TRANSPORT
ISSN 1648-4142 / eISSN 1648-3480
2018 Volume 33(1): 77–91
doi:10.3846/16484142.2015.1062048

INVESTIGATING BOUNDARY EFFECTS OF CONGESTION CHARGING IN A SINGLE BOTTLENECK SCENARIO

Ying-En Ge[1,2], Kathryn Stewart[3], Yuandong Liu[4], Chunyan Tang[5], Bingzheng Liu[6]

[1]*College of Transport and Communications, Shanghai Maritime University, China*
[2,4,5,6]*School of Transportation and Logistics, Dalian University of Technology, China*
[3]*Transport Research Institute, Edinburgh Napier University, United Kingdom*

Submitted 29 October 2013; resubmitted 16 July 2014, 10 October 2014; accepted 23 October 2014; published online 13 July 2015

Abstract. Many congestion charging projects charge traffic only within part of a day with predetermined congestion tolls. Demand peaks have been witnessed just around the time when the charge jumps up or down. Such peaks may not be desirable, in particular (a) when the resulting peaks are much higher than available capacities; (b) traffic speeding up to get into the charging zone causes more incidents just before the toll rises up to a higher level; or (c) traffic slowing down or parking on the roadside decreases road traffic throughput just before the toll falls sharply. We term these types of demand peaks 'boundary effects' of congestion charging. This paper investigates these effects in a bottleneck scenario and aims to design charging schemes that reduce undesired demand peaks. For this purpose, we observe and analyse the boundary effects utilising a bottleneck model under three types of toll profiles that are indicative of real charging schemes. The first type maintains a constant toll across the charging period, the second type allows the toll to increase from zero to a given maximum level and then decrease back to zero and the third type allows the toll to rise from zero to a given maximum level, remain at this level for a fixed period and then fall down to zero. This investigation shows that all three types of toll profiles can produce greater boundary peak demands than the bottleneck capacity. A significant contribution of this work is that instead of designing an optimal traffic congestion pricing scheme we analyse how existing sub-optimal congestion pricing schemes could be improved and suggest how observed problems may be overcome. Hence, we propose a set of extra requirements to supplement existing principles or requirements for design and implementation of congestion charging, which aim to reduce the adverse consequences of boundary effects. Concluding remarks are made on implications of this investigation for the improvement of existing congestion charging projects and for future research.

Keywords: bottleneck models; congestion charging; boundary issues.

Notes: Current paper is an entirely updated version of our earlier paper:
Ge, Y.-E.; Stewart, K. 2010. Investigating boundary issues arising from congestion charging in a bottleneck scenario, in C. M. J. Tampère, F. Viti, L. H. Immers (Eds.). *New Developments in Transport Planning: Advances in Dynamic Traffic Assignment*, 303–327. http://dx.doi.org/10.4337/9781781000809.00025

In this new version, we have rewritten the introductory section. Secondly, we are proposing a set of extra requirements for road traffic congestion pricing design and implementation in a newly-added section. Thirdly, three new tables have been generated and the data in them are used to support our analysis while four new figures are added to show the advantages of the adopted bottleneck model over the existing ones. Besides, we significantly improved the presentation and analysis of numerical results as well as corrected the errors and typos in the earlier version throughout the paper, including new Figs 1–2.

Introduction

To mitigate urban traffic congestion from the demand side, a number of congestion charging projects have been introduced since the 1970s and are now operating on a permanent basis around the world, including projects in London, Singapore, Stockholm and Milan. These projects charge traffic only within one or a few time periods during the day as opposed to time-varying and continuous (potentially optimal) 24-hour-a-

Corresponding author: Ying-En Ge
E-mail: *yege@shmtut.edu.cn*

day charging schemes. Some only toll a fixed amount for the use of a charging zone or for passing through a charging point during the charging period, such as in London and in Milan. Some set a different fixed toll for each time interval of the charging period, such as in Singapore and in Stockholm. One reported problem for such stepwise congestion charging schemes is that travellers tend to depart earlier or later to avoid congestion charging or to pay less. For example, materials which presented in Report (TfL 2008) shows the traffic flow profiles throughout a charging day, in which the peaks for inbound traffic appear around the start of the charging period and those for outbound traffic appear just after the end of the charging period. In addition, in both Singapore and Stockholm, travellers tend to intentionally choose their times to pass through the boundary of the charging zone or control points so that they can pay less. Hence, demand peaks have been observed just around the time when the charge jumps up or down. These peaks are often caused by intentional traveller behaviour (e.g. speeding up or slowing down or waiting while they are close to the charging zone) so that they could avoid congestion charging fees or pay less (Salmon 2010). Hence, such peaks may not be desirable, in particular when:

- the demand during these peak periods is much higher than available capacities;
- traffic speeding up causes more incidents;
- traffic slowing down or parking on the roadside decreases road traffic throughput.

We term these types of demand peaks 'boundary effects' of congestion charging. By means of our methodology based on the kinematic wave model of traffic flow (e.g. Newell 1993a, 1993b, 1993c; Daganzo 1994, 1995a), this work investigates the boundary effects in a bottleneck scenario with the aim to identify the characteristics of charging schemes that can reduce or remove these undesired boundary demand peaks and to gain new insights into the real-life problem derived from existing congestion pricing projects.

As the fundamental instrument to investigate the bottleneck congestion charging problem, bottleneck models can be grouped into three categories. The first is the point-queue model, which is attributed to Vickrey (1969) and often termed the Vickrey model. In such a model, traffic congestion takes a point-queue form, i.e. cars queuing behind a capacitated bottleneck point without occupying any physical space (which has been recognised to be a key limitation of bottleneck models of this kind). Due to its elegant simplicity, it has subsequently been discussed or used widely in the literature, including Small (1982), Arnott *et al.* (1990, 1993) and Braid (1989). However, this approach is too simplified to investigate boundary issues or many other practical issues in real-life traffic because it only captures a fraction of the interactions in traffic flows. The second category of bottleneck models is to assume that the travel time a traveller would experience traversing a road segment is a function of the departure rate at the time when he or she enters the road segment and that there are no interactions between departure flows at different times or that the outflow rate of a link is a function of the average density over the whole link (Mahmassani, Herman 1984). As pointed out in Newell (1988), 'This model [proposed in (Mahmassani, Herman 1984)], in effect, admits an infinite wave velocity since any input flow has an immediate influence on the output' and also violates the causality principle. Daganzo (1995b) shows that this type of link travel time models can violate the First-In-First-Out (FIFO) queueing principle. This type of models was employed in Dynamic Traffic Assignment (DTA) modelling at network level by applying them to estimate the travel times on each link. Under certain conditions the FIFO principle can be preserved (Friesz *et al.* 1993). Carey and Ge (2003, 2004, 2005) show that, if this functional relation preserves the FIFO queueing principle, it may be acceptable to apply it to a very short road section; otherwise, it may produce an outflow profile quite different from one from the kinematic wave model of traffic flow. In contrast to the above two categories of bottleneck models, Newell (1988) assumes 'that the highway is homogeneous and that the flow $q(x,t)$ at location x at time t depends on the density $k(x,t)$ at the same location and time according to the theory proposed by Lighthill and Whitham (1955a, 1955b) and Richards (1956)' (also known as the Lighthill–Whitham–Richards or LWR model). This approach has been used commonly in the existing practice of DTA modelling (Lo, Szeto 2002; Szeto, Lo 2004; Friesz *et al.* 2011; Carey, Ge 2012; Ge, Zhou 2012). Different from Newell (1988), for the facilitation of model solving these references use a discrete version of the theoretical model of traffic flow that are proposed in Daganzo (1994, 1995a) rather than the characteristic line-based solution method proposed in Newell (1993a, 1993b, 1993c). This current paper treats a bottleneck as a real road segment or link with a limited capacity and assumes that traffic propagates along the bottleneck following the LWR model, which formulates a unique bottleneck model consisting of two parts. The downstream part is a link with a limited capacity and the upstream part is a deterministic point-queue model at the entry of the link; the deterministic queue exists only when the departure rate is greater than the receiving capacity of the downstream part or the queue in the downstream road segment has grown upstream beyond the entry of the link. We use the finite difference approximation technique developed in Daganzo (1995a) to solve the LWR model numerically.

In the bottleneck scenario, we assume that all vehicles only go through the bottleneck once, hence each vehicle is charged once only. It is also assumed that travellers have perfect information on traffic conditions while choosing their departure-times and aim to achieve the least generalised travel costs. The generalised travel cost consists of travel time cost, schedule delay cost and congestion charging toll. The total demand within the time horizon under study is assumed to be known and constant. It is also assumed that traffic satisfies flow conservation, causality and FIFO queueing principles. Further, we consider a preferred arrival time window rather than

a common work start time. These assumptions enable a stable state of a bottleneck scenario to exist, at which all travellers receive identical generalised travel costs.

The investigation in this paper is carried out under three types of illustrative toll profiles. The first defines a constant toll across the charging period, which is termed a 'coarse toll' in the literature (Xiao *et al.* 2011; Van den Berg 2014). A typical example of this is the London Congestion Charging Scheme (TfL 2013) where a driver pays a charge of £10 on the day of travel if he or she is driving within the charging zone between 07:00 am and 6:00 pm, Monday to Friday. The other typical example is the congestion charging scheme in the northern Italian city of Milan where drivers currently pay a fixed amount of 5€ to drive into the congestion charging zone (termed Area C) on weekdays between 7:30 am and 7:30 pm. So, in either of the two cities, drivers only need to pay a fixed amount of charge once and then can drive in and out the charging zone within a day as many times as they want. The second Type of toll profiles allows the toll to increase linearly from zero to a maximum level and then decrease linearly to zero (triangular). For the third type, the toll grows from zero up to a maximum level, remains invariant for a period and then falls back to zero (trapezoidal). The second and third types of toll profiles may be regarded as two abstract versions of the charging schemes implemented in Stockholm (when a vehicle goes through a congestion tax control point, a time-dependent cost will have to be paid but the total amount a vehicle pays on a single day is no more than SEK60. Such a congestion tax is charged on Swedish-registered vehicles that are driven into or out of central Stockholm, Mondays to Fridays during 06:30 am – 6:30 pm; the tax is not applicable on public holidays, on a day preceding a public holiday or during the month of July) and in Singapore (this toll profile in Fig. 1 applies to the control point on Beach Road (16)). As shown in Fig. 1, in Stockholm and Singapore the tolls begin at zero and return to zero over a sub-period of a day and both cities charge the most in the two peak periods during the day; the toll goes down to a non-zero level after the morning peak in Stockholm whereas it falls to zero on Beach Road in Singapore. The scenario presented in this paper corresponds to the morning peak period.

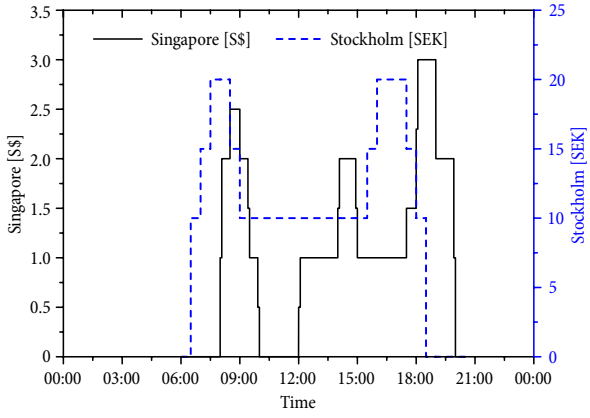

Fig. 1. Staired toll profiles implemented in Stockholm (STA 2013) and in Singapore (LTA 2013)

There are currently a number of papers on coarse and staired toll profiles, both of which are often called step tolls. After a comparison of the two step-toll schemes respectively investigated in Arnott *et al.* (1990) and Laih (1994), Lindsey *et al.* (2012) propose another scheme that allows drivers to 'slow down or stop just before reaching a tolling point, and wait until the toll is lowered from one step to the next step'. The further investigation in this reference shows that each of the three charging schemes is associated with a different optimal toll schedule, which compresses the total social cost to a different degree and that such differences do not vanish as the number of steps tends to infinity. Van den Berg (2012) shows that, given variable demand, the more steps a toll profile has the more the users gain in terms of social welfare. Given varied queueing models for commuters on congested travel-to-work routes, Laih (2004) investigates the optimal single-step and double-step toll structures. It also models 'commuter behaviour where there is an optimal multi-step toll, with regular departure times and time switching decisions'. Xiao *et al.* (2011) show that the optimal coarse toll scheme is Pareto-improving while Xiao *et al.* (2012) find that a coarse toll for managing the morning commuters may induce them to wait tactically at the entry of the bottleneck, which means that the capacity is not utilised sufficiently during certain periods.

There is also much literature that focuses on establishing a set of tolls to achieve a set of system optimal or second-best congestion pricing tolls (e.g. Chu 1995; Doan *et al.* 2011; Tsekeris, Voß 2009; Yang, Huang 1997, 2005; Zhang, Ge 2004) and utilising financial derivatives to facilitate the implementation of congestion charging (Nie 2012; Teodorović *et al.* 2008; Tian *et al.* 2013; Yang, Wang 2011; Yao *et al.* 2010; Xiao *et al.* 2013). Differing from the work on designing or optimising traffic congestion pricing schemes, one contribution of the current work is that we aim to analyse and overcome derived problems of the existing congestion pricing schemes operating across the world by means of analytical techniques, specifically, focusing on the undesired boundary effect problem arising from the implementation and operations of congestion charging. Consequently, this paper is interested in analysing the impacts of the aforementioned three toll profiles on departure rate patterns. Following this section, Section 1 describes the bottleneck scenario and the methodology used in this research. Section 2 presents and analyses demand peaks obtained numerically under the aforementioned three types of toll profiles. Section 3 discusses and proposes a set of extra requirements as supplement to the existing principles or requirements for design and implementation of congestion charging, which aims specially to reduce the undesired boundary effects of congestion charging. Some concluding remarks are given in last section.

1. Bottleneck Scenario and Methodology

The bottleneck scenario under investigation is illustrated in Fig. 2 and D travellers/vehicles are assumed to depart A for D in the time horizon $\left[0,T\right]$, where D is a con-

stant. They all have to travel through a bottleneck, i.e. link from B to C with a limited capacity of q^{max} and length equal to L. Hence, as the demand increases there will be congestion in the bottleneck and the travel time through the bottleneck depends on departure times and congestion states, denoted as $\tau(t,\mathbf{f})$ or, if it does not cause any confusion, τ, where t and $\mathbf{f}=\{f(t): t\in[0,T]\}$ represent departure time and departure rate pattern, respectively. It is also assumed that no congestion takes place on the segment from A to B or from C to D, hence the travel times for the two road segments are constant, denoted as t_{AB} and t_{CD}, respectively. Since travellers are assumed to minimize their travel costs and t_{AB} and t_{CD} are constant, without loss of generality we let both t_{AB} and t_{CD} be zero.

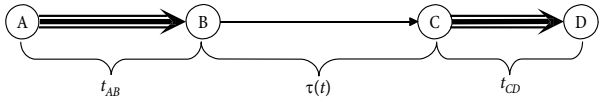

Fig. 2. A bottleneck scenario

Here is a list of assumptions made for the following model formulation:

- *Assumption 1*: Traffic follows the FIFO and causality principles;
- *Assumption 2*: All travellers have perfect information on traffic conditions;
- *Assumption 3*: A traveller chooses to depart at a departure time that incurs the minimum generalised cost.

1.1. Scenario Settings

Charging period and generalised travel costs. We consider a charging period $[t^s, t^e]\subset[0,T]$ and let $C_{toll}(t)$ be the toll (per vehicle) collected at the entry of the bottleneck or at the time a vehicle starts to queue, whichever is earlier. $C_{toll}(t)$ satisfies the following condition:

$$C_{toll}(t)=\begin{cases}\geq 0, & \text{if } t\in[t^s, t^e];\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

and the generalised travel cost a traveller would experience is given by:

$$C_g(t,f)=\alpha\tau(t,\mathbf{f})+C_s(t+\tau)+C_{toll}(t), \tag{2}$$

in which the first term on the Right Hand Side (RHS) of Eq. (2) represents the travel time cost, the second denotes the schedule delay cost and α is the monetary value per unit time.

Preferred arrival window and schedule delay costs. It is assumed that $[w^l, w^r]$ is the preferred arrival window. Suppose that a traveller arrives within the window and then he/she will receive no schedule delay costs; otherwise, due to early or late arrival at D, a penalty $C_s(t+\tau)$ will be incurred, i.e. a schedule delay cost. We adopt the following widely used form of schedule delay costs from the literature (e.g. Hendrickson, Kocur 1981; Braid 1989; Arnott *et al.* 1993):

$$C_s(t+\tau)=\begin{cases}\beta(w^l-(t+\tau)), & 0\leq t+\tau\leq w^l;\\ 0, & w^l\leq t+\tau\leq w^r;\\ \gamma((t+\tau)-w^r), & w^r\leq t+\tau\leq w^l,\end{cases} \tag{3}$$

where: β and γ respectively represent the shadow costs of early and late arrivals and satisfy $\beta\leq\alpha\leq\gamma$.

Travel time generation. Here we use the kinematic wave model of traffic flow to describe the movement of traffic through the bottleneck and a discretisation approximation to the model, i.e. Finite Difference Approximation (FDA) model (Daganzo 1995a) is implemented. We obtain travel times τ by comparing the cumulative traffic at the entry and exit of the bottleneck, represented by $Q(t)$ and $E(t)$ respectively. Specifically, the travel time a traveller departing at time t would experience is equal to the horizontal gap between the two N-curves: $Q(t)$ and $E(t)$ (Ge, Carey 2004; Long *et al.* 2012).

The following quadratic flow-density $(q-k)$ relationship is used in this research:

$$q=\begin{cases}\left(q^{\max}-v^f k^c\right)\left(\frac{k}{k^c}\right)^2+v^f k, & \text{if } 0\leq k\leq k^c;\\ q^{\max}\left(\frac{-k^2+2k^c\left(k-k^j\right)+\left(k^j\right)^2}{k^c-k^j}\right), & \text{if } k^c\leq k\leq k^j,\end{cases} \tag{4}$$

where: $q^{\max}$ represents the capacity of the bottleneck; v^f free-flow speed; k^c critical density; k^j jam density.

Additionally, it is required that $k^c=\frac{2q^{\max}}{v^f}$ hold, which gives $\frac{dq}{dk}=0$ at $k=k^c$. Such a $q-k$ relationship is also used in (Lin, Ge; Carey 2006, Ge 2012; Ge, Zhou 2012).

Feasible set of departure rate patterns. $f(t)$ is subject to the flow conservation and non-negativity constraints, i.e.:

$$\int_0^T f(t)dt=D; \tag{5}$$

$$f(x)\geq 0,\ \forall t\in[0,T] \tag{6}$$

and we have the following proposition:

Proposition 1: By letting $y(t)=\int_0^t f(s)ds$ then the flow-conservation constraint (5) is equivalent to the following two-point boundary value conditions:

$$\dot{y}(t)=f(t); \tag{7a}$$

$$y(t)=0; \tag{7b}$$

$$y(T)=D. \tag{7c}$$

Remark 1: This is a simplified version of network-level two-point traffic dynamics presented at Friesz *et al.* (2008, 2011). Hence, the proof of this proposition is omitted.

All flow patterns satisfying the constraints (6)–(7) compose a feasible set of $\mathbf{f}$, denoted as:

$$\Lambda=\{\mathbf{f}: \text{Eqs}(6)-(7)\ \text{hold}\}.$$

1.2. Methodology

Following the previously-defined notations, we define the equilibrium state of the bottleneck as follows:

Definition 1: Let $\mu = \min\left\{C_g\left(t,\mathbf{f}\right): \forall t \in \left[0,T\right]\right\}$ *and a pattern* $\mathbf{f} \in \mathbf{\Lambda}$ *is in equilibrium if:*

$$C_g\left(t,\mathbf{f}\right) = \mu \quad \text{if } f\left(t\right) > 0; \tag{8a}$$

$$C_g\left(t,\mathbf{f}\right) \geq 0 \quad \text{if } f\left(t\right) = 0. \tag{8b}$$

Remark 2: The equilibrium condition says that a traveller only leaves at a time which produces the least travel costs μ.

When $C_g\left(\cdot,\mathbf{f}\right): \left[0,T\right] \to R^+$ is measurable on $\left[0,T\right]$ for each **f**, following Friesz *et al.* (1993, 2011) the above equilibrium condition (8) can be reformulated as the variational inequality (VI) problem below:

Finding $\mathbf{f}^* \in \mathbf{\Lambda}$ *such that:*

$$\int_0^T \left[f\left(t\right) - f^*\left(t\right)\right] C_g\left(t,\mathbf{f}^*\right) dt \geq 0\,, \ \forall \mathbf{f} \in \mathbf{\Lambda}. \tag{9}$$

The VI problems given in Friesz *et al.* (1993, 2011) are equivalent to a simultaneous route and departure choice Dynamic User Equilibrium (DUE) problem and the above VI problem containing the departure choice only can be regarded as a special case of the VI formulations presented in the two references. For such formulations, Friesz, Mookherjee (2006) and Friesz *et al.* (2011) propose a fixed-point algorithm. Yang and Huang (1997) propose an optimal control variable-demand bottleneck model, which determines the optimal congestion pricing tolls as well as departure rate pattern. These methods all require continuous travel times but this is not always the case in this piece of research work since the coarse toll makes the generalised travel cost function discontinuous at least at both ends of the charging period. To avoid potential discontinuity-related issues, the following pairwise swapping method has been used in our numerical experiments:

Step 0: Initialisation. Divide the time horizon $\left[0,T\right]$ into I intervals of length δ, indexed by $i = 1,2,\ldots,I$. Set the initial departure rate pattern $\mathbf{f^0}$ by letting $f_i^0 = \dfrac{D}{I\delta}$, $i = 1,2,\ldots,I$. Set the iteration number to $n = 0$, the maximum number of iterations to N and the tolerance to ϵ_0.

Step 1: Travel time and cost generation. Load traffic $\mathbf{f^n}$ into the bottleneck, generate the travel times τ_i, $i = 1,2,\ldots,I$, and then compute the generalised travel costs as follows:

$$C_g^{ni} = \alpha\tau_i + C_a\left(t+\tau_i\right) + C_{toll}^i\,, \ \forall i \in 1,2,\cdots,I.$$

Step 2: Time interval pairing and departure swapping. Ranking all C_g^{ni} in ascending order, $i = 1,2,\ldots,I$ gives:

$$C_g^{n1'} \leq C_g^{n2'} \leq \ldots \leq C_g^{nI'}\,.$$

Suppose that the time interval i' is the first one with non-zero departure rate from the right side of the above ordered travel costs. Pair the time intervals numbered from $1'$ to i' in the following manner:

$$\left(1',i'\right), \left(2',\left(i'-1\right)\right), \ldots, \left(\left(\frac{i'}{2}\right), \left(\frac{i'}{2}+1\right)\right) \text{ if } i' \text{ is even;}$$

$$\left(1',i'\right), \left(2',\left(i-1\right)'\right), \ldots, \left(\left(\frac{i'-1}{2}\right), \left(\frac{i'-1}{2}+1\right)\right) \text{ if } i' \text{ is odd.}$$

For each pair, a certain proportion of traffic from the more expensive interval is transferred to the cheaper one. Consider a pair $\left(j',k'\right)$ with travel costs $C_g^{nj'} \leq C_g^{nk'}$ and swap a fraction h_n of the flow $f_{k'}^n$ from k' to j' in the following manner:

$$\begin{aligned} f_{j'}^{n+1} &= f_{j'}^n + h_n f_{k'}^n; \\ f_{k'}^{n+1} &= f_{k'}^n - h_n f_{k'}^n, \end{aligned} \tag{10}$$

where: $h_n = \alpha_n \dfrac{C_g^{nk'} - C_g^{nj'}}{\sqrt{\left(C_g^{nk'}\right)^2 + \left(C_g^{nj'}\right)^2}}$ and α_n is a parameter that can vary at each iteration n and should be chosen so that $h_n \leq 1$ and normally the simplest way to ensure this requirement is to choose $0 \leq \alpha_n \leq 1$.

Step 3: Convergence test.

If $\epsilon = \sum_{i=0}^{I} f_i^{n+1}\left(C_g^{ni} - C_g^{n1'}\right)^2 < \epsilon_0$ or $n > N$, let $\mathbf{f}^* = \mathbf{f^{(n+1)}}$ and then stop. Otherwise, $n = n+1$ and go to Step 1.

Clearly, the swapping rule Eq. (10) ensures that traffic is always switched from the more costly time interval to the less costly one and that no traffic switches between a pair of intervals if the generalised travel costs corresponding to them are equal. Certainly, there are other forms of h_n, as suggested in Subsection 2.3 in Carey and Ge (2012). Also, $\mathbf{f^{(n+1)}}$ from Eq. (10) is feasible if $\mathbf{f^n}$ is feasible, since reallocating departure rates as defined in Eq. (10) ensures that $\sum_{i=0}^{I} f_i^{n+1} = \sum_{i=0}^{I} f_i^n = D$ and f_i^{n+1} is nonnegative for all $i = 1,2,\ldots,I$.

Remark 3: This method equilibrates departure rates by swapping flows between each pair of time intervals. A similar method is proposed in Carey and Ge (2012) to search for a route-choice DUE solution by swapping flows between each pair of paths; as implied in this reference, the above method may have other forms, for example we can remove departure rates from time intervals in proportion to their deviation $\left(C_g^{nk} - C_g^{n1'}\right)$ from the time interval with the minimum generalised travel cost and reassign all of these removed departure rates to the time interval with the minimum generalised travel cost. Mounce and Carey (2011) investigates the convergence of the path-oriented swapping method and the speed of convergence.

1.3. An Illustrative Example

This example is intended to illustrate that the above pairwise swapping method can find the equilibrium solution to the bottleneck problem. The scenario for this example is also to be used in the rest of the paper.

The bottleneck of interest was chosen to be 5.25 kilometres or km long, i.e. L = 5.25 km with the free-flow travel time being tt^{ff} = 7.50 minutes (min), hence the free-flow speed on the bottleneck road segment is $v^f = \frac{L}{tt^{ff}} = 0.7$ km/min. The other parameter values for the bottleneck were defined as: k^c = 56 vehicles/km and k^j = 160 vehicles/km so that consequently $q^{\max}$ = 19.60 vehicles/min. Additionally, N =15000 iterations and $\alpha_0 = 10^{-6}$.

The time horizon under study was defined to be [0,90] (in minutes), and the preferred arrival window was [75,85] (in minutes). The time step size utilised was 0.25 minutes, that is, the time horizon divided into 360 intervals. A smaller time step size of 0.125 minutes was also tried, resulting in nearly the same solution profiles as presented in this paper. The total demand was assumed to be 625 vehicles. Without loss of generality, the value-of-time parameter value was set to 1, i.e. α = 1 monetary unit per minute. In addition, $\beta = \frac{\beta}{\alpha} = 0.22$ monetary units per minute and $\gamma = \frac{\gamma}{\alpha} = 2.00$ monetary units per minute.

Fig. 3 displays the solution profiles from this example. It can be seen that there was a spike in the cost profile in the vicinity of t = 80 but that the departure rates during the period corresponding to this spike are 0; hence the equilibrium condition (8) was not violated. The solution profiles of departure rates and travel costs show that all travellers departed during $[t_1, t_5]$, in which their generalised travel costs were identical and equal to the minimal travel cost. Therefore, the equilibrium condition (8) has been satisfied across the time horizon, which demonstrates that the above pairwise swapping method successfully found the equilibrium solution to this bottleneck problem.

As seen in Fig. 3a, there were two sharp drops in equilibrium departure rates respectively at t_3 and t_4. These were due to the acceleration fans, which arise when the downstream density is low and the immediate upstream traffic starts to accelerate. Accordingly, the rates of change in travel times had a rapid drop at t_3 and t_4; as shown in Fig. 3b, the change rates become 0 from t_3 and go negative from t_4. The gap between the profiles of travel times and generalised travel costs represents the penalties due to earlier or later arrival.

Fig. 3a also shows that the departure rates were greater than the capacity of the bottleneck road segment between $[t_2, t_3]$, hence a queue started to form at t_2 and remained after t_3, which demonstrates the necessity of implementing congestion charging. Fig. 4 gives the two cumulative flow curves $N(t)$ respectively at the entry and exit of the bottleneck road segment. It can be seen that the rate of outflow does not always equal the bottleneck capacity and that, in such bottleneck settings, neither departure nor arrival N-curves increase linearly, which differs from those results based on the point-queue model (Arnott *et al.* 1990, 1993; Lindsey 2006, 2010; Lindsey *et al.* 2012; Small, Verhoef 2007; Van den Berg 2012; Van den Berg, Verhoef 2011).

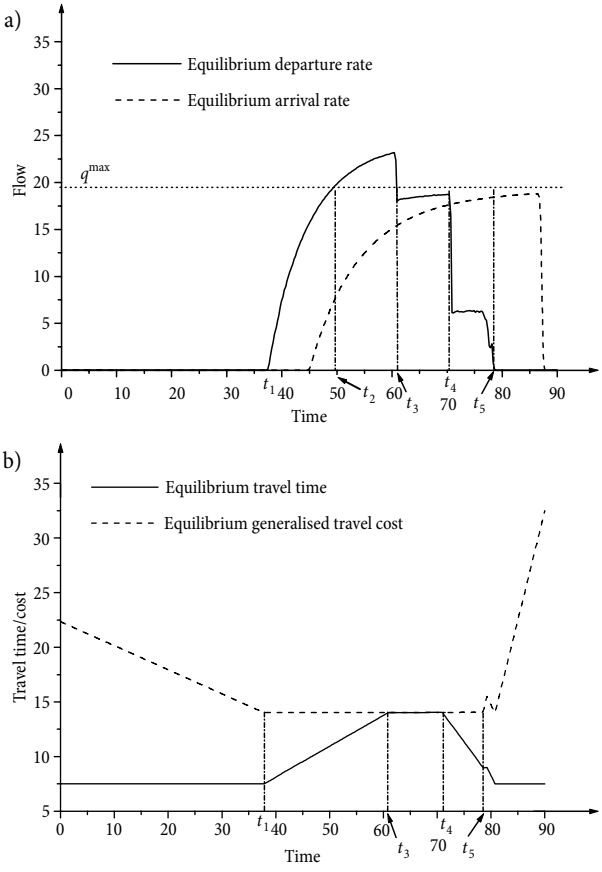

Fig. 3. Profiles of solutions to the example in Subsection 1.3: a – departure and arrival rates; b – travel times and costs

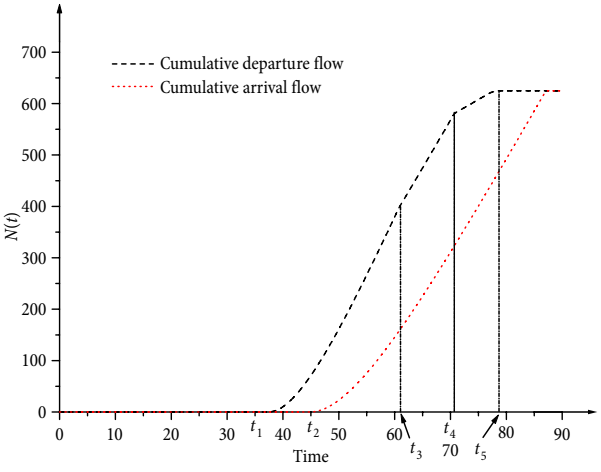

Fig. 4. Cumulative N-curves for the example in Subsection 1.3

2. Demand Peaks under Congestion Charging

This section presents a series of observations on boundary issues under the three types of toll profiles below:

$$C_{toll}^1(t) = \begin{cases} u, & \text{if } t \in [t^s, t^e]; \\ 0, & \text{if otherwise;} \end{cases}$$

$$C^2_{toll}(t)=\begin{cases}\dfrac{u\left(t-t^s\right)}{t^m-t^s}, & \text{if } t\in\left[t^s,t^m\right];\\ \dfrac{u\left(t^e-t\right)}{t^e-t^m}, & \text{if } t\in\left[t^m,t^e\right];\\ 0, & \text{if otherwise;}\end{cases}$$

$$C^3_{toll}(t)=\begin{cases}\dfrac{u\left(t-t^s\right)}{t^l-t^s}, & \text{if } t\in\left[t^s,t^l\right];\\ u, & \text{if } t\in\left[t^l,t^r\right];\\ \dfrac{u\left(t^e-t\right)}{t^e-t^r}, & \text{if } t\in\left[t^r,t^e\right];\\ 0, & \text{if otherwise,}\end{cases} \tag{11}$$

where: the parameter u represents the toll level and is a constant (in monetary units). In our later experiments, u is specified at three levels: low, medium and high, represented respectively by u = 2, 4 or 6. As previously mentioned, the congestion charging schemes in London and in Milan are typical examples of the first type of toll profiles while the second and third types of profiles may be considered abstract versions of those toll profiles implemented in Singapore and in Stockholm.

The key reason for choosing the abstract rather than real-life step toll profiles is that such continuous charging profiles (with no sharp changes in congestion charging tolls) make it possible to find equilibrium flow patterns or solutions (Ge *et al.* 2014). By the same token, our numerical experiments have actually used the following continuous toll profile to approximate $C^1_{toll}(t)$:

$$C^1_{toll}(t)=\begin{cases}\dfrac{1+\sin\left(\dfrac{t-t^s}{\delta}\pi\right)}{2}u, & \text{if } t\in\left[t^s-\dfrac{\delta}{2},t^s+\dfrac{\delta}{2}\right];\\ u, & \text{if } t\in\left[t^s-\dfrac{\delta}{2},t^s+\dfrac{\delta}{2}\right];\\ \dfrac{1+\sin\left(\dfrac{t^e-t}{\delta}\pi\right)}{2}u, & \text{if } t\in\left[t^s+\dfrac{\delta}{2},t^e-\dfrac{\delta}{2}\right];\\ 0, & \text{if otherwise,}\end{cases} \tag{12}$$

where: the constant $\dfrac{\delta}{2}$ is very small in comparison to the length of the charging period. This approximation makes the toll profile at the start and end of the charging period $\left[t^s,t^e\right]$ not change suddenly from 0 to u or from u to 0, hence smoothing out the discontinuities of toll profiles at both ends of the charging period. Otherwise, there would be large fluctuations in resulting departure flow rates at both ends of $\left[t^s,t^e\right]$ or no equilibrium solutions would exist.

2.1. Impacts of Choice of Charging Periods on Boundary Effects

The solution profiles of departure rates in Fig. 3 show that the departure rates in $\left[t_2,t_3\right]\subset\left[50,65\right]$ are higher than the bottleneck capacity. To reduce the congestion in this period, a charging period wider than this one would be desirable. The first trial was $\left[t^s,t^e\right]=\left[45,65\right]$ with u = 2 for all. Other parameter values used are: $\delta=2$ for profile 1, $t^m=\dfrac{t^s+t^e}{2}=55$ for profile 2 and $\left[t^l,t^r\right]=\left[52.5,57.5\right]$ for profile 3. The solution profiles of departure rates and travel costs/times are given in Fig. 5.

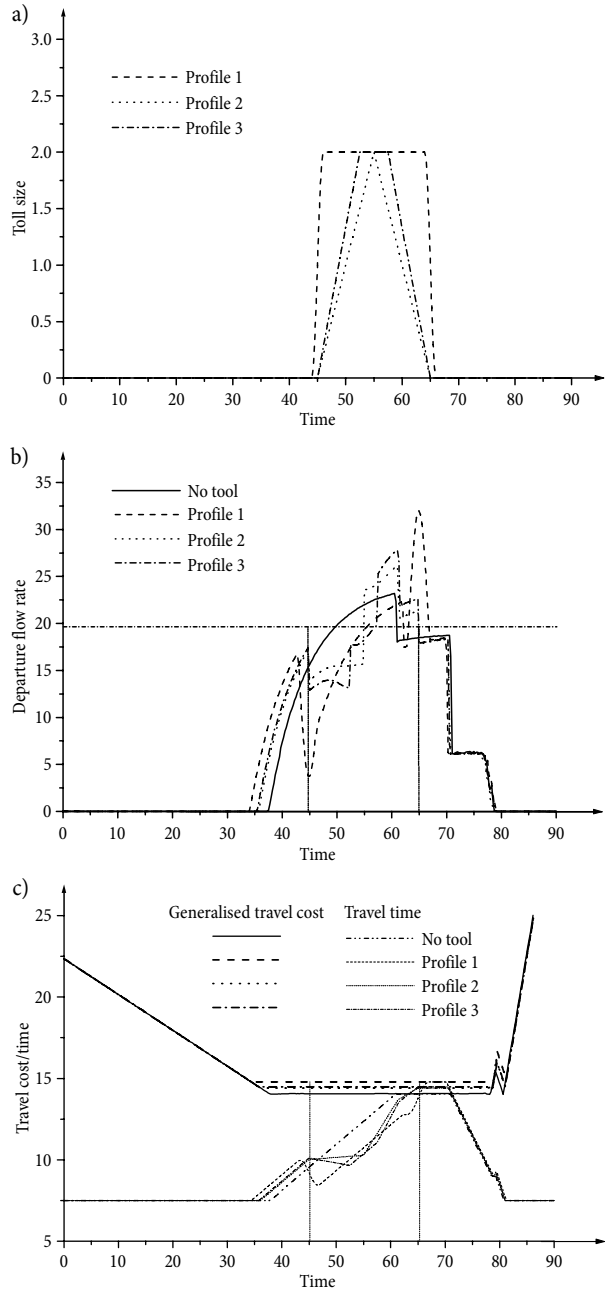

Fig. 5. Solution profiles for charging period choice: a – toll profiles; b – departure travel times; c – travel times and costs

As seen in Fig. 5b, imposing the charge on traffic departing in the charging period [45, 65] made many travellers depart before the start of the charge (i.e. $t = 45$), so there is a boundary demand peak corresponding to each toll profile but (importantly) none of them is higher than the bottleneck capacity; it should be noted that the low toll level $u = 2$ was used. For each toll profile, there is a significant drop in the resulting departure rates at $t = 45$; among all three solution profiles of departure rates, the drop corresponding to toll profile 1 is the largest, which results from the biggest rate of change in the toll at or around $t = 45$.

It is interesting to see no significant change in departure rates after the charge end if full consideration is given to the fact that the toll profile 1 did not fall down to 0 until $t^e + \frac{\delta}{2} = 66$. However, just before the end of the charge every one of the three solution profiles of departure rates has a peak greater than the bottleneck capacity and all of these peaks are higher than the original no-toll peak, which is not what was anticipated before implementing this charging scheme. The primary reason could be that the charging period was too small; specifically the stopping end of the period should have been set to a much later moment. In the early stage of Singapore congestion charging, i.e. the Area Licensing Scheme (ALS), the charging period for the morning was initially set as from 7:30 am to 9:30 am daily. In order to control the peak demand resulting from those waiting to enter just after 9:30 am, the charging period was, shortly after opening, extended from 7:30 am to 10:15 am (Seik 1997).

Based on the above observations, we extended the charging termination time from 65 to 75. We also delayed the start of congestion charging by 5 time units. The changes in the start and stopping time of charging have defined the second trial of the charging period, i.e. [50,75]. However, this 'late-start' charging led to the resulting demand peaks being greater than the bottleneck capacity at the start end of the charging period, which called for the extension of the charging period to an earlier start. So throughout the rest of the experiments, we selected [45,75] as the charging period.

We did not attempt to carry out a sensitivity analysis of the charging period but tended to show that a wrong choice of charging period could make the situation worse. A final point to note in this subsection is that the solution profiles of departure rates and travel costs in Fig. 5b, c show that the pairwise swapping method can indeed find the equilibrium solutions even when such toll profiles as Profile 1 in Fig. 5a, and accordingly travel cost profiles, change very quickly at both ends of the charging period. In the rest of the paper we will no longer present the solution profiles of travel times and travel costs and it is assumed that convergence does indeed occur.

2.2. First Type of Toll Profiles

First type of toll profiles: $\left[t^s, t^e\right] = \left[45,75\right]$, $\delta = 4$.

As shown in Table 1, the higher the toll level the more travellers choose to depart before the charge starts or after it stops. This is generally consistent with the original objective of traffic congestion charging to spread the total travel demand, both spatially and temporally, so that the road capacity can be utilised more efficiently.

Table 1. Percentages of demand departing before and after the charging period

Toll level	Pre-charging	Post-charging	In-between
No-toll	11.35%	2.72%	85.92%
$u = 2$	17.29%	3.66%	79.05%
$u = 4$	27.74%	4.59%	67.67%
$u = 6$	41.17%	5.57%	53.26%

To observe the boundary effects of congestion charging, we examine the demand peaks of the solution profiles in Fig. 6b. At $u = 2$, the demand peaks at both ends of the charging period are lower than the bottleneck capacity and the demand peak inside the charging period was also reduced. When u was increased to 4, the demand peaks at both ends of the charging period also increased and the demand peak inside the charging period was further reduced although the resulting peak was still higher than the bottleneck capacity. At this toll level, the demand peak at the start of the charge exceeds the bottleneck capacity for a very short period. When u continues to increase up to 6, the resulting demand peak inside the charging period was lower than the capacity but the demand peaks at both ends of the charging period were higher than the capacity. Hence, undesired boundary effects occurred.

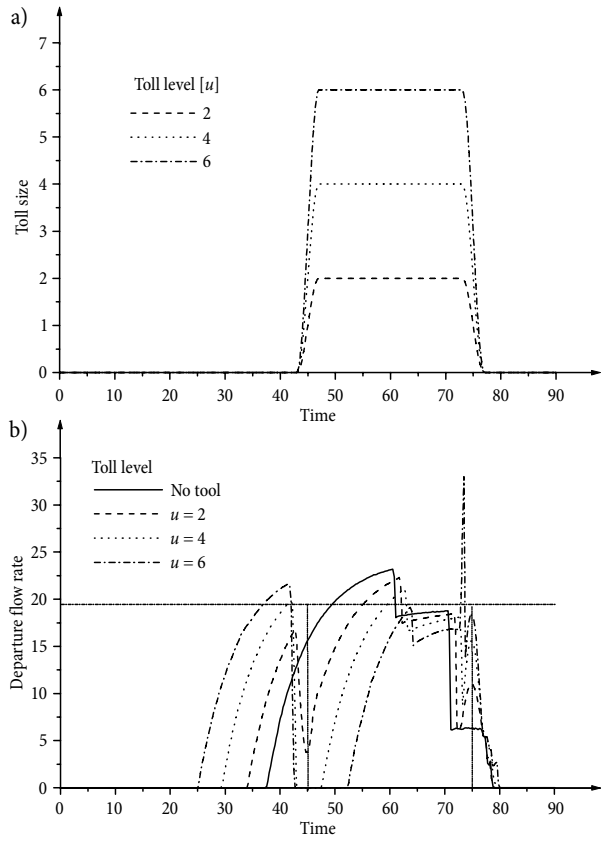

Fig. 6. Solution profiles from the first type of toll profiles: a – toll profiles; b – departure rates

This illustrates that an excessively high toll level can produce boundary demand peaks higher than the bottleneck capacity whereas a low toll level may not reduce congestion inside the charging period sufficiently. This implies that the choice of toll levels is a critical issue in congestion charging design when we use this type of toll profiles.

Note that in Fig. 6b there is quite a long interval in which no traffic appeared at either $u = 2$ or 4, which made both time and road capacity unexpectedly underutilised. This is not only due to the excessively high toll level but also because of the sharp change in the toll. These observations suggest that gradually changing toll profiles be a better choice than flat toll profiles because we may be able to reduce the boundary effects by lowering the tolls at both ends of the charging periods, as implemented in Singapore and in Stockholm.

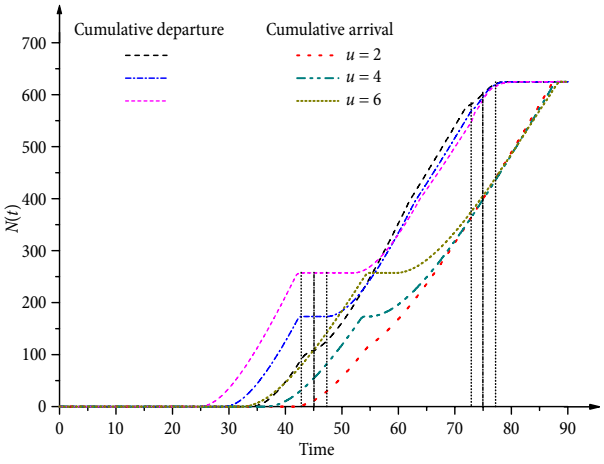

Fig. 7. Cumulative N -curves for the first type of toll profiles

Fig. 7 displays the cumulative N-curves under the first type of toll profiles and we have to acknowledge that it is not that intuitive to observe the boundary effects by means of such cumulative N-curves unless a linear cumulative curve whose slope equals the bottleneck capacity is also drawn; that the slope of an N-curve is greater than the bottleneck capacity at the time when the toll jumps up or down means the appearance of undesired boundary effects. In addition, it is difficult to tell from this figure that congestion charging has driven more traffic to depart after the charging period. However, these curves are widely used in equilibrium analysis of the bottleneck scenario and consequently we included them here.

2.3. Second Type of Toll Profiles

Second type of toll profiles: $\left[t^s, t^e\right] = \left[45, 75\right]$, $t^m = 60$.

As shown in Table 2, under the type of toll profiles given in Fig. 8a, the higher the toll level the more travellers choose to depart before the charge starts or after it stops. However, when we focus on the proportions of travel demand departing in the post-charging period we can see that the total departure at $u = 2$ is even less than in the no-toll case. This may be due to congestion

Table 2. Percentages of demand departing before and after the charging period

Toll level	Pre-charging	Post-charging	In-between
No-toll	11.35%	2.72%	85.92%
$u = 2$	13.98%	2.50%	83.52%
$u = 4$	20.33%	2.77%	76.90%
$u = 6$	29.93%	2.98%	67.09%

charging making traffic less crowded inside the charging period and because the toll close to the end of the charging period was low, travellers tended to depart before the charging ended rather than waiting outside the charging point or cordon. This can be confirmed by the last peaks of the departure rate profiles in Fig. 8b.

It is interesting to see that, as the toll increased from one level to the next, more travellers departed before the charge started but there were no significant changes in the departure rates after the charge stopped.

As shown in Fig. 8, when the charge started to grow linearly from 0 at $t = 45$, the departure rates dropped quickly to a level below the bottleneck capacity; but as the toll started to decrease linearly at $t = 60$, the departure rates increased quickly up to a level above the capacity. As the toll continues to fall, the departure rates started to fluctuate and eventually dropped to a level below the capacity before the end of the charging period.

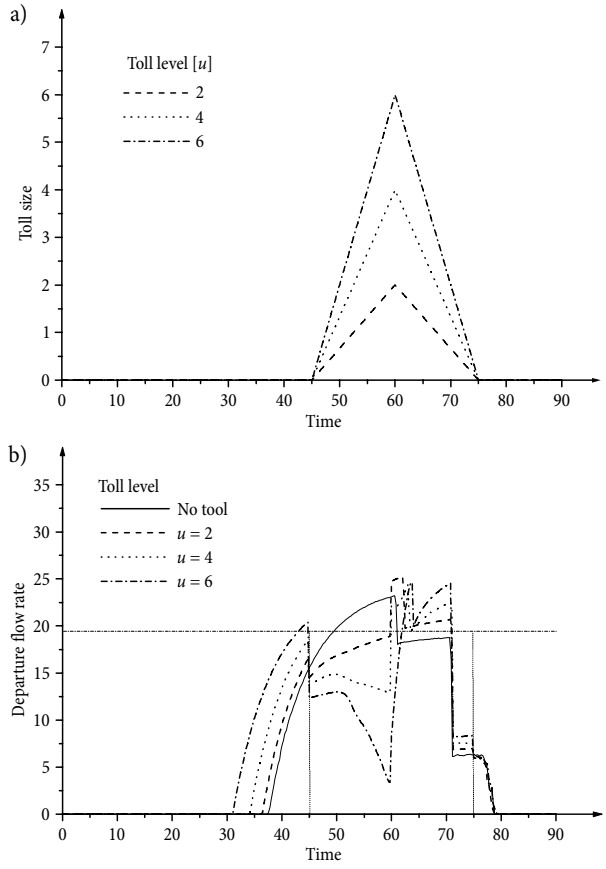

Fig. 8. Solution profiles from the second type of toll profiles: a – toll profiles; b – departure rates

Some of these demand peaks under congestion charging were even higher than the original demand peak associated with the no-toll case. The primary reason may be due to the rapid fall in the toll. This shows that if a toll profile decreases too fast, the boundary issues may shift from the end of the charge period to a point inside the period. The resulting demand peak may be higher than the original peak with no toll. There may be more than one demand peak because of the fluctuations in departure rates when the toll profile falls.

It is interesting to see in Fig. 1 that the toll implemented in Singapore, which was associated with the morning peak fell more slowly than it rose whereas the toll associated with the evening peak fell more quickly than it rose. Again, in Fig. 1, the toll implemented in Stockholm and associated with the morning peak increased to a maximum level and then decreased but did not fall to zero; the toll remained at this non-zero level; when the evening peak arrived, the toll increased again up to a maximum level and then fell to zero. Based on the numerical results above, it may be concluded that such toll profiles would help to reduce the demand peaks inside and outside the charging period.

These experiments also show that an excessively high toll level can produce boundary demand peaks higher than the bottleneck capacity (as the toll level of $u = 6$ produced a boundary demand peak greater than the capacity at the start of the charge).

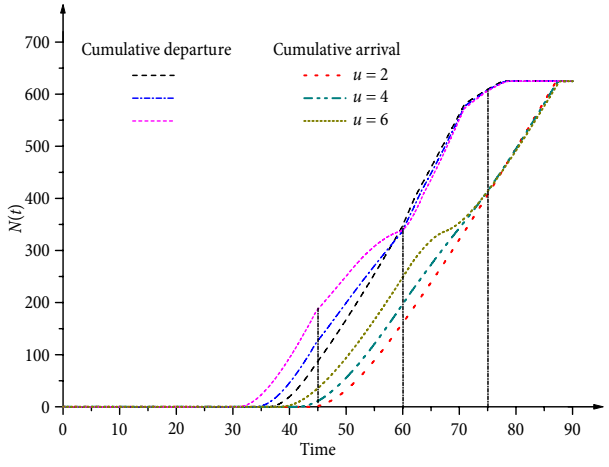

Fig. 9. Cumulative *N*-curves for the second type of toll profiles

Analysing the *N*-curves in Fig. 9, we can clearly see that the arrival rate may be time-varying, given the bottleneck scenario set up previously, which is different from the Vickrey (1969) or point-queue bottleneck model, whose arrival rate is constant and always equal to the bottleneck capacity.

2.4. Third Type of Toll Profiles

Third type of toll profiles: $[t^s, t^e] = [45,75]$, $t^l = 55$ and $t^r = 65$.

The solution profiles under this type of toll profiles are given in Fig. 10, based on which a general point can be made that the higher the toll level the more travellers choose to depart outside the charging period. This is substantiated by the data in Table 3. As discussed previously, this point is also true for the other two types of toll profiles.

The second point consistent with the previous experiments is that there were demand peaks just before the start of congestion charging but they dropped sharply once the charge started. Different from the experiments on the flat toll (i.e. first type of toll profiles) but similar to the experiments on the second type of toll profiles, the departure rates did not drop down to zero but remained far above zero.

The third point shared by the experiments on all three types of toll profiles is that there were no significant changes in departure rates after the termination of the charge.

Furthermore, the largest of the three toll levels $u = 6$ produced a problematic solution profile of departure rates. Firstly, the demand peaks were greater than

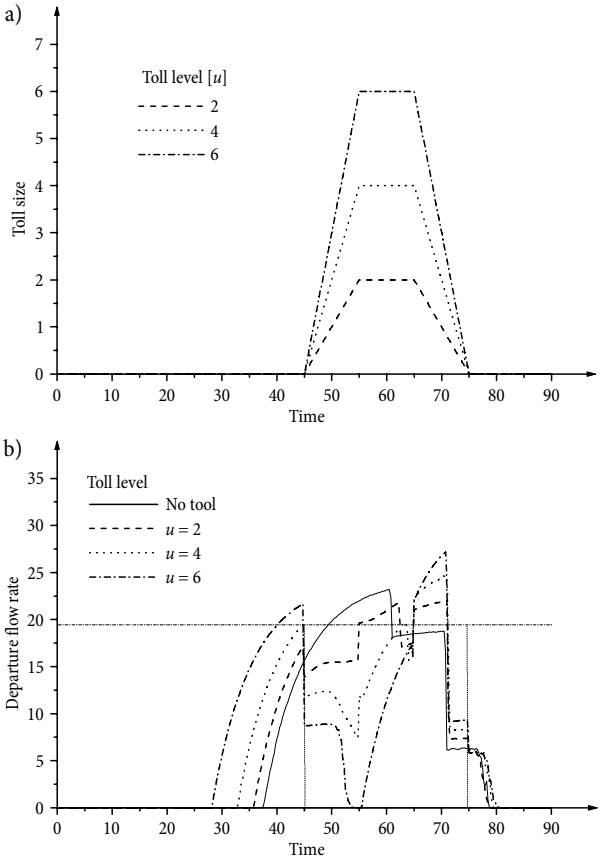

Fig. 10. Solution profiles from the second type of toll profiles: a – toll profiles; b – departure rates

Table 3. Percentages of demand departing before and after the charging period

Toll level	Pre-charging	Post-charging	In-between
No-toll	11.35%	2.72%	85.92%
$u = 2$	15.40%	2.41%	82.19%
$u = 4$	23.53%	3.01%	73.47%
$u = 6$	37.75%	2.93%	59.32%

the bottleneck capacity, both inside and outside the charging period. Secondly, the departure rates stayed at zero for a while when the toll reached the maximum, which results in the under-utilisation of time and road capacity resources. Thirdly, the demand peak before the charging end was much higher than the demand peak associated with the no-toll case. The primary reason for this problematic solution profile of departure rates is an excessively high toll level. This again shows that the choice of toll levels is of critical importance to the success of a congestion charging project.

At u = 2 or 4, the demand peaks at either end of the charging period have been significantly reduced and are lower than the bottleneck capacity. In [63.5,70.5], the departure rates associated with all three toll levels are above the bottleneck capacity and the lower toll level the lower the peak of departure rates. While the toll was falling, the departure rates jumped to a level above the bottleneck capacity. Again, the slower rate of decrease in the toll the lower the demand peaks. In Subsection 2.3, we discussed briefly the toll profiles given in Fig. 1, which helps to prevent too many people from delaying their trips and causing undesired demand peaks.

We can also argue about this using the solution profiles associated with the second and third types of toll profiles. For each toll level u, the RHS of toll profiles of the third type decreases much faster than that of the second type and the resulting most-right demand peak from the third type of toll profile was much higher than that from the second type.

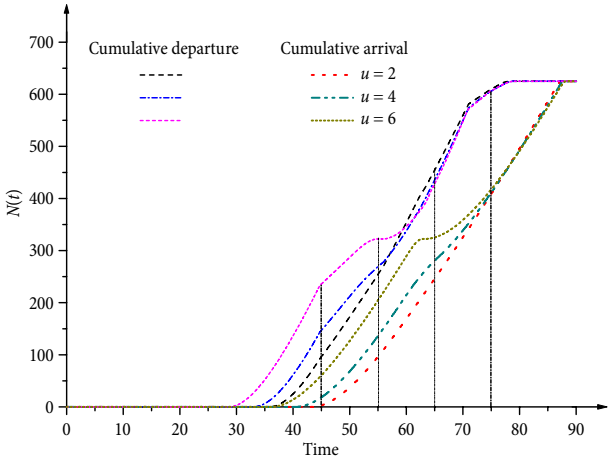

Fig. 11. Cumulative N-curves for the third type of toll profiles

Fig. 11 gives the cumulative N-curves under this type of toll profiles at the three toll levels, which may be seen to be different from those based on the point-queue bottleneck model (Vickrey 1969; Zhang *et al.* 2013) but is similar to those in Newell (1988).

2.5. Toll Profile Assessment

Whilst further examination of the charging sub-period choice is a requirement for our ongoing work, here we include a description of a preliminary assessment of the relative benefits of the charging schemes investigated in the previous subsections. Fig. 12 shows the total travel

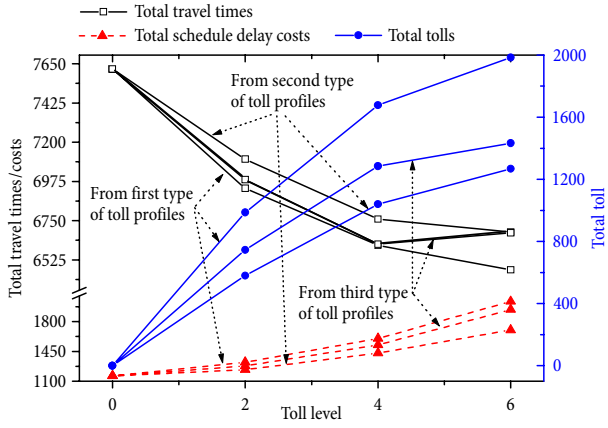

Fig. 12. Total travel times, total schedule delay costs and total tolls

times, total schedule delay costs and total tolls collected under each of the toll levels (low/mid/high) for each toll profile considered previously. As shown in this figure, all three types of toll profiles show reductions in total travel times as the tolls increased from 0 to 2 (low level) and to 4 (mid level). Whilst corresponding increases in the total schedule delay may be observed, the ratio of the rate of change in total travel time to total schedule delay is greater than 1, which indicates that increasing the toll to such levels is beneficial. The total travel time increased when the toll level went up from 4 to 6 for the third type of toll profiles (trapezoidal), so it would not be beneficial to increase tolls to this level for toll profiles of this type. For the first and second toll profile types (fixed and triangular), the travel times continued to fall when the toll was raised to the specified highest level, but the relative benefit of this change was reduced, the ratio of the rate of change of total travel time to total schedule delay only being of the order of 0.3 (significantly less than unity).

3. Extra Requirements on Congestion Pricing Design

Having observed and analysed the undesired boundary effects of congestion charging, it is pertinent to consider the objectives of congestion pricing. The general aim of traffic congestion charging or travel demand management is to spread all travel demand spatially and temporally so that the social cost is minimised or social welfare is maximised. However, in designing a congestion charging scheme it is extremely difficult, if not impossible, to get sufficient precise data to compute real social costs or welfare, which is due to market distortions, hidden information and hidden actions on users, etc. It is due to these factors that only second-best congestion pricing tolls can be determined in real-life transportation systems. The following set of criteria is proposed to make resulting toll profiles meet our needs and generate fewer undesired side-effects:

(a) the social cost should be minimised or social welfare maximised;

(b) the resulting flows on the edge or boundary of the congestion charging zone(s) should not be greater than the available capacities for longer than a certain (relatively short) period;

(c) no user should tend to accelerate due to the avoidance of congestion charging tolls at the sacrifice of traffic safety while nearing the charging zone(s);
(d) no user should tend to slow down or idle outside the charging zone(s) to wait for the end of the charging period;
(e) the charging amount should be transparent and also change from time to time at a reasonable pace (e.g. monthly or quarterly) in response to emerging undesired boundary effects as well as changes in travel demand levels, travel patterns, land use patterns, travel behaviour, etc.

An online note (Vickrey 1992) lists Nobelist William S. Vickrey's 12 principles for efficient congestion pricing and Hau (2006) summarises the existing criteria for a 'good' road pricing system from the perspective of operational requirements. It is noteworthy that the requirements we have proposed in this paper are not contradictory but supplementary to these principles or criteria. The first of Vickrey's principles suggests that 'Charges should equal or exceed the marginal social cost of each trip'. In a distorted market, it may not be feasible to determine the correct marginal social cost of each trip. As a compromise, we suggest minimising social costs or maximising social welfare, which should be calculated on the basis of current market prices. Although the same problem of market distortion exists, this optimisation approach may be able to offer a set of tolls consistent with this objective (minimising social cost or maximising social welfare). The second of Vickery's principles is 'Charges should vary in sufficiently small increments to avoid creating mini-peaks'. He might have noticed from the practice of congestion charging in Singapore that large sharp changes in tolls can cause such 'mini-peaks'. In September 1998, Singapore took the bold step to upgrade their flat-toll-based ALS congestion pricing system to an Electronic Road Pricing (ERP) system based on multi-step tolls, as shown in Fig. 1, which, it is reported, has reduced the boundary effects or 'mini-peaks'. Our criteria (b)–(e) are proposed to specify how to reduce the boundary effects. In future research on overcoming the undesired boundary effects of congestion charging, it is essential to explicitly take into account these constraints in an optimisation model/formulation. The rest of the Vickrey's principles work together with the extra requirements proposed above.

Hau (2006) categorises the existing criteria into three groups, which are listed respectively from the view points of users, transport authority, and the society. The first set of four criteria is given for the sake of users, including user-friendliness (simplicity), transparency (via ex ante pricing), anonymity (protection from invasion of privacy), and prepayment/postpayment options for charging. The subsequent seven criteria have been given from the standpoint of a transport authority, including enhanced efficiency via direct charging, flexibility (responsiveness to demand), reliability, security and enforcement, provision for occasional visitors, 'market' price as an investment signal, and passage of revenue-cost tests. The last set of nine criteria is presented from the societal point of view, i.e. passage of benefit-cost tests, minimum of road work and environmental intrusion, provision for mixed traffic, handling of transitional phase, compatibility with other systems, modularity to add-on options, tolerance to culture of non-compliance, tolerance to varied geography, fairness and the availability of alternatives. It is noteworthy that none of these 20 criteria are put forward to directly solve the problems of undesired boundary effects of congestion pricing. Our criterion e) may be associated with two of the 20 criteria, i.e. transparency and flexibility (responsiveness to demand). It is well known that congestion pricing can influence traveller choices of departure times as well as trip routes, the reason being their travel costs would increase if travellers drive into a charging zone in the charging period. Only when they know how much to pay can they make their best decision on such choices. Therefore, we need 'transparency'. It is reasonable to say that the existing flow forecast techniques may not be able to meet transparency if a real-time time-varying congestion charging scheme is implemented. The second part of criterion e) calls to adapt a congestion charging scheme to changed travel demand patterns, which requires that congestion pricing system be flexible to travel demand that may evolve due to many factors, including implementation or operation of congestion pricing.

Concluding Remarks

In summary, the purpose of this paper is at least twofold. One is to highlight and investigate the boundary effects of congestion charging and the other is to examine the implications of such effects for the design and implementation of congestion charging projects in the future. To fulfill the purpose, we did not use the widely-used point-bottleneck/queue model but employed the bottleneck model proposed in Newell (1988) that consists of two parts: one is a road segment with a limited capacity and the other is a deterministic point-queue model at the entry of the segment, in which the queue exists only when the departure rate is greater than the receiving capacity of the downstream bottleneck link.

A series of numerical experiments on the impacts of choice of charging periods on the boundary effects of congestion charging have been carried out to show that a wrong choice of traffic congestion charging periods may make the situation worse rather than better. The investigation of effects of toll levels implies that an excessively high toll level can produce a boundary demand peak unacceptably higher than the bottleneck capacity and could also reduce the departure rates inside the charging period to a level so unexpectedly low that the time and road capacity resources could not be utilised efficiently. We were also aware of the rate of change, in particular the decrease rate, in the toll over time. Our preliminary observations demonstrate that a slower rate of decrease in the toll produced a lower demand peak. Finally, none of the toll profiles tested in this paper produced a departure rate profile entirely below the bottleneck capacity

across the time horizon. In addition, a possible reason for part of the peaks in the departure rate profiles under investigation is the assumption of perfect knowledge on traffic conditions and travellers' values of travel and schedule delay times; in fact, a stochastic bottleneck model allowing for user uncertainty would reduce the severity of the departure time or rate peaks observed in the previous experiments.

Our observations also imply that, under all three aforementioned types of toll profiles, the use of congestion charging as the only transport management instrument may be unable to entirely remove congestion (i.e. demand greater than the capacity), both inside and outside the charging period. Therefore, it may be more efficient to implement a congestion charging project together with other transport policy instruments, such as traveller information provision, improved public transport services and other demand management tools. These are part of our ongoing research.

It has to be acknowledged that the nature of a constant toll being introduced at a precise time means that under Assumptions 1–3 the flow effect on the network will necessarily be to produce a surge of inflow just prior to the charging period commencing and a dip in inflow just after (the effects being reversed at the termination of the charging period). In practice, Assumption 2 (perfect information) will not hold and Assumption 3 (cost minimisation) will only hold partially. It is then of importance to investigate how the relaxation of these assumptions will affect the undesired boundary effects investigated in this paper. Another hidden assumption that all travellers share the same set of values of time (i.e. α, β and γ) may also affect boundary effects (but these were not touched on in this paper since it is beyond the intended scope of the investigation and the reader who is interested in this issue may refer to Ramadurai *et al.* (2010), Tian *et al.* (2013), Van den Berg (2014)). Whilst it may be hypothesised that the relaxation of Assumptions 2 and 3 or consideration of heterogeneity (in the body of travellers) would produce less sharp path inflow fluctuations, it is unlikely that such relaxations will mitigate these boundary effects entirely; in fact, in real-life traffic where these assumptions are not satisfied, such undesired boundary effects of congestion charging have been recorded and discussed (e.g. TfL 2008; Salmon 2010). Therefore, an efficient removal of the undesired boundary effects will still require a well-designed time-varying toll profile, although the adverse effects reported in this paper are likely to be somewhat reduced under more realistic assumptions, whish is also supported by our latest investigation of boundary effects of congestion charging on road networks (Ge *et al.* 2015).

Acknowledgements

We gratefully acknowledge support from the National Natural Science Foundation of China (Grant No: 71171026).

Great thanks are given to Prof. Malachy Carey of the University of Leeds (United Kingdom) who provided useful comments on the earlier version of this paper.

References

Arnott, R.; De Palma, A.; Lindsey, R. 1993. A structural model of peak-period congestion: a traffic bottleneck with elastic demand, *The American Economic Review* 83(1): 161–179.

Arnott, R.; De Palma, A.; Lindsey, R. 1990. Economics of a bottleneck, *Journal of Urban Economics* 27(1): 111–130. http://dx.doi.org/10.1016/0094-1190(90)90028-L

Braid, R. M. 1989. Uniform versus peak-load pricing of a bottleneck with elastic demand, *Journal of Urban Economics* 26(3): 320–327. http://dx.doi.org/10.1016/0094-1190(89)90005-3

Carey, M.; Ge, Y. E. 2012. Comparison of methods for path flow reassignment for dynamic user equilibrium, *Networks and Spatial Economics* 12(3): 337–376. http://dx.doi.org/10.1007/s11067-011-9159-6

Carey, M.; Ge, Y. E. 2005. Convergence of a discretised travel-time model, *Transportation Science* 39(1): 25–38. http://dx.doi.org/10.1287/trsc.1030.0083

Carey, M.; Ge, Y. E. 2004. Efficient discretisation for link travel time models, *Networks and Spatial Economics* 4(3): 269–290. http://dx.doi.org/10.1023/B:NETS.0000039783.57975.f0

Carey, M.; Ge, Y. E. 2003. Comparing whole-link travel time models, *Transportation Research Part B: Methodological* 37(10): 905–926. http://dx.doi.org/10.1016/S0191-2615(02)00091-7

Chu, X. 1995. Endogenous trip scheduling: the Henderson approach reformulated and compared with the Vickrey approach, *Journal of Urban Economics* 37(3): 324–343. http://dx.doi.org/10.1006/juec.1995.1017

Daganzo, C. F. 1995a. A finite difference approximation of the kinematic wave model of traffic flow, *Transportation Research Part B: Methodological* 29(4): 261–276. http://dx.doi.org/10.1016/0191-2615(95)00004-W

Daganzo, C. F. 1995b. Properties of link travel time functions under dynamic loads, *Transportation Research Part B: Methodological* 29(2): 95–98. http://dx.doi.org/10.1016/0191-2615(94)00026-V

Daganzo, C. F. 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory, *Transportation Research Part B: Methodological* 28(4): 269–287. http://dx.doi.org/10.1016/0191-2615(94)90002-7

Doan, K.; Ukkusuri, S.; Han, L. 2011. On the existence of pricing strategies in the discrete time heterogeneous single bottleneck model, *Transportation Research Part B: Methodological* 45(9): 1483–1500. http://dx.doi.org/10.1016/j.trb.2011.05.019

Friesz, T. L.; Bernstein, D.; Smith, T. E.; Tobin, R. L.; Wie, B. W. 1993. A variational inequality formulation of the dynamic network user equilibrium problem, *Operations Research* 41(1): 179–191. http://dx.doi.org/10.1287/opre.41.1.179

Friesz, T. L.; Kim, T.; Kwon, C.; Rigdon, M. A. 2011. Approximate network loading and dual-time-scale dynamic user equilibrium, *Transportation Research Part B: Methodological* 45(1): 176–207. http://dx.doi.org/10.1016/j.trb.2010.05.003

Friesz, T. L.; Mookherjee, R. 2006. Solving the dynamic network user equilibrium problem with state-dependent time shifts, *Transportation Research Part B: Methodological* 40(3): 207–229. http://dx.doi.org/10.1016/j.trb.2005.03.002

Friesz, T. L.; Mookherjee, R.; Yao, T. 2008. Securitizing congestion: the congestion call option, *Transportation Research Part B: Methodological* 42(5): 407–437. http://dx.doi.org/10.1016/j.trb.2007.10.002

Ge, Y. E.; Carey, M. 2004. Travel time computation of link and path flows and first-in-first-out, in B. Mao, Z. Tian, Q. Sun, (Eds.). *Proceedings of the 4th International Conference on Traffic and Transportation Studies*, 2–4 August 2004, Dalian, China, 326–335.

Ge, Y. E.; Stewart, K.; Sun, B.; Ban, X. G.; Zhang, S. 2015. Investigating undesired spatial and temporal boundary effects of congestion charging, *Transportmetrica B: Transport Dynamics*. http://dx.doi.org/10.1080/21680566.2014.961044

Ge, Y. E.; Sun, B. R.; Zhang, H. M.; Szeto, W. Y.; Zhou, X. 2014. A comparison of dynamic user optimal states with zero, fixed and variable tolerances, *Networks and Spatial Economics*. http://dx.doi.org/10.1007/s11067-014-9243-9

Ge, Y. E.; Zhou, X. 2012. An alternative definition of dynamic user optimum on signalised road networks, *Journal of Advanced Transportation* 46(3): 236–253. http://dx.doi.org/10.1002/atr.207

Hau, T. D. 2006. Congestion charging mechanisms for roads, part I – conceptual framework, *Transportmetrica* 2(2): 87–116. http://dx.doi.org/10.1080/18128600608685658

Hendrickson, C.; Kocur, G. 1981. Schedule delay and departure time decisions in a deterministic model, *Transportation Science* 15(1): 62–77. http://dx.doi.org/10.1287/trsc.15.1.62

Laih, C.-H. 2004. Effects of the optimal step toll scheme on equilibrium commuter behaviour, *Applied Economics* 36(1): 59–81. http://dx.doi.org/10.1080/0003684042000177206

Laih, C.-H. 1994. Queueing at a bottleneck with single- and multi-step tolls, *Transportation Research Part A: Policy and Practice* 28(3): 197–208. http://dx.doi.org/10.1016/0965-8564(94)90017-5

Lighthill, M. J.; Whitham, G. B. 1955a. On kinematic waves. I. Flood movement in long rivers, *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences* 229: 281–316. http://dx.doi.org/10.1098/rspa.1955.0088

Lighthill, M. J.; Whitham, G. B. 1955b. On kinematic waves. II. A theory of traffic flow on long crowded roads, *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences* 229: 317–345. http://dx.doi.org/10.1098/rspa.1955.0089

Lin, J.; Ge, Y. E. 2006. Impacts of traffic heterogeneity on roadside air pollution concentration, *Transportation Research Part D: Transport and Environment* 11(2): 166–170. http://dx.doi.org/10.1016/j.trd.2005.12.001

Lindsey, R. 2010. Reforming road user charges: a research challenge for regional science, *Journal of Regional Science* 50(1): 471–492. http://dx.doi.org/10.1111/j.1467-9787.2009.00639.x

Lindsey, R. 2006. Do economists reach a conclusion on road pricing? The intellectual history of an idea, *Econ Journal Watch* 3(2): 292–379.

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. http://dx.doi.org/10.1016/j.jue.2012.02.001

Lo, H. K.; Szeto, W. Y. 2002. A cell-based variational inequality formulation of the dynamic user optimal assignment problem, *Transportation Research Part B: Methodological* 36(5): 421–443. http://dx.doi.org/10.1016/S0191-2615(01)00011-X

Long, J.; Gao, Z.; Szeto, W. Y. 2011. Discretised link travel time models based on cumulative flows: formulations and properties, *Transportation Research Part B: Methodological* 45(1): 232–254. http://dx.doi.org/10.1016/j.trb.2010.05.002

LTA. 2013. *ONE.MOTORING Services*. Land Transport Authority (LTA). Available from Internet: http://www.onemotoring.com.sg/publish/onemotoring/en/imap.html

Mahmassani, H.; Herman, R. 1984. Dynamic user equilibrium departure time and route choice on idealized traffic arterials, *Transportation Science* 18(4): 362–384. http://dx.doi.org/10.1287/trsc.18.4.362

Mounce, R.; Carey, M. 2011. Route swapping in dynamic traffic networks, *Transportation Research Part B: Methodological* 45(1): 102–111. http://dx.doi.org/10.1016/j.trb.2010.05.005

Newell, G. F. 1993a. A simplified theory of kinematic waves in highway traffic, part I: general theory, *Transportation Research Part B: Methodological* 27(4): 281–287. http://dx.doi.org/10.1016/0191-2615(93)90038-C

Newell, G. F. 1993b. A simplified theory of kinematic waves in highway traffic, part II: queueing at freeway bottlenecks, *Transportation Research Part B: Methodological* 27(4): 289–303. http://dx.doi.org/10.1016/0191-2615(93)90039-D

Newell, G. F. 1993c. A simplified theory of kinematic waves in highway traffic, part III: multi-destination flows, *Transportation Research Part B: Methodological* 27(4): 305–313. http://dx.doi.org/10.1016/0191-2615(93)90040-H

Newell, G. F. 1988. Traffic flow for the morning commute, *Transportation Science* 22(1): 47–58. http://dx.doi.org/10.1287/trsc.22.1.47

Nie, Y. 2012. Transaction costs and tradable mobility credits, *Transportation Research Part B: Methodological* 46(1): 189–203. http://dx.doi.org/10.1016/j.trb.2011.10.002

Ramadurai, G.; Ukkusuri, S. V.; Zhao, J.; Pang, J.-S. 2010. Linear complementarity formulation for single bottleneck model with heterogeneous commuters, *Transportation Research Part B: Methodological* 44(2): 193–214. http://dx.doi.org/10.1016/j.trb.2009.07.005

Richards, P. I. 1956. Shock waves on the highway, *Operations Research* 4(1): 42–51. http://dx.doi.org/10.1287/opre.4.1.42

Salmon, F. 2010. *The Congestion Pricing Debate*. 4 June 2010. Available from Internet: http://blogs.reuters.com/felix-salmon/2010/06/04/the-congestion-pricing-debate-cont

Seik, F. T. 1997. An effective demand management instrument in urban transport: the area licensing scheme in Singapore, *Cities* 14(3): 155–164. http://dx.doi.org/10.1016/S0264-2751(97)00055-3

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

Small, K. A.; Verhoef, E. T. 2007. *The Economics of Urban Transportation*. Routledge. 296 p.

STA. 2013. *Congestion taxes in Stockholm and Gothenburg*. Swedish Transport Agency (STA). Available from Internet: http://www.transportstyrelsen.se

Szeto, W. Y.; Lo, H. K. 2004. A cell-based simultaneous route and departure time choice model with elastic demand, *Transportation Research Part B: Methodological* 38(7): 593–612. http://dx.doi.org/10.1016/j.trb.2003.05.001

Teodorović, D.; Triantis, K.; Edara, P.; Zhao, Y.; Mladenović, S. 2008. Auction-based congestion pricing, *Transportation Planning and Technology* 31(4): 399–416. http://dx.doi.org/10.1080/03081060802335042

TfL. 2013. *Congestion Charge*. Transport for London (TfL). Available from Internet: http://www.tfl.gov.uk/roadusers/congestioncharging

TfL. 2008. *Central London Congestion Charging: Impacts Monitoring*. Sixth Annual Report. Transport for London (TfL). 227 p. Available from Internet: https://www.tfl.gov.uk/cdn/static/cms/documents/central-london-congestion-charging-impacts-monitoring-sixth-annual-report.pdf

Tian, L.-J.; Yang, H.; Huang, H.-J. 2013. Tradable credit schemes for managing bottleneck congestion and modal

split with heterogeneous users, *Transportation Research Part E: Logistics and Transportation Review* 54: 1–13. http://dx.doi.org/10.1016/j.tre.2013.04.002

Tsekeris, T.; Voß, S. 2009. Design and evaluation of road pricing: state-of-the-art and methodological advances, *Netnomics: Economic Research and Electronic Networking* 10(1): 5–52. http://dx.doi.org/10.1007/s11066-008-9024-z

Van den Berg, V. A. C. 2014. Coarse tolling with heterogeneous preferences, *Transportation Research Part B: Methodological* 64: 1–23. http://dx.doi.org/10.1016/j.trb.2014.03.001

Van den Berg, V. A. C. 2012. Step-tolling with price-sensitive demand: why more steps in the toll make the consumer better off, *Transportation Research Part A: Policy and Practice* 46(10): 1608–1622. http://dx.doi.org/10.1016/j.tra.2012.07.007

Van den Berg, V.; Verhoef, E. T. 2011. Congestion tolling in the bottleneck model with heterogeneous values of time, *Transportation Research Part B: Methodological* 45(1): 60–78. http://dx.doi.org/10.1016/j.trb.2010.04.003

Vickrey, W. 1992. *Principles of Efficient Congestion Pricing*. Available from Internet: http://www.vtpi.org/vickrey.htm

Vickrey, W. S. 1969. Congestion theory and transport investment, *The American Economic Review* 59(2): 251–260.

Xiao, F.; Qian, Z.; Zhang, H. M. 2013. Managing bottleneck congestion with tradable credits, *Transportation Research Part B: Methodological* 56: 1–14. http://dx.doi.org/10.1016/j.trb.2013.06.016

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. http://dx.doi.org/10.1007/s11067-010-9141-8

Xiao, F.; Shen, W.; Zhang, H. M. 2012. The morning commute under flat toll and tactical waiting, *Transportation Research Part B: Methodological* 46(10): 1346–1359. http://dx.doi.org/10.1016/j.trb.2012.05.005

Yang, H.; Huang, H.-J. 2005. *Mathematical and Economic Theory of Road Pricing*. Elsevier Science. 486 p.

Yang, H.; Huang, H.-J. 1997. Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory, *Transportation Research Part B: Methodological* 31(6): 425–440. http://dx.doi.org/10.1016/S0191-2615(97)00005-2

Yang, H.; Wang, X. 2011. Managing network mobility with tradable credits, *Transportation Research Part B: Methodological* 45(3): 580–594. http://dx.doi.org/10.1016/j.trb.2010.10.002

Yao, T.; Friesz, T. L.; Wei, M. M.; Yin, Y. 2010. Congestion derivatives for a traffic bottleneck, *Transportation Research Part B: Methodological* 44(10): 1149–1165. http://dx.doi.org/10.1016/j.trb.2010.03.002

Zhang, H. M.; Ge, Y. E. 2004. Modeling variable demand equilibrium under second-best road pricing, *Transportation Research Part B: Methodological* 38(8): 733–749. http://dx.doi.org/10.1016/j.trb.2003.12.001

Zhang, H. M.; Nie, Y.; Qian, Z. 2013. Modelling network flow with and without link interactions: the cases of point queue, spatial queue and cell transmission model, *Transportmetrica B: Transport Dynamics* 1(1): 33–51. http://dx.doi.org/10.1080/21680566.2013.785921