van den Berg, Vincent

Working Paper

Self-financing roads under coarse tolling and heterogeneous preferences

Tinbergen Institute Discussion Paper, No. TI 2022-045/VIII

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: van den Berg, Vincent (2022) : Self-financing roads under coarse tolling and heterogeneous preferences, Tinbergen Institute Discussion Paper, No. TI 2022-045/VIII, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/263965

TI 2022-045/VIII
Tinbergen Institute Discussion Paper

Self-financing roads under coarse tolling and heterogeneous preferences

Vincent A.C. van den Berg[1,2]

1 Vrije Universiteit Amsterdam

2 Tinbergen Institute

Self-financing roads under coarse tolling and heterogeneous preferences

Version of 13-07-2022

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 if a road is self-financing under flat or step tolling and optimized capacity while incorporating preference heterogeneity, bottleneck congestion and linear capacity cost. Previous work has shown that a *sufficient condition* for the toll revenue to equal the capacity cost is that the toll to equals the marginal external costs (MECs) of all types of user at all moments when their users travel. However, under 'ratio heterogeneity' between values of time (VOT) and schedule delay, an anonymous second-best coarse toll must differ from the heterogeneous MECs. This paper derives that this toll will be a weighted average of the MECs with the weights depending on the derivatives of the demand and travel cost functions. 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 was 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, as this raises the toll. In our numerical model, the Low-VOT type must be almost twice as price sensitive than the High-VOT type for there not to be loss; and, typically, there is a 5% to 15% loss. Imposing self-financing only causes a small welfare loss of 0% to 1.5%.

We also analyze other forms of heterogeneity: proportional heterogeneity, heterogeneity in the preferred arrival time and heterogeneity between values of schedule delay early and late.

Keywords: Self-financing, road pricing, flat toll, step toll, coarse toll, heterogeneity, second best

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), the toll revenue will exactly equal the cost of the optimized road capacity. This 'first-best' capacity minimized total cost, which is the sum of capacity and total travel cost. Self-financing is important for the social acceptability of congesting pricing. It means that people have to pay to use the roads, but this goes toward road capacity and so they get something in return. It means that there is no cross-subsidization between modes and that no other distortive taxes are needed to pay for our roads. Finally, road capacity is mostly set to deal with peak demand; so it can be seen as fair that users of the centre peak pay more. All this is why self-financing has been extensively studied, especially in settings that are more realistic: with, for example, uncertainty, dynamics with the day or the long run (see the extended literature review in Section 2).

Mohring and Harwitz (1962) studied homogeneous users and static congestion. Their conditions for self-financing—which we assume throughout this paper—are that capacity and number of users are continuous, capacity cost is homogeneous to degree one in capacity and that per car travel cost only depends on the ratio of number of cars and capacity.[1] Arnott et al. (1992, 1993) showed that the self-financing result carries over to dynamic bottleneck congestion, both with a fine toll that can vary every microsecond as 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 some US pay-lanes. Under preference heterogeneity, Arnott and Kraus (1995) found that a *sufficient condition* for self-financing is that the (coarse) toll equals the MEC throughout the peak. They consider two-type 'ratio heterogeneity' between the value of time (VOT) and values of schedule delay.[2] This heterogeneity means that people differ in their preference for trading off travel time and schedule delay (i.e. arriving at an earlier or later than the preferred arrival time). Thus, it implies differences in flexibility. This heterogeneity could stem from differences in type of job, trip or family status (Van den Berg and Verhoef, 2011a; Hall, 2018, 2021ab).[3] The marginal external congestion cost (MEC) is higher for the Low-VOT users than for High-VOT users. Hence, an anonymous coarse toll must be second best as it cannot equal the MEC throughout the peak (Arnott and Kraus, 1995). Only few papers have looked at user heterogeneity and self-financing in a dynamic congestion model,[4] and this is the aim of the present paper. We study coarse tolls

[1] The assumption on capacity cost implies neutral scale economies, which later empirical research found to hold more or less, at least at the network level (de Palma and Lindsey, 2002; Small and Verhoef, 2007).

[2] The value of time 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. Note that unlike the current paper, they assumed heterogeneity in the value of schedule delay for a fixed value of time, whereas we use a heterogeneous value of time and fixed values of schedule delay. Their set-up in fact in implies that there is both ratio and proportional heterogeneity, but as we will see only the presence of the ratio heterogeneity affects the self-financing.

[3] Income differences may have little to nothing to do with ratio heterogeneity, as they should affect values of time and schedule delay similarly.

[4] Yet, in really, there certainly is heterogeneity: different people have different preferences (e.g., Small et al., 2005; Small, 2012). Moreover, preference heterogeneity affects the optimal levels of policy instruments and their effects on consumer surplus and societal welfare. Heterogeneity also means that policies have distributional effects (e.g., Arnott er al., 1988; Small and Yan, 2002).

and capacity setting under preference heterogeneity: how large a deficit or profit will there be and what is the welfare loss of imposing self-financing?

What Arnott and Kraus (1995) did not consider—and what we will show—is that, under two-type ratio heterogeneity, the capacity rule will also be second best. The second-best capacity is higher than following the first-best rule that minimized total cost. This raises welfare by increasing the number of High-VOT users as they face a toll that exceeds their MEC and thus have too little travel. Even with the extra second-best capacity, the scheme may have a zero or even positive profit if the (average) toll well exceeds the average MEC. Our toll rule shows that, for this to happen, the Low-VOT type must be much more price sensitive than the High-VOT type. Low-VOT users have the higher MEC, and them being more price sensitive means that a second-best flat or step toll is closer to their MEC. The single-step toll tends to have a lower loss or profit than the flat toll and it is less likely to have a loss. This is because the step toll has a smaller second-best capacity adjustment.

Ratio heterogeneity seems the most interesting dimension of heterogeneity as it affects the self-financing result. Consistent with Van den Berg and Verhoef and Van den Berg (2014), 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. Finally, heterogeneity between values of schedule delay early and late means that there must also be ratio heterogeneity, and thus it has similar effects as ratio heterogeneity.

Our methodological contribution is deriving, under preference heterogeneity, explicit formulas for the optimal flat and single-step toll as well as for the capacity. The flat toll and the flat part of the step toll are a weighted average of the MECs, with the weights depending on the derivatives of the demand and cost functions. The weights are *independent* of the numbers of users of the types. 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 road by considering coarse tolls under heterogeneity and dynamic congestion.

The next 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 does extensive sensitivity analyses. Sections 6 and 7 look at other forms of heterogeneity. Section 8 concludes. The below nomenclature box summarises the notation.

Nomenclature

α_i	Value of time (VOT) of user of type *i*. It is the cost of an hour of travel time.
β_i	Value of schedule delay early of user of type *i*. It 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*. It is the cost of arriving an hour later than the preferred arrival time t^*.
δ_i	Compound preference parameter for type *i*: $\delta_i \equiv \beta_i\gamma_i/(\beta_i+\gamma_i)$
η_i	Relative preference parameter for type *i*: $\eta_i \equiv \beta_i/\gamma_i$
$\tau[t]$	Toll, τ, that varies of arrival time t
ρ	Step part of the toll. In a step toll scheme, 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	Travel cost for a type *i* user. It is the sum of the queuing time cost plus the schedule delay cost.
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. Total capacity cost is $k \cdot s$.
MEC_i	Marginal external cost of a type *i* user. It 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 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	(Generalised) 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.
t^-	Moment when the step part of the toll is turned off.
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_i	Total travel cost of type *i*: $TC_i = N_i \cdot c_i$
TC	Total cost including capacity cost: $TC=TC_1+TC_2+k \cdot s$
W	Welfare equals Consumer benefits of the two types, B_i, minus total cost: $W=B_1 + B_2 - TC_1 - TC_2 - k \cdot s$
w_i	The weight attached to type *i*'s marginal external cost in the toll rule
Π	Profit, which equals toll revenue, TR, minus 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}	Centre peak period with a step toll when the toll equals $\mu+\rho$. It lasts from t^+ to t^-.

2. Extended literature review

The self-financing results 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 degree one in capacity (i.e. doubling capacity doubles capacity costs), and iii) that per car travel cost only depends on the ratio of number of cars and capacity (i.e. doubling both usage and capacity leaves travel cost unchanged). 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 road 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., Kraus (1982), D'Ouville and McDonald (1990), Lindsey and de Palma (2014), Lu and Meng (2017), and Fu et al. (2018)), that input prices may vary with the amounts used (e.g., Small (1999)), and to considered networks (e.g., Yang and Meng (2002)).[5]

These papers focused on static congestion, the extension to dynamic congestion was introduced by Arnott et al. (1990, 1993) and Arnott and Kraus (1993, 1995, 1998a). The bottleneck model is the work horse 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 important component of our setting as it can cause the self-financing result to break down. It was first introduced to the bottleneck model by Vickrey (1973) and extended by Newell (1987), and Arnott et al. (1988, 1990, 1993) and many others. Van den berg and Verhoef (2011b) introduced the distinction between ratio heterogeneity and proportional heterogeneity. Ratio heterogeneity means that there is heterogeneity in the ratio of value of time to value of schedule delay, and 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, and it affects the outcome under fully time variant and step tolling (Van den Berg, 2014). Later papers such as Liu et al (2015), Chen et al. (2015a) and Hall (2018, 2021ab) have looked at many more multiple dimensions of heterogeneity.[6]

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)—i.e., how the difference between marginal social cost and private travel cost—where this MEC is constant under flat tolling.

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 road, lanes and bridges in the USA. Various models have been proposed that differ in how the ensure that the generalised price is constant over time. These include the ADL model (Arnott et al.,1990, 1993), 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 Ren et al. (2016) and Li 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. Xiao et al. (2011) added proportional heterogeneity to the ADL step toll model.

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

[6] Hall (2021ab) shows that in particular adding heterogeneity in the preferred arrival time to ratio and proportional heterogeneity has large effects. Conversely, Arnott et al (1988, 1994) found that if there is only heterogeneity in the preferred arrival time, and not also other heterogeneity, this does very little.

Under homogeneous users and bottleneck congestion, a fully-time-variant congestion toll will remove all travel delays and leaves the generalised prices unchanged. Xiao et al. (2011) find that their ADL step toll lowers generalised prices. This also occurs with homogeneity due to the Mass departures (Lindsey et al, 2012), but this is strengthened by the proportional heterogeneity which makes tolling more beneficial for users (Van den Berg and Verhoef, 2011a). 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 and preferences.

Finally, there is a small literature that investigates coarse tolling under other dynamic congestion models. Chu (1999) used his dynamic flow congestion model to study a flat and a single-step toll. Börjesson and Kristoffersson (2012) used their model for the Stockholm step toll system. Zheng et al. (2015) used a macroscopic fundamental diagram to model a flat cordon change for the city of Zurich. Ge and Stewart (2010a,b) and Ge et al. (2016) used a cell transmission models to study step tolling.

3. Model Set-up

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 easy of presentation. A type is defined by its users having the same preferences for travel time and schedule delay. Within a type there will be differences in willingness to pay for the trip, as for each type there is an independent price sensitive demand.

We will not go into details on how our bottleneck model works. See Small and Verhoef (2007), Small (2015) and Li et al. (2020) for overviews.

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

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

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. The α_i is the value of time (VOT) for type i. We consider bottleneck congestion, and normalize the free-flow travel time to zero.[7] Travel time equals the number of cars in the queue before reaching it divided by the capacity s. The queue is assumed to be at a single point.

[7] 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.

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

$$TC = k \cdot s + \sum N_i \cdot c_i. \qquad (2)$$

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 users arrives; the price is no lower on all moments that there are no arrivals of type i users:

$$D_i[N_i] = P_i \equiv c_i[t] + \tau[t] \qquad (3)$$

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

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

$$W = \sum B_i - TC\,. \qquad (5)$$

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

Now, we briefly discuss the general outcome when the coarse toll equals the (potentially heterogeneous) marginal external costs (MECs) throughout the peak. The results directly follow from Arnott and Kraus (1995) and Van den Berg and Verhoef (2011a,b).

Lemma 1: First-best capacity

Consider M discrete types of heterogeneous users, were the travel cost, $c_i[N_1, .., N_M, s]$, of type i increases with the number of users of the different types, N_j, and decreases with the bottleneck capacity, s. Suppose that the toll equals the marginal external cost (MEC) at all moments. Then, the first-best optimal capacity minimizes the total cost from (5) and is set by the following ***first-best*** *condition: $-\sum N_i \cdot \partial c_i/\partial s = k$.*

Proof: The social optimal toll optimizes the number of users of each type, even if the toll is flat or has a single step. Therefore, maximizing welfare to capacity, s, is equivalent minimizing total cost, as consumer benefit is solely determined by the numbers of users. The f.o.c. of minimizing total cost to s is $k + \sum N_i \cdot \partial c_i/\partial s = 0$. 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. This we will show for ratio heterogeneity later on. Although we do not study this, without tolling, the capacity rule is adjusted downward to limit latent demand, as Small and Verhoef (2007) show for static congestion.

Lemma 2: Self-financing

With a total cost minimizing capacity,[8] a ***sufficient condition*** *for exact self-financing—i.e. that toll revenue equals road capacity cost—is that the (coarse) toll equals the marginal external cost (MEC_i) of a type i at each moment that a type i user travels.*

Proof: See Kraus and Arnott (1995, p. 279-280), who proof this for any dynamic congestion model with total cost that are homogeneous to the degree zero in number of users of each type and *s*, which is true with bottleneck congestion.[9] □

Even if a second-best optimal toll differs from the marginal external cost at some point in time, 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 set-up as used throughout this paper. It also presented previous results in our notation which will prove useful 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. The equations for toll and capacity setting will turn out very complex. Adding more realism with many types or multiple dimension of heterogeneity would make analytical analysis difficult if not impossible.[10] With flat tolling and without any tolling, the travel costs, MECs and MSCs are constant over time.

With a flat toll, the toll equals μ throughout the peak:

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

A flat toll leads to the same user equilibrium as no toll at all. So we can use the cost functions from Arnott et al. (1988) and Van den Berg and Verhoef (2011a,b):

$$c_L^F = \delta \frac{N_L + \frac{\alpha_L}{\alpha_L} \cdot N_H}{s}, \qquad (6a)$$

$$c_H^F = \delta \cdot \frac{N_L + N_H}{s}; \qquad (6b)$$

[8] As noted, throughout this paper, we focus on bottleneck congestion and capacity costs that follow $k \cdot s$.

[9] To directly see this in their proof, remember that a toll that equals MEC_i means that a typei's inverse demand, D_i, equals their marginal social cost.

[10] Of course, numerical analysis for specific parameters would still be possible, but this would not give general insights.

with $\delta=(\beta+\gamma)/(\beta\cdot\gamma)$ being a compound preference parameter. A subscript *L* indicates the Low-VOT type who has the lower value of time. The *H* is for the High-VOT type who cares relative more about travel time than schedule delay. Hence, 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 centre peak. This leads to a total cost of

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

where superscript ^{F} indicates the flat toll equilibrium.

Marginal external cost,[11] 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 MEC_L of the Low-VOT users exceeds the MEC_H of the High-VOT users:

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

The difference in MECs depends on the ratio of values of time, and 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:

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

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 & the flat toll

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

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

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

[11] Users are assumed to be selfish and hence only consider their own travel cost.

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

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. In optimum, $\lambda_L^F = -\lambda_H^F$, and so 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 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:

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

where $\lambda_L^F > 0$ is the multiplier for the Low-VOT user-equilibrium in (10). Since $\lambda_L^F > 0$ and $\left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right) > 0$, for given number of users, the capacity with a flat toll is set higher than following the first-best capacity rule (which is $-\sum N_i \cdot \partial c_i^F/\partial s = k$).

Proposition 2: Self-financing with a flat toll and two-type ratio heterogeneity

With a second-best flat toll and capacity, the profit is

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

where w_L^F is the weight of the Low-VOT users in the toll rule of Proposition 1 and $f_L = N_L/(N_L+N_H)$ 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. To be done and then in Appendix A.□

Proposition 1 implies that the volume-capacity ratio with flat tolling is higher 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 over pricing the High-VOT users—as they see a toll above their MEC_i—but raises distortion from underpricing Low-VOT users. Nevertheless, the second-best addition raises welfare as the first effect is stronger.

With ratio heterogeneity, it is impossible to say explicitly when there will be a profit and when a loss. In eq. (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.[12] Hence, with the second-best capacity, the flat toll must well 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).[13] 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 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. Typically, 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 has a single step in it. Of type i, $V_i\geq 0$ users travel in the centre period when 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:

$$\tau[t]=\begin{cases}\mu+\rho & \text{if } t^-\leq t\leq t_2\\ \mu & otherwise\end{cases}.$$

[12] 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.

[13] In the profit equation there is a direct effect of $-\partial D_L/N_L$, but a smaller $-\partial D_L/N_L$ also raises the Low-VOT type's weight, which raises the toll and thereby further raises profit

With single-step tolling, we find that travel costs and MECs are constant with a period, but differ between periods. Travel costs are lower in the centre peak period, whereas MECs are higher.

For simplicity, we only consider a single step in the toll and use the Laih (1994) equilibrium model. Obvious extensions would be multiple steps and alternative equilibrium models. See Lindsey et al. (2012) for an overview. The analysis would be more complicated in these models, as the exact distribution of user types over time is uncertain and costs depend on the preferences and number of users of each type (Van den Berg, 2014). Nevertheless, also in these models, the results would be similar as here, since costs would be similar in magnitude, and depend in a similar way on the preferences, capacity and numbers of users of each type.

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 last from t_s until t^+ at which time the toll is increased. The late shoulder period lasts from t^- to t_e. The below results follow from Van den Berg (2014). Therefore in text, we will only summarise the results, Appendix B gives a detailed derivation under two-type heterogeneity.

High-VOT users are less willing to queue since they value travel time more. Hence, both in the shoulder periods as in the centre period, the High-VOT users will arrive further from t^* when travel times are lower 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 centre as in the shoulder periods. The optimal step part of the toll minimizes the total travel cost for given numbers of users. For this, its level, ρ, equates the generalized prices in the centre 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 centre period: $V_i=N_i/2$. This allows us to write total cost as

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

where superscript ^{SS} indicates the single-step toll equilibrium. The costs per trip simplify to

$$c_L^{SS} = \begin{cases} c_L^{cp} = \dfrac{1}{2}\delta\dfrac{N_L + \dfrac{\alpha_L}{\alpha_H}\cdot N_H}{s} & \text{if } t^- \le t \le t^+ \\ c_L^{sh} = \dfrac{1}{2}\delta\dfrac{N_L + \dfrac{\alpha_L}{\alpha_H}\cdot N_H}{s} + \dfrac{1}{2}\delta\dfrac{N_L+N_H}{s} & \text{if } t_s \ge t > t^- \text{ or } t^+ \ge t > t_e \end{cases} \qquad (15a)$$

$$c_H^{SS} = \begin{cases} c_H^{cp} = \dfrac{1}{2}\ \delta\dfrac{N_H+N_L}{s} & \text{if } t^- \le t \le t^+ \\ c_H^{sh} = \delta\dfrac{N_H+N_L}{s} & \text{if } t_s \ge t > t^- \text{ or } t^+ \ge t > t_e \end{cases} \qquad (15b)$$

Here, superscript ^{cp} indicates the centre peak period and ^{sh} the shoulder periods.[14]

[14] Van den Berg (2014) found basically the same for *M*-type ratio heterogeneity in the Laih step toll model. For the ADL and braking model, the V_i is different for each type and is a function of the N_i and preference parameters.

The marginal external cost of the Low-VOT users exceeds that of the High-VOT users in each period. The differences in MECs between the centre peak and shoulder periods are the same for both types:

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

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

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

The ρ equals the difference in cost between the centre peak and shoulder period, and this turns out to also be the difference in MEC between these periods:

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

Using all this, we can show that maximizing welfare is equivalent to maximizing the following Lagrangian:

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

The user-equilibrium multiplier, λ_i^{sh}, ensures that type *i*'s sum of travel cost and toll equals its inverse demand, D_i. Problem (18) is akin to problem (13) 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 number of users.[15]

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 centre peak and shoulder periods of both types. The flat part of the toll, μ, balances the underpricing of Low-VOT users with the overpricing of High-VOT users:

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

where the weight of type L is smaller than in the flat toll case:

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

Just as with the flat toll, the step toll balances the overpricing of the High-VOT users, with the underpricing of the Low-VOT users. The toll generally differs from the average *MEC*

[15] We get the same analytical results if we do the full problem at once. Moreover, for the numerical model, we diricly 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.

(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 cost for type i averaged over time[16] *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 $\lambda_L^{sh} > 0$ is the Low-VOT type's multiplier from (18), w_L^{SS} its weight in the toll rule from (20) and $f_L = N_L/(N_L+N_H)$. [17]

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

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

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

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

where 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 loss would be twice as large with the flat toll (F) as with the step toll (SS). However, we will see that the difference will tend to be somewhat smaller. The second term in (23) will be somewhat smaller

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

[17] 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.

than 1. This is because: the volume capacity-ratio, $\left(N_L^j + N_H^j/s^j\right)$ 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.[18] Hence, the profit or loss with a step toll will tend to be somewhat more than half that of the flat toll.

Eq. (23) also indicates that the two tolling forms will 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 $\partial D_L/N_L$ 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 and thus α_L approaches α_H. As the step toll has the smaller second-best capacity expansion, it is slightly 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 value of time to value of schedule delay. This heterogeneity has also been called flexibility heterogeneity. With a flat toll, the toll is a constant throughout the peak. With a step toll, the toll is μ in the early and late parts of the peak; in the centre peak, it is higher and is $\mu + \rho$.

Both the flat toll as the flat part of 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. The general formulas of the two tolls are the same, but, as equilibrium cost functions differ, they will result in different tolls. The step toll lowers travel costs and MECs for the same numbers of users by reducing queuing, and hence also the flat part of the toll is lowered.

For given numbers of users, the capacity is set higher with the flat toll than the step toll, which in turn has a higher capacity than under the first-best fully-time-variant toll. This is done to limit the welfare reducing effects of overpricing High-VOT users, where this issue is most 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

[18] This follows from three points.

Point 1, N_L^F/s^F *vs* N_L^{SS}/s^{SS} is uncertain, as $\left(N_L^F + N_H^F\right)/s^F < \left(N_L^{SS} + N_H^{SS}\right)/s^{SS}$ and $f_L^F > f_L^{SS}$ as step tolling is relatively more detrimental for Low-VOT users. But on the whole, we 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 direction and mostly cancel 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 oppositive directions and will mostly cancel out.

sensitive or the High-VOT users are less price sensitive. The profit or loss goes towards zero as the degree of heterogeneity goes to zero. As the step toll has the 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 over twice that of the step toll.

5. Numerical model for ratio heterogeneity

Now we turn to our numerical model. It will illustrate our analytical model. It also studies the effects of imposing self-financing, which we cannot do with pure analytics. The sensitivity analysis studies: (i) How likely it is that there will be a loss? (ii) How important is the potential lack of self-financing? (iii) How sensitive is the outcome to parameter values? (iv) What is the effect on welfare if we impose that a flat or step toll has to be self-financing?

For comparison and calibration, the numerical model will also look at the outcome with no tolling and with a first-best toll that is fully time variant. This allows us to put the effects of flat and step tolling into perspective.

5.1 Base case calibration

We aim to keep the set-up comparable with Van den Berg and Verhoef (2011a,b) and Van den Berg (2014). The numeral model thus also consider 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).[19] The marginal capacity cost parameter, k, is set such that the first-best capacity is s=3600, which is the fixed capacity in Van den Berg and Verhoef (2011a,b) and Van den Berg (2014). The capacity in the no-toll equilibrium is by assumption also 3600. The VOTs are α_L=€7.50/h and α_H=€15.00/h. This ensures that the no-toll average value of time is €10.00/hour, which 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 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 1.5 that of the High-VOT type, implying that the Low-VOT type is less price sensitive. One of the things the sensitivity analysis will look at is the changing this ratio of demand slopes.[20]

[19] We use fuel cost for the elasticity as there is large empirical literature on this, while little is known for toll payments. We assume that people do not care where they spend money on.

[20] The remaining parameters are k=14.436, $d0_L$=45.889, $d0_H$=24.053, $d1_L$=0.0059641, $d1_H$=0.00397603 and $\delta=\beta\cdot\gamma/(\beta+\gamma)$= 4.8501.

5.2 Results for the base calibration

Table 1 presents the results for the base calibration. The most important results are that both coarse toll schemes have a large loss. The flat toll has a loss of 480 or 12% of capacity cost; the step toll has a loss of 260 or 6.8% of capacity cost. 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 of with step tolling. With the first-best toll, the toll equals the time-variant MEC, and there is a zero profit. With the flat toll, the average MEC is €7.91 and the toll is €7.01. So the flat toll is below the average MEC, while for a zero profit it would need to exceed it. The flat toll is €1.79 lower than the Low-VOT type's MEC_L, and €1.37 higher than the MEC_H. With the single-step toll, the toll is also below the average *MEC* and the MEC_L, and it 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 equals capacity cost, the welfare is only 0.25% lower while the generalized price increases by 7% for the Low-VOT users and 6% for the High-VOT users. With a step toll, the effect of self-financing is minute: it causes a 0.05% fall in welfare and the prices increase 3.4% and 3.7%.

The flat toll is a blunt instrument that can only limit the number of users, without affecting queuing delays otherwise. With fixed demand, it would have a welfare gain of zero. Now it attains 24.8% of the first-best welfare gain relative to the no-toll case, but both types face very large price increases.[21]

The single-step toll removes—for given number of users—halve the queuing compared to the no-toll case. It is a more precise instrument and is less harmful for consumers. Still, both types see an increase in the price, and this increase is smaller for the High-VOT type. The single-step toll attains 58.4% of the first-best gain in our numerical base case. With fixed demand, this percentage would be 50%, both with homogeneity as with ratio heterogeneity (Laih, 2004; Van den Berg, 2014).

[21] For a given numbers of users, the price with a flat toll would be twice that of the no-toll case; but, off course, the flat toll also reduces the number of users tremendously.

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		
					Centre period	Shoulder period	Combined	Centre period	Shoulder period	Combined
Number of Low-VOT users, N_L	6000	5710.1	5251.5	5080.3	2737.2	2737.2	5474.3	2696.0	2696.0	5391.9
Number of High-VOT users, N_H	3000	3073.4	2072.5	1836.0	1260.6	1260.6	2521.1	1201.1	1201.1	2402.2
Travel cost for the Low-VOT type[e]	10.10	5.92[a]	7.55	7.60	4.22	9.23	6.72[a]	4.23	9.24	6.74[a]
Travel cost for the High-VOT type[e]	12.13	5.92 [a]	8.80	8.77	5.01	10.02	7.51[a]	5.00	10.01	7.51[a]
(Average) MEC_L	**12.13**	**5.92[a]**	**8.80**	**8.77**	**10.02**	**5.01**	**7.51[a]**	**10.01**	**5.00**	**7.51[a]**
(Average) MEC_H	**8.08**	**5.92[a]**	**5.64**	**5.55**	**8.30**	**3.29**	**5.80[a]**	**8.27**	**3.27**	**5.78[a]**
(Average) toll	**x**	**5.92[a]**	**7.01**	**7.99**	**9.02**	**4.01**	**6.52[a]**	**9.50**	**4.49**	**7.00**
Step part of the toll	x	x	x	x	5.01			5.00		
Flat part of the toll	x	0	7.01	7.99	4.01			4.49		
Toll revenue	x	51970	51372	55232	52098			54524		
Capacity cost of $k{\cdot}s$	51970	51970	58279	55232	55882			54524		
Profit	**x**	**0**	**-480.41**	**0**	**-263.23**			**0**		
Profit as percentage of capacity cost	**x**	**0%**	**-11.9%**	**0%**	**-6.8%**			**0%**		
Capacity, s	3600[c]	3600	4037.1[c]	3826.0[c]	3871.0[c]			3776.9[c]		
$(N_L+N_H)/s$: volume-capacity ratio	2.50	2.44	1.81	1.81	2.07			2.06		
Total travel cost cost[e]	97003	51970	57907	54727	71451			69694		
Welfare, W	73275	116007	83872	83665	98218			98169		
Relative efficiency [b]	**0[b]**	**1[b]**	**0.248[b]**	**0.243[b]**	**0.584[b]**			**0.583[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, with 4516.3, be much higher than the in the first-best case, but using this level would complicate comparison with previous papers who used a fixed capacity. The second-best capacity is higher because the no-toll case has slightly more users and mostly because, for a given N_L and N_H, the no-tolling user costs are almost twice that of in the FB case, making capacity building more attractive. Finally, the no-toll case would have a second-best capacity reduction, which corrects for latent demand due to the unpriced congestion (see also Small and Verhoef (2007)). The flat and step toll settings have a small second-best capacity increase compared to the first-best rule, so that they have a slightly lower volume-capacity ratio.

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

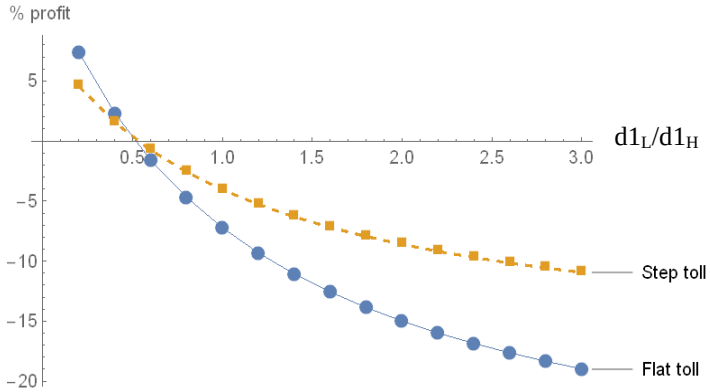
5.3 Sensitivity analysis

We now turn to our sensitivity analysis. The most important one is 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 with looking at the ratio of demands $d1_L/d1_H$. In the base calibration, this ratio was 1.5, and so the Low-VOT type was less price sensitive than the High-VOT type. Fig. 1 shows that, as the Low-VOT users become relatively more price sensitive, the profit falls for both schemes. For both of them, the Low-VOT users would need to be about twice as price sensitive as the High-VOT users in order for there not a loss. As we argued using eq. (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 a bit more likely not to make a loss.

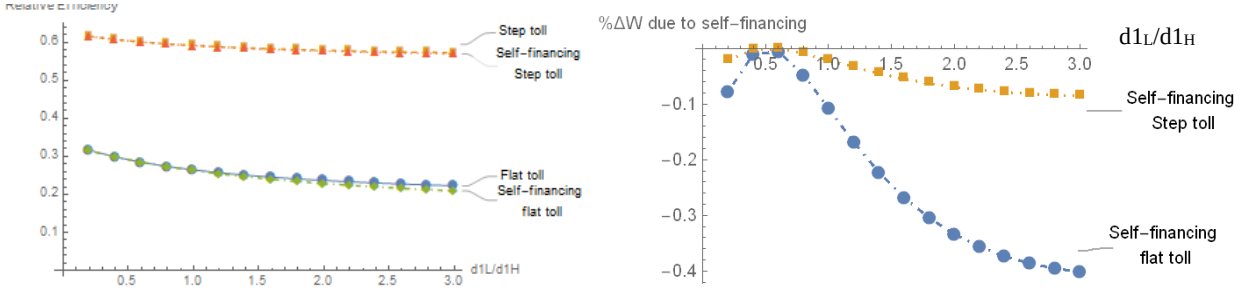
Fig 1: Profit as a percentage of capacity cost over the ratio, $d1_L/d1_H$, of function demand slopes.



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, and thus attains about 60% of the welfare again from the no-toll case that the first-best policy attains. 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 perform decreases slightly. This probably occurs because both policies do not remove all queueing and partly reduce congestion by reducing consumption, which is more difficult when the Low-VOT type—which is more numerous—is less price sensitive.

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 get self-financing; 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. The self-financing has a more noticeable effect on prices, number of users and consumer surplus. For instance, the number of users can fall by up to 10%.

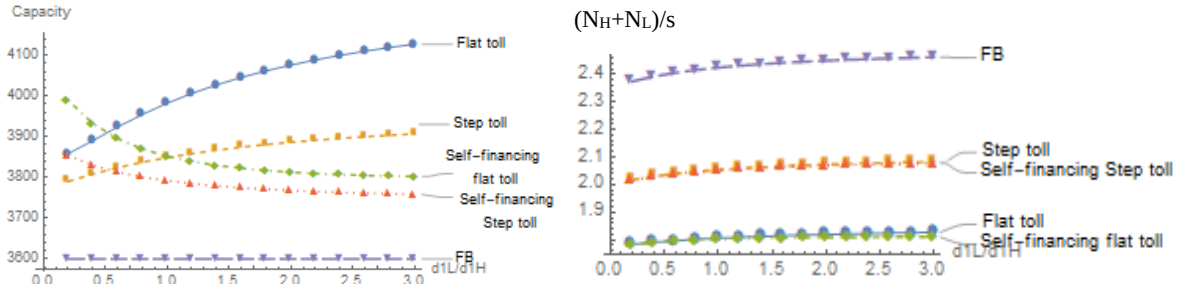
Fig 2: Welfare effects and the ratio $d1_L/d1_H$ of demand slopes: Left panel relative efficiency, right panel: welfare loss due to imposing self-financing



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.

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. First-best toll removes all queuing, and step tolling halves the queueing for given usage numbers. So the travel costs is much higher with the flat toll than with the step toll, who in turn has higher travel costs than the first-best toll. Higher travel costs, for given usage, is a second reason for more capacity and thus a lower volume-capacity ratio.

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



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: $\%\Delta\alpha$. With homogeneity, the coarse toll equals the homogenous *MEC* and, as Fig. 4 shows, profit is exactly zero. As the degree of heterogeneity rises, the *MEC* becomes more heterogeneous and the loss increases.

The degree of heterogeneity has little to no effect on the welfare effect of the step toll. With fixed demand in Van den Berg (2014), 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 mostly 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 %Δα. Still, for the maximum amount of heterogeneity at which the α_L is only just above β,[22] the welfare loss is still only 0.5%.

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

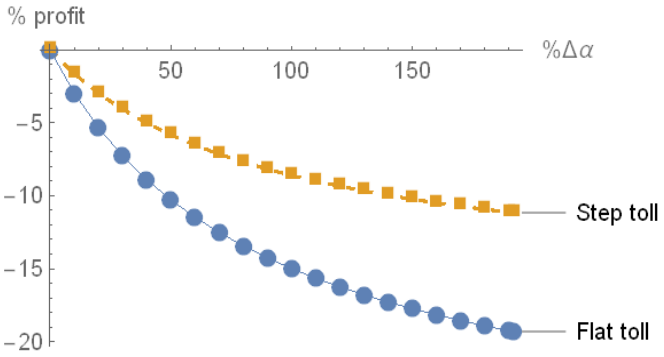
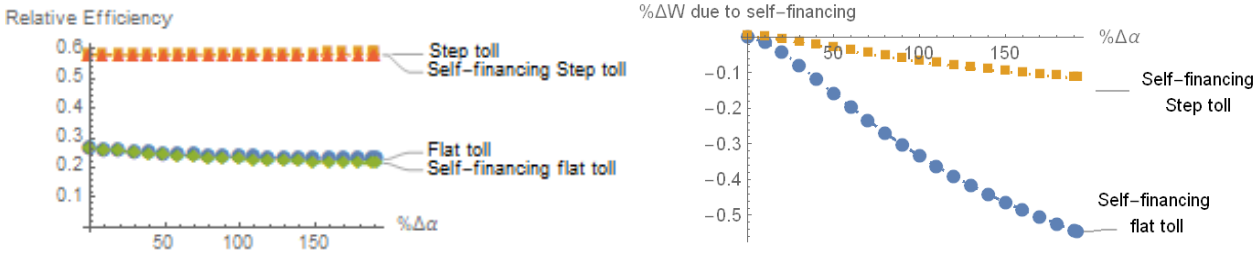


Fig 5: Welfare effects and the degree of heterogeneity (%Δα)

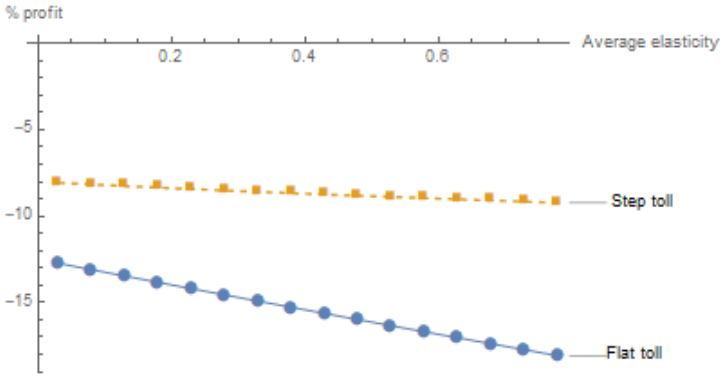


Note: The relative efficiency of a policy is its welfare gain from the no-toll equilibrium divided by the corresponding welfare gain of the first-best social optimum.

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 large empirical literature on this, while less is known for toll payments. Fig. 6 indicates that as demand becomes more elastic, losses are larger in percentage terms and per passenger; although total losses are smaller. With the larger loss in percentage terms and per traveller, self-financing thus needs a larger toll increase, which makes it more harmful for welfare.

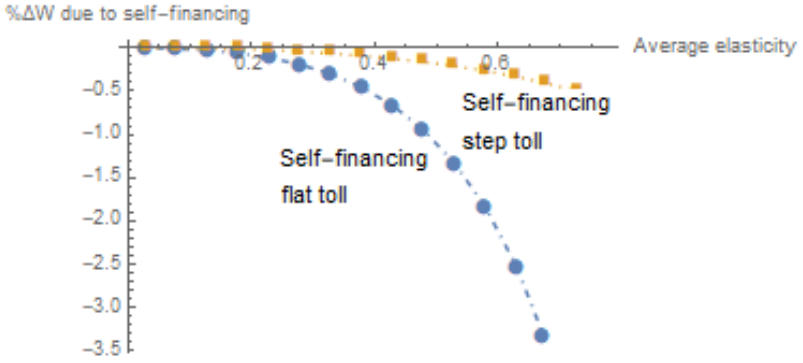
Fig 6: The effect of the average fuel cost elasticity on profit (as a percentage of capacity cost)



Note: we use the absolute of the elasticity, so that a larger number implies more elastic demand

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

Fig 7: Welfare effects and the average fuel cost elasticity



This concludes, our sensitivity analysis. Appendix C gives some further analysis. To summarise, it is most likely that a coarse toll will make a loss. The 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, so that the toll covers the extra capacity cost due to the second-best capacity rule. As overall demand becomes more price sensitive or the degree of ratio heterogeneity increase, the deficit with a coarse toll rises, and the welfare decrease due to imposing self-financing also rises.

6. Proportional heterogeneity

Ratio heterogeneity is 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 proportions. As Vickrey (1973) and Van den Berg (2011a) argued, this heterogeneity could stem from income differences. This section studies proportional heterogeneity with two user types: a low-values type and a high-values type, where all values are an arbitrary percentage higher for the high type. Again, there are separate demands for each type.

6.1 Flat toll

We again first turn to the flat toll. We will not go into details as the equilibrium is the same as the no-toll equilibrium in Van den Berg and Verhoef (2011a). The flat toll cannot remove queueing; it can only remove the persons from the road who have a value of the trip that is below the marginal social cost.

Proposition 5: Proportional heterogeneity and flat tolling

With proportional heterogeneity, the marginal external costs (MEC's) 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 will look at the more general case with M discrete type. Using Van den Berg and Verhoef (2011a), total cost with *M* types of users is:

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

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:

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

as the terms in the summation are independent of what type type *j* is. □

As the toll equals the homogeneous MEC with proportional heterogeneity, Lemmas 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 that exactly equals 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). The analysis proves to be a bit more difficult as there are three possible equilibria depending on the relative sizes of the high-values and low-values types. However, these equilibria only differ in whether they are fully or partially separated. We assume that each type is of positive size, $N_H>0$ 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 centre peak. Off each type, $V_i \geq 0$ users travel in the centre peak and the remaining $N_i - V_i$ in the shoulder periods. In the fully separated equilibrium, all high-values users travel in the centre peak and all low-values users in the shoulder periods. In the two partially separated equilibria, one type travels in both the centre period and the shoulder period, while the other type uses only one.

Lemma 5. Possible user equilibria with proportional heterogeneity and single step tolling
With two-type proportional heterogeneity, the Laih-single step toll has three possible user equilibria depending on the relative sizes of the user types. The equilibria differ in which type uses which period(s):

i. Partially separated equilibrium with $N_L>V_L>0$ and $V_H=N_H$. This occurs if $\delta_L N_L > \delta_H N_H$.
ii. Fully separated equilibrium with $V_L=0$ and $V_H=N_H$. This occurs if $\delta_H N_H \geq \delta_L H_L$ and $N_L \geq N_L$.
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 and single-step tolling, for all three possible user equilibria, the toll $\tau[t]$ equals the $MEC_i[t]$ when type i travels. Consequently, following Lemmas 1-2, the system is self-financing with the toll revenue equaling the cost of the optimal capacity.

Proofs of Lemma 5 and Proposition 6: Lemma 5 follows directly from Van den Berg (2014). The MEC_i is constant within travel period and the same for all type, but now does differ between the shoulder periods and the centre peak. Just as with ratio heterogeneity, the difference in toll between the toll in the shoulders

and the centre 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 Lemma 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 capacity that follows the first-best rule. Van den Berg (2014) considered continuous heterogeneity, which turns out to be easier as there is only one possible equilibrium. Again, the MEC in the centre period is the same for all types and the shoulder periods' MEC is also the same for all types. When there are many types m, multiple equilibria are possible, but also then the higher values types use the centre period and there will be at most one type that uses the centre and shoulders.

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).

Heterogeneity in β/γ_i means that also α/γ_i 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 this means that self-financing is not ensured.

Following Arnott et al. (1989) heterogeneity in t_i^* means that user 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 also be the same for all types, so one can expect self-financing to hold as the toll will equal the MEC.

We do not look at general heterogeneity in multiple dimensions.[23] Such heterogeneity is off course present in really, but complicates analytical analysis of step tolling tremendously due to the explosion of the number of possible equilibria. Yet, using the results of van den Berg (2011a) and Hall (2018, 2021ab), 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. If t* is heterogeneous, we add the extra complication of versus marginal vs infra-marginal users, where infra-marginal users can only arrive at one moment and time but their preferences do not affect the equilibrium. Then the marginal users would probably determine the toll, while one would expect the capacity to depend on the average users; this would imply a second-reason why self-financing may not hold.

8. Conclusion

We studied whether a road with bottleneck congestion is self-financing under flat or step tolling; that is whether the toll revenue cover the bottleneck capacity costs. Self-financing will always hold if the toll can equal the marginal external cost (MEC) throughout the peak. But, with ratio

[23] We do not consider time-variant values of time or schedule delay and possible heterogeneity therein.

heterogeneity between 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. For this result, we derived explicit formulas for toll and capacity setting under ratio heterogeneity with two discrete types of users. We find that the capacity rule has a second-best correction: the capacity is set higher than following the first-best rule as this lessens the distortion from overpricing the drivers with high values of time. Users with a lower VOT cause higher MECs. The 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 more price sensitive a type is, the closer the toll is to its MEC.

The scheme can only be self-financing if the users with low value of time are much more price sensitive than those with high values, and a deficit is most likely to occur. In our numerical model, the users with a low value of time must be almost twice as price sensitive for there not to be loss; and, typically, the loss is 5-15% of capacity costs. Nevertheless, imposing self-financing by adding an extra constraint that toll revenue equals capacity costs only causes a small welfare loss of 0% to 1.5%. Other dimensions of preference heterogeneity do not in themselves lead to a violation of the self-financing result.

All this shows that in reality it unlikely that coarse tolling systems will be exactly self-financing. This unfortunate for the acceptability of congestion pricing. Still, we find that the effect on welfare of adding a self-financing constraint are small, and even minute with a step toll.

An obvious follow-up question is how a multi-step toll—as used in Singapore and Stockholm—would perform. Building on the results of Lindsey et al. (2012) for homogeneous users, we would expect the step toll to approach the time-variant MEC as the number of steps goes to infinity. This is true for the Laih and ADL step toll models, but not for the braking model. Another interesting follow up area is considering multiple dimensions of heterogeneity or different congestion models. In particular, Hall (2021ab) has shown that adding heterogeneity in the preferred arrival time to heterogeneity in values of time and schedule delay can massively change the effects of fully-time-variant tolling. Finally, for any dynamic model, the system will always be self-financing if the toll can always equal the MEC. But this is not the case with ratio heterogeneity, and then 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?

Acknowledgements

We are thankful for the comments of the participant of the ITEA conference in 2018 in Hong Kong and at the Eureka seminar in Amsterdam. Any remaining errors are ours.

Appendix:

A. *Detailed derivations for the flat toll under with ratio heterogeneity*

A.1 Proof of Lemma 3

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

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

The first order conditions to the choice variables are:

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

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

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

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

$$\frac{\partial L}{\partial \lambda_L^F} = 0 = \left(c_L^F + \mu^F - D_L\right) \quad (A.6)$$

$$\frac{\partial L}{\partial \lambda_H^F} = 0 = \left(c_H^F + \mu^F - D_H\right) \quad (A.7)$$

Applying (A.3), we get:

$$\lambda_L^F = -\lambda_H^F \quad (A.8)$$

Using this and (A.2) and (A.7), we can then solve for λ_L^F:

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

$$\lambda_L^F = \frac{MEC_L - \mu}{\frac{\partial c_L^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H} - \frac{\partial c_H^F}{\partial N_H}} > 0 \quad (A.9)$$

Combining (A.1), (A.7) and (A.8), 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.9), we get:

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

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

$$\mu^F = \frac{MEC_L \cdot \left(\frac{\partial c_L^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H} - \frac{\partial c_H^F}{\partial N_H} \right) + MEC_H \cdot \left(\frac{\partial c_L^F}{\partial N_L} - \frac{\partial D_L}{\partial N_L} - \frac{\partial c_H^F}{\partial N_L} \right)}{\left(\frac{\partial c_L^F}{\partial N_L} - \frac{\partial D_L}{\partial N_L} - \frac{\partial c_H^F}{\partial N_L} \right) + \left(\frac{\partial c_L^F}{\partial N_H} - \frac{\partial D_H}{\partial N_H} - \frac{\partial c_H^F}{\partial N_H} \right)} \qquad \textit{(A.10)}$$

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

$$\mu^F = MEC_L^F \cdot w_L^F + MEC_H^F \cdot (1 - w_L^F) \qquad \textit{(A.11)}$$

$$w_L^F = \frac{-\frac{\partial D_H}{\partial N_H} - \frac{\partial c_L}{\partial N_H} + \frac{\partial c_H}{\partial N_H}}{\left(-\frac{\partial D_H}{\partial N_H} - \frac{\partial c_L}{\partial N_H} + \frac{\partial c_H}{\partial N_H} \right) + \left(-\frac{\partial D_L}{\partial N_L} - \frac{\partial c_H}{\partial N_L} + \frac{\partial c_L}{\partial N_L} \right)} \qquad \text{(A.12)}$$

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

A.2 Proof of Proposition 1

The eq. (12) of the proposition follows directly from the f.o.c. in (A.4). That $\lambda_L^F > 0$ is visible in (A.9). 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 the larger cost. So 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 $-\sum N_i \cdot \partial c_i^F / \partial s = k - \lambda_L^F \left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right)$ must imply a higher capacity *s* for given N_L and N_H than the first-best rule that is $-\sum N_i \cdot \partial c_i^F / \partial s = k$. This in turn means that the flat toll has the higher 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, where the second line follow from flogging what the capacity condition from proposition 1 implies *k* must equal:

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

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

$$\Pi^F = \mu^F \cdot \left(N_L + N_H\right) + \left(MEC_L \cdot N_L + MEC_H \cdot N_H\right) - s \cdot \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right) \quad \text{(A.13)}$$

$$= \left(\sum_i \left(\mu^F - MEC_i\right) N_i\right) - s \cdot \lambda_L^F \cdot \left(\frac{\partial c_L^F}{\partial s} - \frac{\partial c_H^F}{\partial s}\right) \quad \text{(A.14)}$$

Eq. (A.14) is the general profit equation given in the proposition, by plugging in the functional forms we get the second line. 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.14) must be negative which only happens when the average toll exceeds the average MEC. This completes the proof of Proposition 2.

B. Timings of the peak with a single-step Laih toll under with ratio heterogeneity

For the (generalized) price not to be higher in the centre period when the toll is ρ higher, the travel time at the start of the centre period, must be lower than for arrival just before is. 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 centre period a bit earlier would reduce travel time of all centre peak users, whilst all users that initially travelled in the shoulders would equal off, thereby lowering total cost. Having a period without arrivals just before t^+ would only raise costs, so the queuing of early shoulder period ends exactly at t^+.

The question then arises how t^- is set. Arrivals just after t^- pay a much lower toll than arrivals just before it. 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 besides the road just before the bottleneck without impeding other drivers. There are hence separate queues. The t^- is then set such that the last centre peak user arrives exactly at t^-. Accordingly, in equilibrium the periods are defined by:

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

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

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

Using the results of van den Berg (2014) this makes the travel costs for both types:

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

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

Under two-type ratio heterogeneity and a Laih single step toll, it is optimal that of each type halve of its users travel when the step toll is turned on and the other halve when it is off. Still, within each periods, the 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 centre peak period, total cost can be shown to be:

$$TC^{ss} = k \cdot s + \frac{((N_L+N_H)(N_H-V_H)+V_H(V_L+V_H))\delta}{s} + \frac{((N_L-V_L)(N_L+V_H+\frac{(N_H-V_H)\alpha_L}{\alpha_H})+V_L(V_L+\frac{\alpha_L}{\alpha_H}V_H))\delta}{s}$$

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.

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

Figures C.1, C.2, C.3, C.4 and C.5 extend the sensitivity analysis by looking at some effects that were omitted in the main text.

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

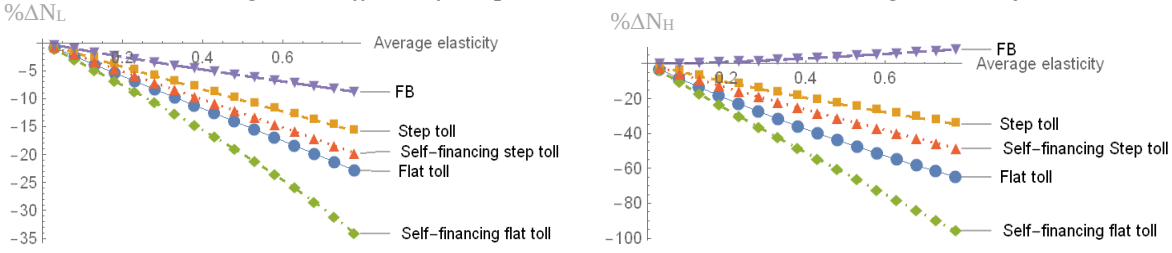


Fig. C.2: Relative efficiencies and the average elasticity

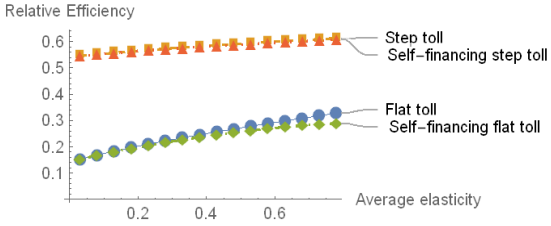
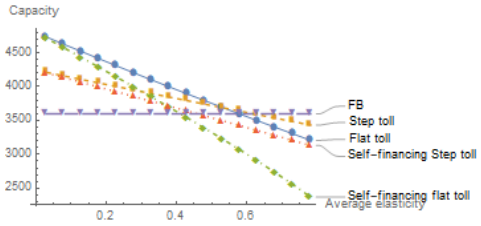


Fig. C.3: The effect of the average fuel cost elasticity on the optimal capacities



Note: we use the absolute of the elasticity, so that a larger number implies more elastic demand

Fig. C.4: Effects of the policies on us · the degree of heterogeneity

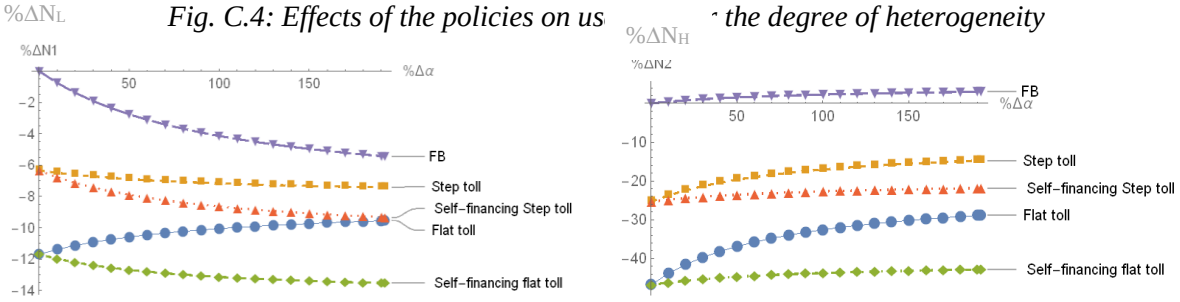
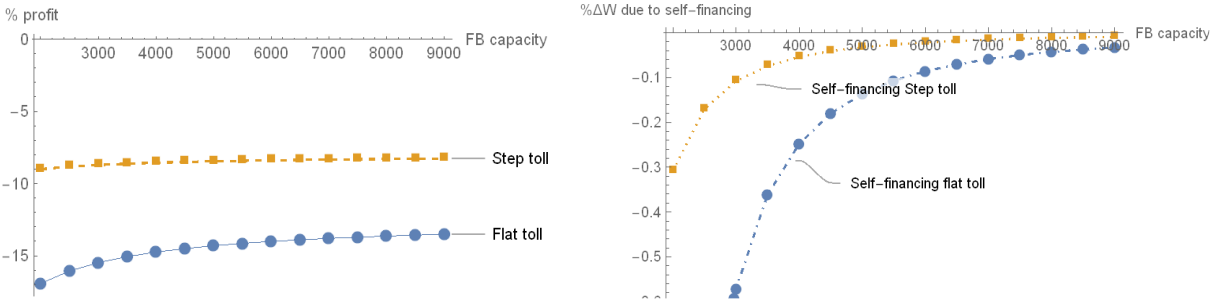


Fig. C.5: Profit and welfare change due to imposing self-financing as the first-best (FB) capacity changes due to changes in the marginal capacity cost.



References

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., 1993. 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., 1994. The welfare effects of congestion tolls with heterogeneous commuters. Journal of Transport Economics and Policy 28 (2), 139–161.

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.

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

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.
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., 2010a. Investigating boundary issues arising from congestion charging in a bottleneck scenario. In: Tampre, C., Viti, F., Immers, L.H.(Eds.), New Developments in Transport Planning: Advances in Dynamic Traffic Assignment. Edward Elgar, Aldershot, pp. 303–326.
Ge, Y.E., Stewart, K., 2010b. Investigating boundary issues arising from congestion charging. In: The Fifth Travel Demand Management Symposium, Aberdeen, Scotland, UK, 26–28 October 2010
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.
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. (2021a). 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. (2021b). Inframarginal Travelers and Transportation Policy. SSRN working paper, 3424097.
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. (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, C.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. (2015). A semi-analytical approach for solving the bottleneck model with general user heterogeneity. Transportation research part B: methodological, 71, 56–70.
Lu, Z. and Q. Meng (2017). Analysis of optimal BOT highway capacityand economic toll adjustment provisions under tra_c demand uncertainty. Transportation Research Part E 100, 17–37.
Mohring H, Harwitz M. 1968. 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. (2012). Valuation of travel time. Economics of transportation, 1(1-2), 2-14.
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.
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 PartA 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.
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.

Xiao, F., Qian, Z.S. and Zhang, H.M., 2013. Managing bottleneck congestion with tradable credits. Transportation Research Part B: Methodological, 56, 1–14.
Xiao, L. L., Huang, H. J., & Liu, R. (2015). Tradable credit scheme for rush hour travel choice with heterogeneous commuters. Advances in Mechanical Engineering, 7(10), 1687814015612430.
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 anagent-based traffic model. Transportation Research Part A 46 (8), 1291–1303.