Transportation Research Part B 83 (2016) 59–73

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Bottleneck congestion: Differentiating the coarse charge

Jasper Knockaert[a,*], Erik T. Verhoef[a,b], Jan Rouwendal[a,b]

[a] VU University Amsterdam, Netherlands
[b] Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam, Netherlands

ARTICLE INFO

Article history:
Received 27 September 2010
Revised 2 November 2015
Accepted 2 November 2015
Available online 11 December 2015

Keywords:
Traffic congestion
Road pricing
Bottleneck model
Step tolls

ABSTRACT

The traditional bottleneck model for road congestion promotes the implementation of a triangular, and time varying, charge as the optimal solution for the road congestion externality. However, cognitive and technological barriers put a practical limit to the degree of differentiation real world implementations can handle. The traditional approach to accommodate for this concern has been a step toll, with the single step *coarse* charge as its simplest case.

In this paper we study how efficiency of the coarse charge can be improved by differentiating its level and timing across groups of travellers. We use the traditional bottleneck model to analyse how the coarse charge can be differentiated over two groups of travellers assuming inelastic peak-hour demand.

The results of our analysis indicate that differentiating the coarse charge across two groups of travellers considerably improves its efficiency without increasing cognitive effort and decision making costs for the individual traveller. A numeric illustration reveals a welfare gain of 69% of the first-best charge, up from 53% for the generic coarse charge. This increase is similar to what is obtained by moving from the coarse charge to a generic two step toll. Once different groups have been defined, one could in fact achieve the same gains by temporal separation of drivers, for example by use of licence plate numbers.

The presented charging regime has a considerable degree of flexibility with respect to the share of travellers to attribute to each scheme, which further adds to its merits in practical applicability.

1. Introduction

The bottleneck model, first introduced by Vickrey (1969) and elaborated by Arnott et al. (1990, 1993) has become the reference model for peak hour road congestion. The model is simple, appeals to intuition and its first-best solution is analytically tractable. The suggestion of the model is that a time-varying toll can completely eliminate queuing before the bottleneck, while the revenue generated by the toll is a pure welfare gain.

The subsequent literature on bottleneck charging focused on a number of extensions, including heterogeneous travellers, interactions between parallel or serial bottlenecks, the presence of untaxed alternatives, and so on. Overviews are presented by Arnott et al. (1998), Lindsey and Verhoef (2001), among others.

* Corresponding author. Tel.: +31205986090.
E-mail address: jasper.knockaert@vu.nl (J. Knockaert).

http://dx.doi.org/10.1016/j.trb.2015.11.004

Recent developments in the literature on bottleneck congestion include a more general specification of commuter preferences and more attention to the spatial context of the bottleneck. It is conventional in the bottleneck model to assume that preferences are linear in travel time, scheduling delay early and scheduling delay late (sometimes referred to as $\alpha - \beta - \gamma$ preferences).[1] One implication of this is that departing 1 min earlier from home always implies the same additional cost, whereas in reality it may be more problematic to depart 1 min earlier when one departs at seven than at half past eight when the children have left for school. Vickrey (1973) already proposed a specification in which utility is the sum of a function of departure time, which can be interpreted as the utility of spending time at home, and a function of arrival time, which can be interpreted as referring time spent at work. Tseng and Verhoef (2008) provide evidence, based on stated preference research, in favour of this more general specification. Fosgerau and de Palma (2012) use an even more general specification in which utility is a concave function of departure and arrival time. The conventional $\alpha - \beta - \gamma$ preferences are a special case that may still provide a useful approximation for the analysis of specific cases. However, the more general framework and the associated empirical evidence make clear that one should not expect the values of these $\alpha - \beta - \gamma$ parameters to be equal in all circumstances.

The bottleneck model as described by Arnott et al. (1990) explicitly assumes any spatial context away by setting travel time before and after passing the bottleneck equal to zero. The queue that the model predicts under imperfect tolling is therefore sometimes referred to as *vertical*. Arnott et al. (1991) already relaxed this assumption by combining the bottleneck with a model of the use of parking space in the vicinity of the work location that is reached through the bottleneck. It seems indeed natural to locate the bottleneck between the residential quarters and the CBD and some recent contributions do so, but concentrate attention on the other side of the bottleneck. Fosgerau and de Palma (2012) consider welfare maximising tolling in the context of an urban area where commuters have to pass the bottleneck in order to arrive at work. The differences in residential locations – which the authors take as given – cause heterogeneity among the commuters, whose preferences are of the generalised type discussed above. The authors show that the queue is still unimodal, as in the standard bottleneck model, but that commuters living close to the bottleneck will in general lose from the introduction of the optimal toll. Gubins and Verhoef (2014) use a similar spatial setting but integrate the preferences for commuting with those for land, assuming that an individual derives more utility from housing when spending more time in it, ceteris paribus leading to a higher willingness to pay for housing and land. This allows them to reconsider the implications of urban congestion for urban structure, which had formerly only been considered for static congestion models. These authors use the general setup for preferences for commuting suggested by Vickrey (1973) to connect commuting with the demand for space, but they use the special case of that model that makes it behaviourally equivalent to the conventional case of $\alpha - \beta - \gamma$ preferences.

Other aspects that have recently received attention are the impact of uncertainty (e.g. de Palma and Fosgerau, 2013), the generalisation to heterogeneous commuters (e.g. Liu et al., 2015), the connection with parking space (Yang et al., 2013) and the impact of teleworking (Gubins and Verhoef, 2011) this suffices to show that the bottleneck model is still an active area for research in urban economics. While the existing analyses of bottleneck charging are generally illuminating, barriers have been identified in implementing its policy recommendations straightforwardly. As noted by Arnott et al. (1990), the implementation of a first-best time varying charge is technically demanding. The continuous time variation of the charge probably also requires a considerable cognitive effort by the traveller. It is to be expected that effectiveness of charging reaches a limit or even decreases when search and decision costs become prohibitive as a result of excessively complex differentiation (see for instance Norwood, 2006).

Arnott et al. (1990) present an optimal coarse charge which yields about half of the efficiency of the first-best scheme (Arnott et al., 1993), but which will be more easy to understand and communicate. In this equilibrium, the queue falls to a zero length just before the toll is raised in the early part of the peak. The average queue length is therefore shorter than in the no-toll equilibrium. Also in the late part of the peak, the queue drops to a zero length before everyone has passed, and right before the toll is lifted, but then a mass departure follows which makes the analytical solution somewhat complicated (Arnott et al., 1990). With multi-step tolls, the queue falls to zero more often during the peak, and average travel delay decline further, and social welfare rises. But of course, the toll schedule becomes increasingly complex.

In this paper we present an approach that allows for an improved efficiency of coarse bottleneck congestion charging, without increasing the complexity of the toll schedule for an individual traveller, by allowing for differentiation across groups of travellers. The general idea is that different groups face different moments at which their group specific coarse toll is switched on and off, and different corresponding toll levels. The resulting temporal separation of travellers allows an equilibrium in which the queue falls back to a zero length multiple times in the early part of the peak, just like it would in a multi-step toll equilibrium. And similarly, there will be multiple smaller mass departures in the late part of the peak. The traffic pattern and social welfare replicates what would be achieved with a multi-step toll, while each traveller faces a simple coarse toll schedule.

Interestingly, once the scheme is in operation, group members will not travel in the period where the coarse toll for their group is positive. A rationing scheme, banning them from that same time period rather than removing them through a toll they choose not to pay, would thus lead to the same outcome.

This is a remarkable result, as it is an example of how a non-price measure could be as efficient as a second-best tax instrument. Given the social opposition against the use of tax instruments in the management of congestion, this is a finding of potentially great policy relevance.

[1] Small (1982) provides empirical support.