types. The scheme can only be self-financing if the users with a low value of time are much more price sensitive than those with a high value: this raises the coarse toll to well above the (average) MEC, and this covers the extra cost of second-best capacity being higher than what the first-best would be. The system is most likely to have a deficit. In our numerical model, the users with a low value of time must be almost twice as price sensitive for there not to be a loss, and, typically, the loss is 0–10% of capacity costs.

Returning to our policy question, our analysis shows that coarse tolling systems are unlikely to be precisely self-financing but that any loss will likely be small. This is important for the acceptability of congestion pricing. The effect on welfare of imposing self-financing tends to be small and even minute with a step toll.

An obvious follow-up question is how a multi-step toll would perform. Building on the results of Lindsey et al. (2012) for homogeneous users, we may expect the step toll to approach the time-variant MEC as the number of steps goes to infinity.[27] Other interesting follow-up research areas are multiple dimensions of heterogeneity. In particular, Hall (2023) has shown that adding heterogeneity in the preferred arrival time to ratio heterogeneity can massively change the effects of tolling. Considering capacity setting in networks also seems important if only because raising one road's capacity can raise costs on some others (see, e.g., Arnott et al. (1993b) and Wang et al. (2022)). And what are the effects of nonlinear schedule delay costs as in Lindsey (2004)? Finally, the system is self-financing for any dynamic model if the toll always equals the MEC. However, this is not the case with ratio heterogeneity. This raises the question: how does the congestion model affect the welfare and distributional effects of step tolling, the lack of self-financing and the effects of imposing self-financing?

Acknowledgments

We are thankful for the comments of the participants of the ITEA conference in Hong Kong and at the Eureka seminar in Amsterdam. We also thank the editors, reviewers and Erik Verhoef for helpful comments. Any remaining errors are ours. There has been no external financial support for this research.

Appendix:

Appendix A. Detailed derivations for the flat toll under ratio heterogeneity

A.1 Proof of Lemma 3

Maximizing welfare is equivalent to maximizing the below Lagrangian to N_H, N_L, μ^F, s, λ_L^F and λ_H^F :

[27] This is true for the Laih and ADL step-toll models, but not for the Braking model.

$$L\ = B_L + B_H - \left(c_H^F \cdot N_H + c_L^F \cdot N_L + k \cdot s\right) + \lambda_L^F \cdot \left(c_L^F + \mu^F - D_L\right) + \lambda_H^F \cdot \left(c_H^F + \mu^F - D_H\right). \tag{9}$$

The first order conditions are:

$$\frac{\partial L}{\partial N_L} = 0 = D_L - \left(c_L^F + MEC_L\right) + \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial N_L} - \frac{\partial D_L}{\partial N_L}\right) + \lambda_H^F \cdot \left(\frac{\partial c_H^F}{\partial N_L}\right) \tag{A.1}$$

$$\frac{\partial L}{\partial N_H} = 0 = D_H - \left(c_H^F + MEC_H\right) + \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial N_H}\right) + \lambda_H^F \cdot \left(\frac{\partial c_H^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H}\right) \tag{A.2}$$

$$\frac{\partial L}{\partial \mu} = 0 = \lambda_L^F + \lambda_H^F \tag{A.3}$$

$$\frac{\partial L}{\partial s} = 0 = -\left(\frac{\partial c_H^F}{\partial s} \cdot N_H + \frac{\partial c_L^F}{\partial s} \cdot N_L + k \cdot s\right) + \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial s}\right) + \lambda_H^F \cdot \left(\frac{\partial c_H^F}{\partial s}\right), \tag{A.4}$$

$$\frac{\partial L}{\partial \lambda_L^F} = 0 = \left(c_L^F + \mu^F - D_L\right) \tag{A.5}$$

$$\frac{\partial L}{\partial \lambda_H^F} = 0 = \left(c_H^F + \mu^F - D_H\right) \tag{A.6}$$

Applying (A.3), we get:

$$\lambda_L^F = -\lambda_H^F \tag{A.7}$$

Using this, (A.2) and (A.6), we can then solve for λ_L^F:

$$0 = \mu^F - \left(MEC_L\right) + \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H} - \frac{\partial c_H^F}{\partial N_H}\right)$$

$$\lambda_L^F = \frac{MEC_L - \mu}{\dfrac{\partial c_L^F}{\partial N_H} - \dfrac{\partial D_H}{\partial N_H} - \dfrac{\partial c_H^F}{\partial N_H}} > 0 \tag{A.8}$$

Combining (A.1), (A.6) and (A.7), we find

$$0 = \mu^F - \left(MEC_L\right) + \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial N_L} - \frac{\partial D_L}{\partial N_L} - \frac{\partial c_H^F}{\partial N_L}\right).$$

And then using (A.8), we get:

$$\mu^F = MEC_L + \frac{\dfrac{\partial c_L^F}{\partial N_L} - \dfrac{\partial D_L}{\partial N_L} - \dfrac{\partial c_H^F}{\partial N_L}}{\dfrac{\partial c_L^F}{\partial N_H} - \dfrac{\partial D_H}{\partial N_H} - \dfrac{\partial c_H^F}{\partial N_H}} \cdot \left(MEC_H - \mu^F\right)$$

$$\mu^F \cdot \left(\frac{\partial c_L^F}{\partial N_L} - \frac{\partial D_L}{\partial N_L} - \frac{\partial c_H^F}{\partial N_L} + \frac{\partial c_L^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H} - \frac{\partial c_H^F}{\partial N_H}\right) = MEC_L \cdot \left(\frac{\partial c_L^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H} - \frac{\partial c_H^F}{\partial N_H}\right) + \left(\frac{\partial c_L^F}{\partial N_L} - \frac{\partial D_L}{\partial N_L} - \frac{\partial c_H^F}{\partial N_L}\right) \cdot MEC_H$$

$$\mu^F = \frac{MEC_L \cdot \left(\frac{\partial c_L^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H} - \frac{\partial c_H^F}{\partial N_H} \right) + MEC_H \cdot \left(\frac{\partial c_L^F}{\partial N_L} - \frac{\partial D_L}{\partial N_L} - \frac{\partial c_H^F}{\partial N_L} \right)}{\left(\frac{\partial c_L^F}{\partial N_L} - \frac{\partial D_L}{\partial N_L} - \frac{\partial c_H^F}{\partial N_L} \right) + \left(\frac{\partial c_L^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H} - \frac{\partial c_H^F}{\partial N_H} \right)} \qquad \textit{(A.9)}$$

From this we get, the general equation of Lemma 3:

$$\mu^F = MEC_L^F \cdot w_L^F + MEC_H^F \cdot (1 - w_L^F) \qquad \textit{(A.10)}$$

$$w_L^F = \frac{-\frac{\partial D_H}{\partial N_H} - \frac{\partial c_L}{\partial N_H} + \frac{\partial c_H}{\partial N_H}}{\left(-\frac{\partial D_H}{\partial N_H} - \frac{\partial c_L}{\partial N_H} + \frac{\partial c_H}{\partial N_H} \right) + \left(-\frac{\partial D_L}{\partial N_L} - \frac{\partial c_H}{\partial N_L} + \frac{\partial c_L}{\partial N_L} \right)} \qquad \textit{(A.11)}$$

Finally, the bottleneck-specific equations are found by inserting the cost equations. Using these, you can also show that the second-order conditions hold. This completes the proof of Lemma 3. □

A.2 Proof of Proposition 1

The equation (12) of the proposition follows directly from the f.o.c. in (A.4). That $\lambda_L^F > 0$ is visible in (A.8). Finally, using the cost equation in (6), we see that the High-VOT type's cost decreases faster with *s* than that of the Low-VOT type, as it has a higher cost. This implies $\left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right) > 0$.

Hence, the capacity rule with flat tolling of $-\sum N_i \cdot \partial c_i^F / \partial s = k - \lambda_L^F \left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right)$ implies a higher capacity *s* for given N_L and N_H than the first-best rule of $-\sum N_i \cdot \partial c_i^F / \partial s = k$. And thus, the flat toll has a lower volume–capacity ratio $(N_L+N_H)/s$. This completes the proof of Proposition 1. □

A.3 Proof of Proposition 2

The profit equals toll revenue minus capacity cost:

$$\begin{aligned} \Pi^F &= \mu^F \cdot (N_L + N_H) - s \cdot k \\ &= \mu^F \cdot (N_L + N_H) + \left(\frac{\partial c_L^F}{\partial s} \cdot N_L + \frac{\partial c_H^F}{\partial s} \cdot N_H \right) \cdot s - s \cdot \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s} \right), \end{aligned}$$

where the second line follows from inserting the capacity condition from Proposition 1.

By plugging in the functional forms, one can show that with the linear travel cost of the bottleneck model: $\frac{\partial c_i^F}{\partial s} \cdot s = MEC_i$. This makes profit:

$$\Pi^F = \mu^F \cdot (N_L + N_H) + (MEC_L \cdot N_L + MEC_H \cdot N_H) - s \cdot \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s} \right) \qquad \text{(A.12)}$$

$$= \left(\sum_i (\mu^F - MEC_i) N_i \right) - s \cdot \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s} \right) \qquad \text{(A.13)}$$

Equation (A.13) is the general profit equation given in the proposition. As $-s\cdot\lambda_L^F\cdot\left(\frac{\partial c_L^F}{\partial s}-\frac{\partial c_H^F}{\partial s}\right)$ must be negative, the first term of (A.13) must be positive for profit to be zero, which only happens when the average toll exceeds the average MEC. This completes the proof of Proposition 2. □

Appendix B. Timings of the peak and total cost with a step toll under ratio heterogeneity

For the (generalized) price not to be higher in the center period when the toll is ρ higher, the travel time at the start of the center period must be lower than for an arrival just before it. Therefore, at some moment, departures stop and the queue starts shrinking. To maximize the reduction in queuing time, t^+ and ρ are chosen so that the queue reaches zero at t^+. If, for instance, there were still some queueing at t^+, then starting the center period a bit earlier would reduce the travel time of all center peak users without affecting shoulder period users, thereby lowering total cost. On the other hand, having a period without arrivals just before t^+ would only raise costs, so the queuing in the early shoulder period ends exactly at t^+.

The question then arises of how t^- is set. Arrivals just after t^- pay a much lower toll than arrivals just before t^-. So, for a constant price, the travel time must be much higher for arrivals just after t^-. In the Laih model, this is attained by having the users who arrive after t^- wait beside the road just before the bottleneck without impeding other drivers. Hence, there are separate queues. The t^- is then set so the last center peak user arrives exactly at t^-. Accordingly, in equilibrium, the periods are:

$$t^+ = -\frac{\gamma}{\beta+\gamma}\frac{V_L+V_H}{s}$$

$$t^- = \frac{\beta}{\beta+\gamma}\frac{V_L+V_H}{s}$$

The early shoulder period before t^+ will automatically have the same price as the late shoulder period since, in user equilibrium, a fraction $\gamma/(\beta+\gamma)$ of the shoulder users travels early. For users to be willing to travel in the center and shoulders, the step part of the toll, ρ, must equal the difference in cost between the center period and shoulders of the peak (from t_s to t^+ and from t^- to t_e).

All this makes the travel costs for both types:

$$c_L^{SS} = \begin{cases} c_L^{cp} = \delta\dfrac{V_L+\dfrac{\alpha_L}{\alpha_H}\cdot V_H}{s} & \text{if } t^+\le t\le t^- \\[2ex] c_L^{sh} = \delta\dfrac{V_L+V_H}{s}+\delta\dfrac{N_L-V_L+\dfrac{\alpha_L}{\alpha_2}\cdot\left(N_H-V_H\right)}{s} & \text{if } t^+\ge t\ge t_s \text{ or } t^-\le t\le t_e \end{cases}$$

$$c_H^{SS} = \begin{cases} c_H^{cp} = \delta\dfrac{V_L+V_H}{s} & \text{if } t^+\le t\le t^- \\[2ex] c_H^{sh} = \delta\dfrac{N_L+N_H}{s} & \text{if } t^+\ge t\ge t_1 \text{ or } t^-\le t\le t_e \end{cases}$$

As we will show, it is optimal that, of each type, half of its users travel in the center peak when the toll is higher, and the other half in the shoulder periods. Still, within each period, the user types travel separated, with the Low-VOT users arriving closer to the preferred arrival time. As the queue reaches zero at the start and end of the center peak period, the total cost is:

$$TC^{ss} = k \cdot s + \frac{((\mathrm{N_L}+\mathrm{N}_H)(\mathrm{N}_H-\mathrm{V}_H)+\mathrm{V}_H(\mathrm{V}_L+\mathrm{V}_H))\delta}{s} + \frac{((\mathrm{N}_L-\mathrm{V}_L)(\mathrm{N}_L+\mathrm{V}_H+\frac{(\mathrm{N}_H-\mathrm{V}_H)\alpha_L}{\alpha_H})+\mathrm{V}_L(\mathrm{V}_L+\frac{\alpha_L}{\alpha_H}\mathrm{V}_H))\delta}{s}$$

and it is globally convex in V_L and V_H. Taking derivatives, we find that the minimum is at V_L=N_L/2 and V_H=N_H/2.

Appendix C. More sensitivity analyses for the numerical model for ratio heterogeneity

Figures C.1, C.2, C.3, C.4, C.5 and C.6 extend the sensitivity analysis by looking at some changes omitted in the main text. As overall demand becomes more elastic, all policies lower demand more by raising the price, except that the first-best fully time variant toll lowers the price for the High-VOT users and thus raises their demand. Coarse tolling leads to a relatively higher welfare when total demand is more elastic. These policies partly deal with the congestion externality by raising the generalized price, which is less harmful to consumers with more elastic demands. For all policies except the first-best, more elastic overall demand means a lower optimal capacity as there will be fewer users. Finally, a higher marginal capacity cost, *k*, means that a flat or coarse toll will have a larger loss. However, for the step toll, the loss as a percentage of the capacity cost falls as the effect of *k* on the capacity cost is stronger than its effect on the profit. For both schemes, imposing self-financing hurts welfare more when *k* is larger.

Fig. C.1. Effects of the policies on usages over the average elasticity.

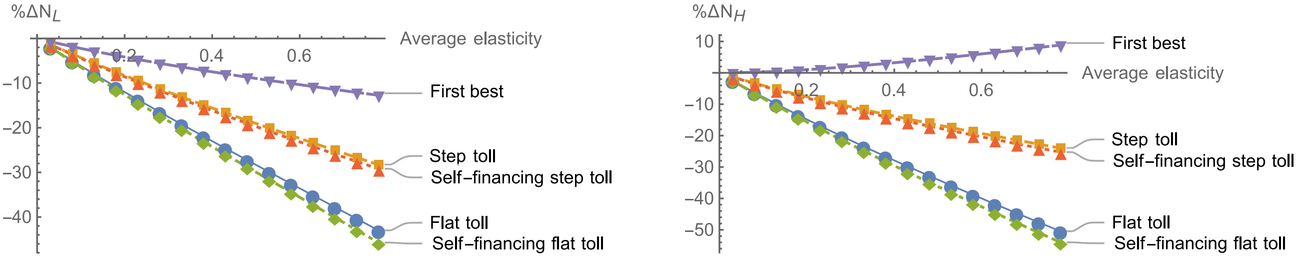


Fig. C.2. Relative efficiencies and the average elasticity.

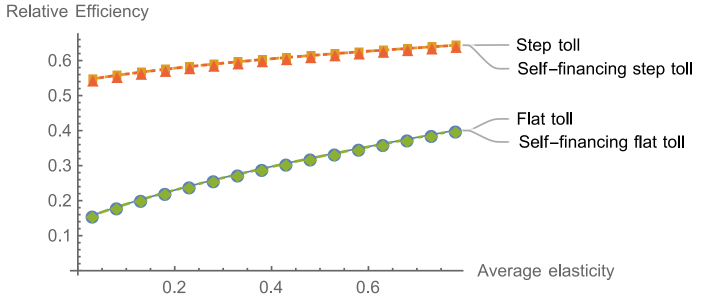
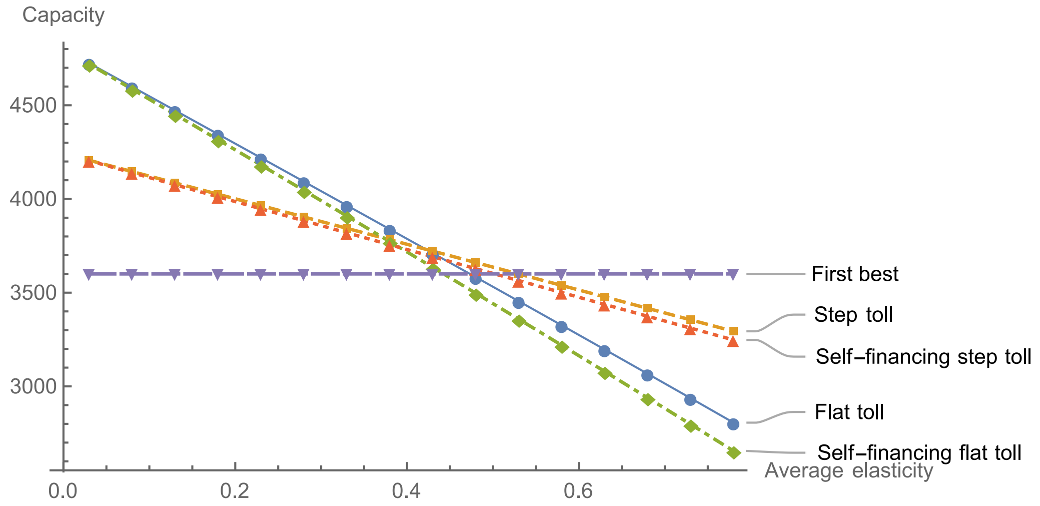


Fig. C.3. The effect of the average fuel cost elasticity on the optimal capacities.



Note: We use the elasticity's absolute value, so a larger number implies more elastic demand.

Fig. C.4. Effects of the policies on usages over the degree of heterogeneity (%Δα).

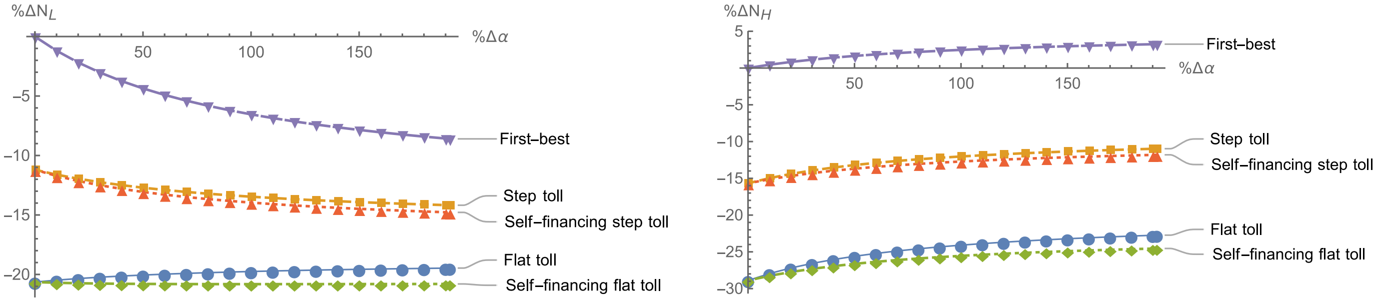


Fig. C.5. Profit and welfare change due to imposing self-financing over the marginal capacity cost (k).

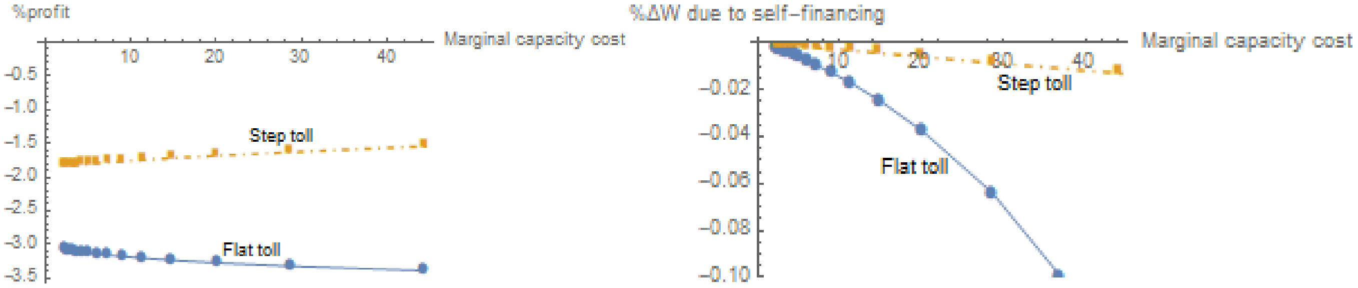
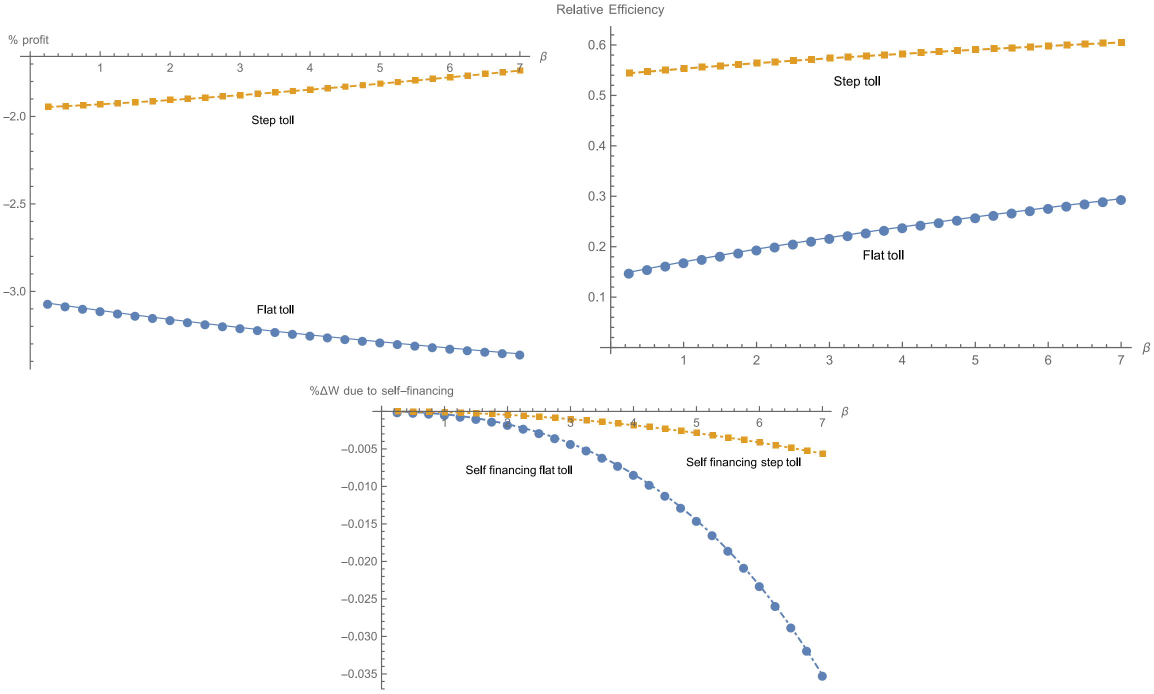


Fig. C.6. Profits(a), relative efficiency(b) and welfare change due to imposing self-financing (c) over the value of schedule delay early (β).



Appendix D: Overview of notation, parameters and variables

Nomenclature

α_i	Value of time (VOT) of user of type *i*. This is the cost of an hour of travel time.
β_i	Value of schedule delay early of user of type *i*. This is the cost of arriving an hour earlier than the preferred arrival time t^*.
γ_i	Value of schedule delay late of user type *i*. This is the cost of arriving an hour later than the preferred arrival time t^*.
δ_i	Compound preference parameter for type *i*: $\delta_i \equiv \beta_i \gamma_i / (\beta_i + \gamma_i)$.
λ_i^F	Lagrange multiplier for the user equilibrium constraint of type *i* for the flat toll (as indicated by superscript *F*). The single-step toll has similar multipliers, but with *SS* indicating the single-step toll.
$\tau[t]$	Toll, τ, that varies over arrival time *t*.
ρ	Step part of the toll. It is levied between t^+ and t^- in addition to the flat toll of μ.
μ	Flat part of the toll, it does not vary over time.
B_i	Consumer benefit. It is the integral of the inverse demand, D_i, from 0 to N_i.
$D_i[N_i]$	Inverse demand of type *i*. It gives the willingness to pay for a trip of the N_i'th user.
$d0_i$	The numerical model uses a linear demand, with $d0_i$ being the demand intercept of type *i*.
$d1_i$	The numerical model uses a linear demand, with $d1_i$ being the demand slope of type *i*.
c_i	(Average) ravel cost for a type *i* user. It is the sum of the queuing time and schedule delay costs.
f_i	Frequency of type *i* with two type heterogeneity: $f_i = N_i/(N_L+N_H)$.
k	Marginal cost per unit of capacity of the bottleneck. The total capacity cost is $k \cdot s$.
MEC_i	Marginal external cost of a type *i* user, which equals the marginal social cost of type *i* ($MSC_i = \partial TC/\partial N_i$) minus the own travel cost: $MEC_i = MSC_i - c_i = \partial TC/\partial N_i - c_i$.
MSC_i	Marginal social cost of a type *i* user is the derivative of the total cost to the number of type *i* users: $MSC_i = \partial TC/\partial N_i$.
N_i	Total number of users of type *i*.
P_i	(Generalized) price for type *i*. It equals the travel cost, c_i, plus the possible toll.
s	Bottleneck capacity.
t	Arrival time.
t^*	Preferred arrival time.
t^+	Moment when the step part of the toll is turned *on* and the toll jumps *up*.
t^-	Moment when the step part of the toll is turned *off* and the toll jumps *down*.
t_e	Moment of the last arrival and, hence, when the peak ends.
t_s	Moment of the first arrival and, hence, when the peak starts.
TT	Travel time.
TC	Total cost is total capacity cost, $k \cdot s$, plus the travel costs summed over all types: $TC = k \cdot s + \sum N_i \cdot c_i$.
V_i	With step tolling, the number of users of type *i* that travel in the center peak when the toll is higher.
W	Welfare equals the summed consumer benefits, B_i, minus total cost: $W = \sum B_i - \sum N_i \cdot c_i - k \cdot s$.
w_i	The weight attached to type *i*'s marginal external cost in the toll rule.
Π	Profit, which equals toll revenue minus the capacity cost of $k \cdot s$.
Indicators used in superscripts	
^{F}	Flat toll.
^{SS}	Single-step toll.
^{sh}	Shoulder periods with a step toll when the toll equals μ. It lasts from t_s to t^+ and from t^- to t_e.
^{cp}	Center peak period with a step toll when the toll equals $\mu+\rho$. It lasts from t^+ to t^-.

References

Akamatsu, T., Wada, K., Iryo, T., Hayashi, S. 2021. A new look at departure time choice equilibrium models with heterogeneous users. Transportation Research Part B: Methodological, 148, 152–182.

Arnott, R., de Palma, A., Lindsey, R. 1988. Schedule delay and departure time decisions with heterogeneous commuters. Transportation Research Record 1197, 56–67.

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. 1993a. A structural model of peak-period congestion: a traffic bottleneck with elastic demand. American Economic Review 83 (1), 161–179.

Arnott, R., de Palma, A., Lindsey, R. 1993b. Properties of dynamic traffic equilibrium involving bottlenecks, including a paradox and metering. Transportation science, 27(2), 148-160.

Arnott, R., Kraus, M. 1993. The Ramsey problem for congestible facilities. Journal of Public Economics, 50(3), 371–396.

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

Arnott, R., Kraus, M. 1998a. When are anonymous congestion charges consistent with marginal cost pricing? Journal of Public Economics, 67, 45–64.

Arnott, R., Kraus, M. 1998b. Self-financing of congestible facilities in a growing economy. In: Pines, D., Sadka, E., Zilcha, I. (eds.). Topics in Public Economics: Theoretical and Applied Analysis. Cambridge, UK : Cambridge University Press , pp. 161–184.

Bichsel, R. 2001. Should road users pay the full cost of road provision? Journal of Urban Economics, 50: 367–383.

Börjesson, M., Cherchi, E., Bierlaire, M. 2013. Within-individual variation in preferences: equity effects of congestion charges. Transportation research record, 2382(1), 92-101.

Börjesson, M., Kristoffersson, I., 2012. Estimating Welfare Effects of Congestion Charges in Real World Settings. CTS Working Paper 2012:13.

Brons, M., Nijkamp, P., Pels, E., Rietveld, P. 2008. A meta-analysis of the price elasticity of gasoline demand. A SUR approach. Energy Economics, 30(5), 2105–2122.

Chen, H., Liu, Y., Nie, Y.M. 2015a. Solving the step-tolled bottleneck model with general user heterogeneity. Transportation Research Part B: Methodological, 81, 210–229.

Chen, H., Nie, Y.M., Yin, Y. 2015b. Optimal multi-step toll design under general user heterogeneity. Transportation Research Part B: Methodological, 81, 775–793.

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

de Palma, A., Lindsey, R. 2007. Transport user charges and cost recovery. Research in Transportation Economics, 19, 29–57.

D'Ouville, E. L., McDonald, J.F. 1990. Effects of demand uncertainty on optimal capacity and congestion tolls for urban highways. Journal of Urban Economics, 28(1), 63–70.

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.

Ge, Y.E., Stewart, K., Sun, B., Ban, X.G., Zhang, S. 2016. Investigating undesired spatial and temporal boundary effects of congestion charging. Transportmetrica B: Transport Dynamics, 4(2), 135–157.

Guo, R.Y., Yang, H., Huang, H.J. 2023. The Day-to-Day Departure Time Choice of Heterogeneous Commuters Under an Anonymous Toll Charge for System Optimum. Transportation Science, Articles in Advance, 1–24

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. 2021. Can tolling help everyone? Estimating the aggregate and distributional consequences of congestion pricing. Journal of the European Economic Association, 19(1), 441–474.

Hall, J.D. 2023. Inframarginal Travelers and Transportation Policy (September 7, 2023) . SSRN working paper, 3424097.

Knockaert, J., Verhoef, E. T., & Rouwendal, J. (2016). Bottleneck congestion: Differentiating the coarse charge. Transportation Research Part B: Methodological, 83, 59-73.

Kouwenhoven, M., de Jong, G.C., Koster, P., van den Berg, V.A.C., Verhoef, E.T., Bates, J.,Warffemius, P.M. 2014. New values of time and reliability in passenger transport in The Netherlands. Research in Transportation Economics, 47, 37–49.

Kraus, M. 1982. Highway pricing and capacity choice under uncertain demand. Journal of Urban Economics, 12(1), 122–128.

Laih, C.H. 1994. Queuing at a bottleneck with single and multi-step tolls. Transportation Research Part A 28 (3), 197–208.

Laih, C.H. 2004. Effects of the optimal step toll scheme on equilibrium commuter behavior. Applied Economics 36 (1), 59–81.

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, Z.C., Lam, W.H., Wong, S.C. 2017. Step tolling in an activity-based bottleneck model. Transportation Research Part B: Methodological, 101, 306–334.

Lindsey, R. 2004. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. Transportation science, 38(3), 293-314.

Lindsey, R. 2012. Road pricing and investment. Economics of transportation, 1(1-2), 49–63.

Lindsey, R., de Palma, A. 2014. Cost recovery from congestion tolls with long-run uncertainty. Economics of Transportation 3(2), 119–132.

Lindsey, R., van den Berg, V.A.C., Verhoef, E.T. 2012. Step tolling with bottleneck queuing congestion. Journal of Urban Economics 72 (1), 46–59.

Liu, Y., Nie, Y. M., Hall, J. 2015a. A semi-analytical approach for solving the bottleneck model with general user heterogeneity. Transportation research part B: methodological, 71, 56–70.

Liu, W., Yang, H., Yin, Y. 2015b. Efficiency of a highway use reservation system for morning commute. Transportation Research Part C: Emerging Technologies, 56, 293-308.

Lu, Z., Meng, Q. 2017. Analysis of optimal BOT highway capacity and economic toll adjustment provisions under traffic demand uncertainty. Transportation Research Part E 100, 17–37.

Mohring H, Harwitz, M. 1962. Highway Benefits: An Analytical Framework. Evanston, IL: Northwestern University Press.

Newell, G.F. 1987. The morning commute for nonidentical travellers. Transportation Science 21 (2), 74–88

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.

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

Small, K.A. 1999. Economies of scale and self-financing rules with noncompetitive factor markets. Journal of Public Economics, 74, 431–450 .

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

Small, K.A., Winston, C., Yan, J. 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. Econometrica 73(4), 1367–1382.

Small, K.A., Verhoef, E.T. 2007. The Economics of Urban Transportation. Routledge, London.

Spence, A.M., 1975. Monopoly, quality, and regulation. The Bell Journal of Economics, 417-429.

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.

van den Berg, V.A.C. 2014. Coarse tolling with heterogeneous preferences. Transportation Research Part B: Methodological, 64, 1–23.

van den Berg, V.A.C., Verhoef, E.T. 2011a. Winning or losing from dynamic bottleneck congestion pricing? The distributional effects of road pricing with heterogeneity in values of time and schedule delay. Journal of Public Economics 95 (7–8), 983–992.

van den Berg, V.A.C., Verhoef, E.T. 2011b. Congestion tolling in the bottleneck model with heterogeneous values of time. Transportation Research Part B 45(1), 60–70.

Verhoef, E.T., Mohring, H., 2009. Self-financing roads. International Journal of Sustainable Transportation, 3(5-6), 293-311.

Vickrey, W.S., 1973. Pricing, metering, and efficiently using urban transportation facilities. Highway Research Record 476, 36–48.

Wang, T., Liao, P., Tang, T. Q., Huang, H.J. 2022. Deterministic capacity drop and morning commute in traffic corridor with tandem bottlenecks: A new manifestation of capacity expansion paradox. Transportation research part E: logistics and transportation review, 168, 102941.

Wu, W.X., Huang, H. J. 2015. An ordinary differential equation formulation of the bottleneck model with user heterogeneity. Transportation Research Part B: Methodological, 81, 34-58.

Xiao, F., Qian, Z., Zhang, H.M., 2011. The morning commute problem with coarse toll and nonidentical commuters. Networks and Spatial Economics 11 (2),343–369.

Xiao, F., Shen, W., Zhang, H.M., 2012. The morning commute under flat toll and tactical waiting. Transportation Research Part B 46 (10), 1346–1359.

Xu, D., Guo, X., Zhang, G. 2019. Constrained optimization for bottleneck coarse tolling. Transportation Research Part B: Methodological, 128, 1–22.

Yang, H., Meng, Q. 2002. A note on ”highway pricing and capacity choice in a road network under a build-operate-transfer scheme”. Transportation Research Part A: Policy and Practice, 36(7), 659–663.

Zheng, N., Waraich, R.W., Axhausen, K.W., Geroliminis, N. 2012. A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model. Transportation Research Part A 46 (8), 1291–1303.