2010

DISCUSSION PAPERS

2

Joint Transport Research Centre

ROUND TABLE
4-5 February 2010, Paris

International Transport Forum

ROAD PRICING WITH COMPLICATIONS

Mogens FOSGERAU, DTU Transport, Denmark & Centre for Transport Studies, Sweden

Kurt VAN DENDER
OECD/ITF Joint Transport Research Centre

OECD

Discussion Paper No. 2010-2

Prepared for the ITF/OECD Round Table of 4-5 February 2010 on

Implementing Congestion Charging

Road pricing with complications

Mogens FOSGERAU

DTU Transport, Denmark &
Centre for Transport Studies, Sweden

Kurt VAN DENDER

OECD/ITF Joint Transport Research Centre

Updated October 2012

The views expressed in this paper are the authors', and do not necessarily represent the opinions of DTU Transport, Denmark, the Centre for Transport Studies, Sweden, the International Transport Forum or the OECD.

INTERNATIONAL TRANSPORT FORUM

The International Transport Forum at the OECD is an intergovernmental organisation with 54 member countries. It acts as a strategic think tank with the objective of helping shape the transport policy agenda on a global level and ensuring that it contributes to economic growth, environmental protection, social inclusion and the preservation of human life and well-being. The International Transport Forum organizes an annual summit of Ministers along with leading representatives from industry, civil society and academia.

The International Transport Forum was created under a Declaration issued by the Council of Ministers of the ECMT (European Conference of Ministers of Transport) at its Ministerial Session in May 2006 under the legal authority of the Protocol of the ECMT, signed in Brussels on 17 October 1953, and legal instruments of the OECD.

The Members of the Forum are: Albania, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bosnia-Herzegovina, Bulgaria, Canada, Chile, China, Croatia, the Czech Republic, Denmark, Estonia, Finland, France, FYROM, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Italy, Japan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Moldova, Montenegro, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, the United Kingdom and the United States.

The International Transport Forum's Research Centre gathers statistics and conducts co-operative research programmes addressing all modes of transport. Its findings are widely disseminated and support policymaking in Member countries as well as contributing to the annual summit.

Discussion Papers

The International Transport Forum's Discussion Paper Series makes economic research, commissioned or carried out at its Research Centre, available to researchers and practitioners. The aim is to contribute to the understanding of the transport sector and to provide inputs to transport policy design. The Discussion Papers are not edited by the International Transport Forum and they reflect the authors' opinions alone.

The Discussion Papers can be downloaded from:

www.internationaltransportforum.org/jtrc/DiscussionPapers/jtrcpapers.html

The International Transport Forum's website is at: www.internationaltransportforum.org

For further information on the Discussion Papers and other JTRC activities, please email:

itf.contact@oecd.org

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

ROAD PRICING WITH COMPLICATIONS[1]

Abstract

The rationale for congestion charges is that by internalising the marginal external congestion cost, they restore efficiency in the transport market. In the canonical model underlying this view, congestion is a static phenomenon, users are taken to be homogenous, there is no travel time risk, and a highly stylised model of congestion is used. The simple analysis also ignores that real pricing schemes are only rough approximations to ideal systems and that inefficiencies in related markets potentially affect the case for congestion charges. The canonical model tends to understate the marginal external congestion cost because it ignores user heterogeneity and trip timing inefficiencies. With respect to the relevance of interactions between congestion and congestion charges and tax distortions and distributional concerns, recent insights point out that there is no general case for modifying charges for such interactions. Therefore the simple Pigouvian rule remains a good first approximation for the design of road charging systems.

[1] We are grateful to Ken Small and two referees for many valuable comments on earlier versions of this paper. Errors and shortcomings are ours. Views expressed are the authors' and not necessarily those of the institutions they are affiliated with. This paper updates and extends some of the material discussed in Fosgerau and Van Dender (2010).

TABLE OF CONTENTS

1. INTRODUCTION

Road tolls can be used as a tool to reduce demand for travel when and where that is thought to be beneficial. They have important advantages over other ways of reducing demand. By adding a toll to the cost of a trip, just those trips are removed that travellers themselves think are not worth the toll. So tolling ensures that the least beneficial trips are eliminated first with drivers themselves assessing the benefits of their trips. Other ways of reducing demand, e.g. banning certain types of vehicles or license plates, or selective network capacity reductions, do not reflect individuals' valuations of travel in the same way, and therefore risk producing much lower or even negative benefits to society.

Tolling allows the individual decisions about whether, when and where to travel to remain decentralised. Drivers compare their benefits from a trip to its total cost, which includes the toll and the time cost of travel. When there is congestion, travel time increases with traffic volumes, and each additional driver imposes extra time costs on other drivers. But these costs are external, so they do not affect individuals' decisions. The result is that there is more travel than is desirable from a social point of view. Tolls can help reduce traffic volumes, but how high should they be? A toll equal to the total cost of delay imposed on other drivers by one additional car, the marginal external congestion cost, ensures that the socially optimal level of congestion results.[2] Exactly those trips are then carried out that are worth the full cost. The toll payment is a loss for drivers, but the money does not disappear so it is not a loss to society. The cost to drivers who choose to pay the toll is offset by the gains of those who receive the toll revenues.[3] The fact that all drivers have to pay in order to deter the least beneficial trips is an obstacle to the acceptance of road pricing, because – except in special cases – drivers as a group incur a loss of benefits when the use of the revenues from road pricing is not taken into account. But when the revenue is accounted for, the welfare gain from pricing congestion can be considerable.

The basic rationale for congestion tolls, which we discuss in more detail in Section 2, is derived in a "canonical model" that (naturally) relies on a range of simplifying assumptions. This paper is an overview of what happens when one opens the door for some real-world complications.

[2] This is true in a first-best world, where all other prices equal marginal social cost.

[3] This is also true in a first-best world.

Each of the five complications that we discuss has been the subject of particularly active research over the past decade or so, with insights now sufficiently mature to assess impacts on recommendations regarding the rationale underlying congestion charges. While our review does not claim to be comprehensive, it does cover the major innovations in understanding congestion and the arguments for congestion charging.

First, a brief discussion of the bottleneck model highlights the importance of schedule delay costs as part of the cost of congestion (Section 3). These costs are overlooked by the canonical model because it adopts a static view of congestion. Second, in contrast to what the simple model assumes, travellers are not identical. On the contrary, the evidence suggests they differ strongly in how they value travel time. The value of travel time varies among individuals, and for the same individual it depends on the exact context of a trip. Section 4 explores the consequences of accounting for such heterogeneity. A third complication, the implications of which are investigated in Section 5, is that travel time is not deterministic but instead tends to become more variable and less predictable as flow approaches the capacity of the road. This means that congestion causes not just the average travel time to increase but also travel time variability. This uncertainty entails a significant additional cost for travellers, which should ideally be accounted for in tolls. But insights on how to do that are only emerging.

The fourth complication, briefly addressed in Section 6, is that the relationship between traffic flow and travel time is considerably more complicated than assumed in the canonical model, which simply posits that travel time increases with volume. Section 7 addresses a fifth complication, which is that actual tolling systems are imperfect approximations to efficient tolls. One problem is that it may be prohibitively costly to implement the theoretically optimal spatial and temporal differentiation of tolls. Detailed traffic models are a good way of finding reasonable approximations. A second issue is that there are inefficiencies in the broader economy that are potentially relevant to determining the optimal toll. For example, literature has highlighted the possibility that tolls exacerbate labour supply distortions and suggest this should lead to low tolls for commuting trips. However, recent work on the optimal provision of public goods strongly suggests such distortions are not relevant, and instead tolls should reflect marginal external congestion costs and nothing else. Section 7 sums up.

Our discussion focuses on innovations in the understanding of congestion and congestion charges, highlighting complexities associated with the principle of “getting the prices right”. It does not in itself clarify how important it is to get the prices right, i.e. how much can be gained from an improved management of external congestion costs.[4] Available estimates of aggregate costs of excessive congestion produce figures of a couple percent of GDP. These numbers have been criticized for their reliance on overly simplistic methods, e.g. when actual travel speeds are compared to speed limits and no mention is made of efficient levels of congestion.

[4] Net benefits also depend on implementation costs. We do not discuss those in this paper, but emerging evidence suggests these costs are high – if the main issue is to raise revenue then cheaper ways of doing so exist (OECD, 2010).

The issues reviewed in this paper do little to inspire more faith in such estimates, but also provide little guidance for alternative ways of establishing aggregate cost estimates. To the contrary, we emphasize the very strong dependence of inefficiencies related to congestion on local circumstances.

That external costs of congestion are high in some places at some times is beyond doubt. External congestion costs are found to dominate other external costs even when aggregated over large areas and time periods (see e.g. the evidence in Parry and Small (2005) and Small and Van Dender (2007)). The case studies discussed in De Borger and Proost (2001) show welfare gains from introducing optimal transport pricing between 0.5% and 1.5% of welfare in a number of European cities, in a model that ignores most complexities discussed in this paper and that therefore likely underestimates the true benefits of optimal congestion charges.

Indeed, we emphasize that gauging congestion costs by referring to travel speed only misses out on important aspects of the problem, including schedule delays, lack of reliability, and heterogeneity of travel time values. Raux (2005) is an illustration, for the London congestion charge, of how applying simple corrections to a basic cost-benefit analysis to account for some of these issues can change a negative assessment into a favourable one.

2. THE STANDARD TEXTBOOK ANALYSIS

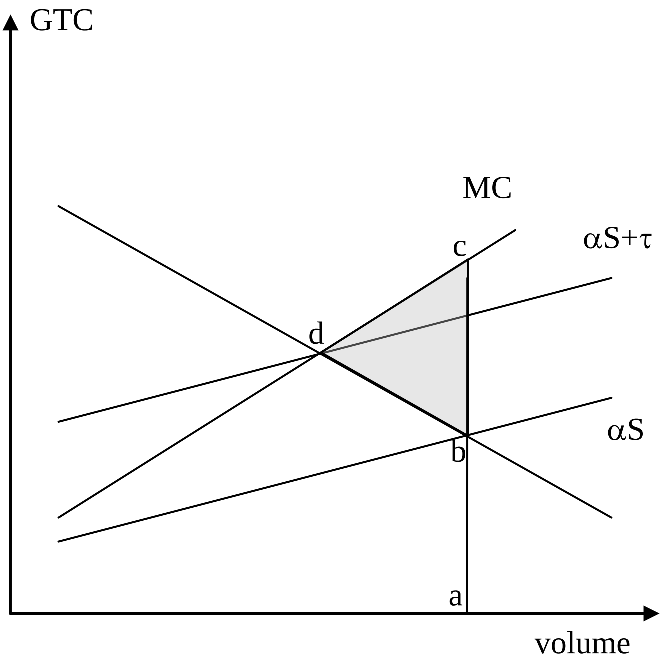
The basic argument for congestion tolls is as follows. When deciding if, when, and how to travel, travellers consider what the various available options will cost them. They choose to drive when the benefit of doing so outweighs the cost of travelling and their personal net gain is at least as large as for other choice options. However, travellers do not take into account the delay they impose on others: the marginal congestion cost is external. That delay can become so large that an additional trip entails a net loss to society, despite the private benefit. Eliminating those trips for which the private benefits are below the social costs of the trip (i.e. the sum of private costs and external costs) increases the aggregate net benefits from travel. The optimal congestion toll achieves this goal by charging the drivers for the costs that they impose on others, at that level of traffic where marginal private benefits equal marginal social trip costs.

The standard analysis is summarized in the following figure. Traffic volume is measured on the horizontal axis, and the vertical axis represents the generalized travel cost (*GTC*).

The curve αS , is the supply curve. The supply S is the travel time as a function of traffic volume, and this is multiplied by the value of travel time (VTT, denoted α) to convert travel time to a money cost.[5] The supply curve slopes upward to reflect a situation with congestion. In the figure it is linear; in general it is convex. All travellers experience the same travel time and the same VTT, so the supply curve can also be understood as an average cost curve. The curve MC indicates the marginal cost. It is the change in total cost arising from an additional traveller. When the supply curve is increasing, the marginal cost curve will always have a larger slope than the supply curve and lie above it. The demand curve slopes downward to reflect that demand decreases in the generalized price GTC.

[5] Other monetary travel costs are ignored.

Figure 2.1. **Graphical summary of the canonical model**



The market equilibrium occurs at the intersection of the demand curve D with the supply curve αS , at the point b, as at this point the cost of the last trip from the private point of view is equal to the benefit of that same last trip. But this equilibrium is not efficient because the marginal cost of all trips between points d and c is higher than the private benefit of those trips: these trips are wasteful from the social perspective but are made nevertheless because travellers consider private costs, not social costs. The total waste is equal to area bcd. The efficient outcome occurs where the marginal benefits given by the demand curve are equal to the marginal social costs, i.e. at point d. The optimal congestion toll τ allows this efficient situation to be attained as a decentralized equilibrium. Compared to the no-toll equilibrium, there is a welfare gain equal to the area bcd, i.e. it is equal to the avoided waste.

The optimal toll τ makes up for the fact that drivers ignore time costs imposed on other drivers, so that the relevant cost curve from the individual user's point of view now is αS+τ instead of just αS. The toll in effect puts a price on time, reflecting the value of the time loss that an additional driver causes for other drivers. To design an efficient toll the first requirement then is to find the VTT, so that time losses can be charged for in money. It turns out, however, that the assumption of homogeneity, which allows one to work with a single supply curve, is difficult to maintain given the strong observed heterogeneity of travel time valuations. This is discussed in detail in Section 4. Section 3 discusses an alternative way of modelling congestion which highlights other shortcomings of the canonical model, namely that people care about when they travel instead of just about how long their trip takes, and that congestion is a dynamic phenomenon.

3. A DYNAMIC VIEW – THE BOTTLENECK MODEL

The bottleneck model captures two features on traffic congestion that are glossed over by the canonical model, which is a static model. First, the static analysis ignores the trip timing aspect of travel demand. It views time merely as a resource of which a traveller can spend more or less in transport, whereas in reality they value a trip starting at, say, 8 AM differently from one at 9 AM. This is because travellers care about when they will arrive at their destination and/or when they leave from the trip origin. The bottleneck model of congestion emphasizes that congestion arises because travellers prefer to travel around the same times, e.g. because their work hours are similar. If they could be spread evenly over time, there would be no congestion. Second, congestion is inherently dynamic, since adding a vehicle to a queue at some instant will affect the evolution of the queue until it is gone. To capture these aspects, the bottleneck model (Vickrey, 1969; Arnott et al., 1993) describes the time dimension more explicitly than the static model. Even in its simplest guise, it reveals a number of important insights regarding the pricing of congestion. One of them is that large efficiency gains are obtained through the effect of pricing on trip timing.

In the simplest incarnation of the bottleneck model, travellers are homogenous with VTT α. They have a common preferred arrival time t^* and prefer not to be early or late at the destination relative to this time. The cost per unit of time of earliness is usually denoted by β and the cost of lateness by γ. For a traveller departing at time t_d and arriving at time t_a, the generalised travel cost is then $GTC(t_d,t_a)=\alpha\cdot(t_a-t_d)+\beta\cdot\min(0,t^*-t_a)+\gamma\cdot\max(0,t_a-t^*)$.

This specification of scheduling preferences was introduced by Vickrey (1969) and later estimated by Small (1982). [6]

Assume next that a total of D travellers have to pass through a bottleneck in order to reach their destination. The bottleneck has a limited capacity of ψ users per minute. Assume for simplicity that travel time is zero before and after the bottleneck. These simplifications allow us to focus on time spent queuing only. Then the first traveller, departing at time t_0, also arrives at the destination at this time, since there is no queue yet at this time. Denote the cumulative arrival rate at the bottleneck by *R*. Then $R(t_0)=0$. Let the time of the last departure be t_1, such that $R(t_1)=D$.

[6] De Palma and Fosgerau (2011) develops the bottleneck model under more general scheduling preferences.

The equilibrium distribution of departure times, in which no traveller has an incentive to change their behaviour, is reached when no traveller can reduce their cost by changing departure time. This implies that there is always a queue during the interval $[t_0, t_1]$ and that the queue has dissipated at time t_1, such that $\psi \cdot (t_1 - t_0) = D$. Equilibrium also requires that the GTC is constant over the interval where travellers depart and higher outside. A queue builds up immediately as the first traveller departs, since travellers initially depart at a higher rate than capacity. The queue has maximum length at the departure time when a traveller would be at the destination exactly at the preferred time t^*: this traveller faces the longest queuing costs but no schedule delay costs. From that point onwards, the departure rate is lower than capacity such that the queue gradually dissipates and is gone at the time of the last departure precisely.

The first and last travellers experience no queue in this model. The first traveller is early at the destination while the last traveller is late. They incur the same *GTC* in equilibrium, which implies that $-\beta \cdot (t^* - t_0) = \gamma \cdot (t_1 - t^*)$. This fixes the interval $[t_0, t_1]$ which allows the equilibrium travel cost to be computed. The total travel cost in equilibrium for all travellers is given by $TC = \frac{\beta\gamma}{\beta + \gamma}\frac{N^2}{\psi}$, such that the marginal external congestion cost (i.e. the difference between the marginal time cost and the average time cost) is $mecc = \frac{\beta\gamma}{\beta + \gamma}\frac{N}{\psi}$. This is the marginal external congestion cost associated with the addition of a marginal user to the equilibrium. It increases in the number of users and decreases in capacity.

We may regard the number of travelers D as being a function of the equilibrium generalised travel cost. Then, connecting the bottleneck model with the simple static analysis discussed in the previous section, we would find that the optimal static toll, or uniform toll, is equal to the mecc. This toll would not remove congestion. The number of travellers would be reduced, but there would still be a queue during $[t_0, t_1]$. But since spending time in a queue is wasteful from the social perspective, the question arises what can be done to eliminate the queue? The answer is to allow the toll to vary over time. The optimal time-varying toll, which eliminates the queue, is zero at time t_0. It increases until the preferred arrival time t^*, then it decreases again, until it is zero at time t_1. The average toll is equal to the optimal static toll. The optimal time-varying toll removes congestion completely, since it modifies departure time choices to ensure that travellers arrive at the bottleneck exactly at the rate ψ, which is the bottleneck capacity.

The toll replaces queueing time by a money cost. This is not better from a driver's point of view (the generalized trip cost is the same) but it is better from the social point of view because queuing time is lost to society whereas toll payments are not.

Whether the bottleneck model is a better representation of the technology of congestion than the static flow model, is a question with no general answer.

The models highlight different features what causes congestion from a supply point of view, and both are partial. Section 6 discusses supply side issues in a bit more detail. A clear merit of the bottleneck model is that it highlights departure time choices and schedule delay costs, as opposed to just the travel time losses emphasized in the static model. The costs of congestion hence are higher than appears from the canonical model.

The bottleneck model can also be used to show that some policies that do not use pricing may nevertheless be effective in dealing with congestion. One such type of policy reserves a part of capacity, e.g. using ramp metering or special lanes for certain classes of road users (Fosgerau, 2011). This can reduce queueing by effectively splitting the queue into two, one of which last for a shorter time than would otherwise have been the case.[7]

[7] A recent paper (Cassidy et al., 2010) indicates that special lanes such as car pool lanes can increase effective road capacity, because they reduce disruptive vehicle lane changing. Even a severely underused carpool lane can in some instances increase a freeway bottleneck's total discharge flow. A theoretical investigation of these issues is undertaken in Menendez&Daganzo (2007).

4. HETEROGENEOUS TRAVELLERS

The discussion up to now has assumed that users have the same value of travel time VTT, but it is clear that in reality users are heterogeneous. This section explores the measurement of heterogeneous VTT (Section 4.1) and the consequences of that heterogeneity for the optimal congestion charge (Section 4.2). It turns out that heterogeneity is large and tends to imply an upward revision of the optimal toll compared to the canonical model.

4.1. Measurement of heterogeneous VTT

How can the VTT of travellers be inferred? The bottom line is to use information on choices where travellers have the option of paying for faster travel, under otherwise equal or at least similar conditions. If a traveller has a choice between two options for making a trip, where one is faster but more expensive than the other, then the traveller faces a trade-off between money and time. Denote travellers' generalised cost of a trip by $GTC = C + \alpha \cdot T$, where C is the monetary cost of the trip, T is the travel time, and α is an individual specific VTT, and the difference in GTC between two trip options for a traveller with VTT by $\Delta GTC = \Delta C + \alpha \cdot \Delta T$. Holding everything else constant, travellers with $\alpha < -\Delta C/\Delta T$ will choose the slow option while travellers with $\alpha > -\Delta C/\Delta T$ will choose the fast option. The trade-off thus entails an implicit price of travel time, namely $\varpi = -\Delta C/\Delta T$. Through his choice, a traveller reveals whether his VTT is larger or smaller than the trade-off price, i.e. whether $\alpha < \varpi$ or $\alpha > \varpi$.

Next, label by Φ the cumulative distribution function describing the distribution of the VTT among users. Observing many travellers at a trade-off price ϖ allows assessment of the share of travellers with $\alpha < \varpi$. This share is the cumulative distribution evaluated at the point ϖ, i.e. $\Phi(\varpi)$. Observing travellers in different choice situations with different ϖ allows assessment of Φ over the range where ϖ varies. To assess Φ completely, it is necessary to observe choice shares for values of ϖ ranging from a point where travel time is cheap and all travellers choose the fast and expensive option to a point where time is expensive and all travellers choose the cheap and slow option.[8]

[8] This may sound easier than it is. Fosgerau (2006) discusses issues related to the identification of the distribution of VTT. De Borger & Fosgerau (2008) discuss extensions to take behavioral anomalies into account.

Data on travel choices can reflect actually observed behaviour (revealed preference) or hypothetical choices presented in surveys (stated preference). In general, revealed preference data are preferable by virtue of relating to real choices where travellers actually feel the consequences of their choices. Suitable revealed preference data could come from situations where travellers face a choice between a slow and cheap route and a fast and expensive route, as found for example in routes combining free access regular freeway lanes with tolled express lanes (Small et al., 2005). With such data, however, it is often difficult to achieve the necessary variation in the price of time needed to reveal the complete distribution of the VTT. This is a reason for relying on stated preference data, where travellers make choices between hypothetical options. It is possible to construct stated preference choice situations to meet many of the demands of econometric modelling. Stated preference data are, however, perennially tainted by doubt whether they represent actual behaviour well.

Figure 4.1. **Confidence band for the cumulative distribution of VTT based on stated preference data. The unit is Danish Crowns (DKK) per hour: 1 EUR ≈ 7.5 DKK.**

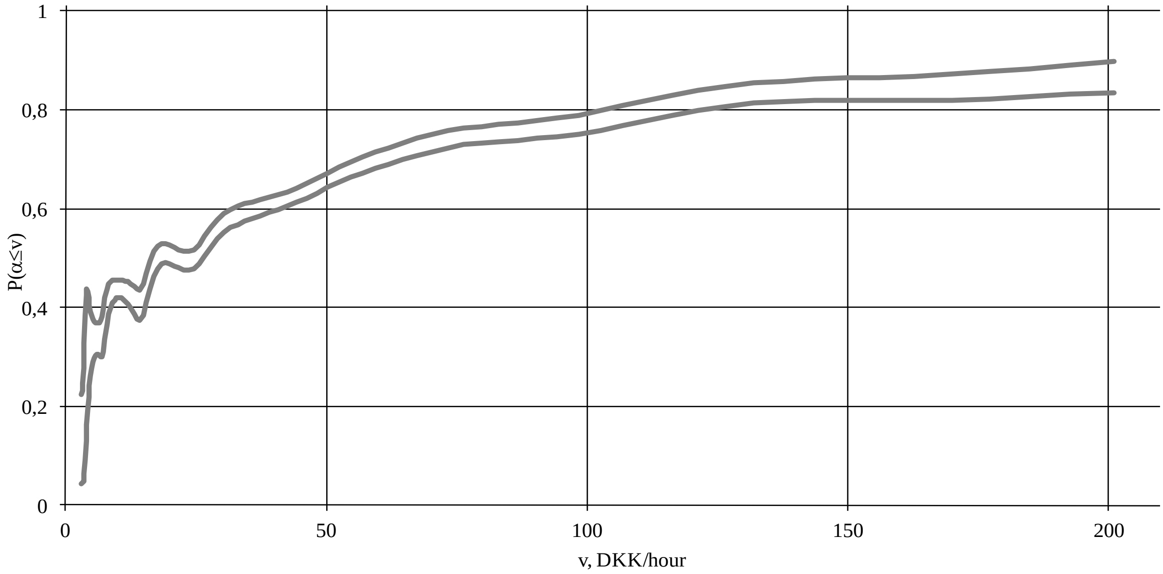


Figure 4.1 shows an estimate of a VTT distribution obtained from stated preference data (Fosgerau, 2006).[9] The shape is broadly typical of many studies on the subject. It shows a right-skewed distribution with many travellers having low VTT and few travellers having large VTT. The median VTT is about 25 DKK/hour, approximately 4.7€/hour, while the mean is considerably larger.

[9] The estimate is computed using a nonparametric technique, which does not impose the restriction that the cumulative distribution should be increasing. It is therefore evidence of the internal validity of the SP data that an increasing function does result.

This estimate of the cumulative distribution of the VTT does not show the maximum of the VTT distribution. The largest trade-off price of time that was offered to respondents in the stated preference exercise was about 200 DKK/hour (approximately 27 Euro/hour). A significant share of respondents, about 15%, indicated that they were willing to pay this amount per hour of travel time saved and hence that their VTT was higher than 200 DKK/hour. How much larger is impossible to say based on the figure, as the survey did not ask about values exceeding 200 DKK/hour. In the absence of observations on the right tail, a complete description of the distribution requires making restrictive assumptions, which may be hard to justify. One popular approach is to assume a specific form for the VTT distribution. This allows calculating the mean VTT, but the result is extremely sensitive to the assumed form of the distribution (Fosgerau, 2006). As will be seen in the next subsection, this poses problems for calculating optimal congestion tolls.

It is clear that there is enormous variation in the VTT with several orders of magnitude from low to high. The VTT depends on observable and unobservable factors. It is generally found to increase with income, although the size of the income elasticity is debated. The VTT is generally thought to vary substantially between individuals but also within individuals depending on the context. In general, a large part of the variation in VTT remains after controlling for observable factors. E.g., Fosgerau (2006) controls for gender, income, trip duration, time difference between alternatives, share of delay time due to congestion in travel time, age and trip purpose, and finds that the remaining variation in VTT has more than a factor 50 between the 20th and 80th percentile of the VTT distribution.

4.2. Road pricing with heterogeneous travellers

Travel time differs strongly among road users. Here we discuss how this means that tolls are usually higher than a direct application of the canonical model purports, and how it implies there are benefits to toll differentiation on parallel lanes or roads.

First, how is the marginal external cost of congestion calculated when the VTT differs among road users? Consider a situation where the population of potential travellers differ in their VTT. Groups of travellers are indexed by their value of travel time α. The group with VTT α has a demand function D(p(α)| α), where p(α) = τ + αt is the generalised cost for travellers α. Treating the VTT α as a random variable, the aggregated demand function is $\overline{D} = E(D(\alpha))$, the average over all groups of travellers. The average VTT in the population is E(α). This is not the same as the average VTT of those who actually travel, which is the weighted average $\overline{\alpha} = E[\alpha \cdot D(\alpha)]/\overline{D}$. Denote travel time as a function of demand by $t = S(\overline{D})$ and the change in travel time resulting from the marginal traveller by $S'(\overline{D})$ Multiplying this by the number of travellers and by the average VTT among travellers indicates the marginal external cost of congestion: $mecc = S'(\overline{D}) \cdot \overline{D} \cdot \overline{\alpha}$.

Comparing this to the canonical model, where there is not travel time heterogeneity, reveals that the difference is in the VTT used to compute the marginal external cost of congestion. There is just a single VTT in the case of homogenous travellers.

With heterogeneous travellers, this single VTT is replaced by the average VTT in the group of actual travellers (which may be difficult to establish empirically, see Section 4.1). Hence the mecc depends in general on the toll, not only – as is the case in the canonical model – through the slope of the supply curve $S'\left(\overline{D}\right),$, but also because the introduction of a toll will change the composition of travellers and this affects the average VTT that co-determines the marginal external congestion cost.

This effect is potentially large. Consider a case where the VTT in the population follows a standard lognormal distribution. Then the mean VTT is E(α) ≈ 1.65. Imagine now a toll that causes a reduction in traffic of 10% and that it is those travellers with VTT above the 10th percentile that remain, i.e. the 10% lowest value of time users are "tolled off the road". The average VTT of road users is then about t $\overline{\alpha}$ ≈ 1.81, which is an increase of about 10% compared to the no-toll situation. Not correcting the marginal congestion cost for this change implies a substantial error.

A toll will discourage some from travelling. If travellers with low VTT are discouraged more, as might be expected and is assumed in the example just given, then the average VTT $\overline{\alpha}$ increases with the introduction of a toll. Then the optimal toll, which is equal to marginal external congestion cost, will be larger when travellers are heterogeneous than when they are homogenous.[10] We are not aware of empirical evidence concerning the likely size of this effect, but it is reasonable to suppose that it is relevant, since the distribution of VTT is generally thought to have a shape similar to that presented in Figure 4.1 with many travellers having a relatively low VTT.

The first insight is that heterogeneity suggests higher optimal tolls. Second, if all travellers have the same value of time, the optimal toll on two parallel roads or lanes of the same capacity is the same. This is no longer true when VTTs differ. To see this, consider still the situation with heterogeneous travellers and imagine that the optimal toll $\tau = mecc = S'\left(\overline{D}\right)\cdot \overline{D}\cdot \overline{\alpha}$ is in operation. Imagine then that road capacity is split in two halves, A and B, with the same toll levied on both parts, and that travellers have to choose which part of capacity they want to use. Then they will divide equally on the two parts of capacity, as this means that generalized costs are the same on both halves so there can be no gains from switching.

[10] Arnott, de Palma & Lindsey (1994), discuss pricing with heterogeneous travellers in the context of the bottleneck model. Arnott & Kraus (1998) discuss marginal cost pricing when travellers are heterogeneous but the differences are not observed.

Now, if the toll is increased slightly on one part of capacity, say part A, then demand will decrease slightly there and shift to part B. Then part A will be faster but more expensive than part B. This will cause rational travellers to sort, such that those with the VTT above some threshold will use part A and those with the VTT below the threshold will use part B. As a consequence, the average VTT is higher on part A and lower on part B. Then, given the properties of the optimal toll under heterogeneity discussed before, the toll can be raised on part A and reduced on part B to produce a net welfare gain. This says that toll differentiation is part of the first-best solution in which there are no constraints on what tolls can be set. It is best to provide travel options that are differentiated in terms of toll – travel time combinations in order to cater as well as possible to the differences in preferences in the population for such combinations.[11]

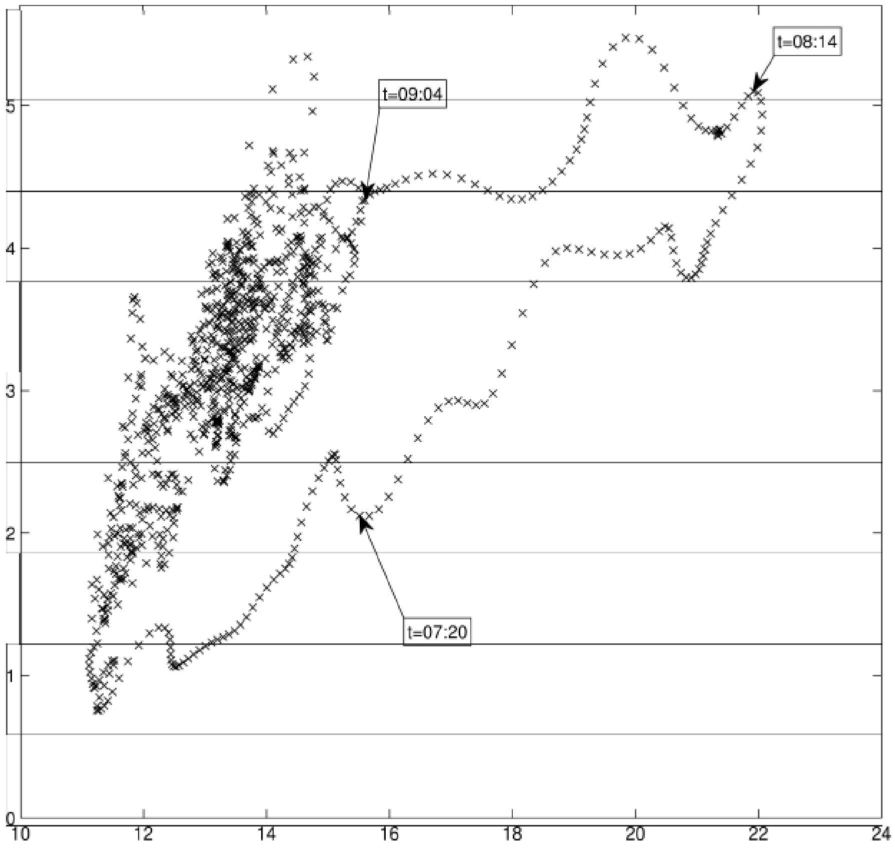
Verhoef & Small (2004) and Small and Yan (2001) consider differentiated tolls in a static network with serial and parallel links and with heterogeneous users . They are particularly concerned with second-best policies whereby only a part of the network is tolled. Value-pricing, as implemented in various places in the US, is an example of such a mechanism. Under value pricing drivers can choose between a faster toll lane and a slower untolled lane. It turns out that these policies are in danger of losing much of their potential effectiveness if heterogeneity is ignored when setting toll levels, i.e. when tolls are set as if all users had the same VTT. Furthermore, ignoring heterogeneity in VTT may cause the welfare benefits of second-best policies to be drastically underestimated, so that policies may erroneously be abandoned. Note also that the importance of recognizing VTT differences implies that speed changes as such are a poor indicator of the benefits of road charging.

[11] This could be regarded as a case of product differentiation (Mas-Colell et al., 1995), since the outcome is that the parts of the road deliver different travel times. In contrast to most goods, the service quality of roads depends strongly on usage.

5. TRAVEL TIME RISK

Increasing travel demand leads to congestion and increasing travel times. But as demand approaches capacity, travel times also become increasingly variable and unpredictable for users. This travel time variability (TTV) may add significantly to the generalised travel cost. How should we account for this when thinking about congestion tolls?

Figure 5.1. **Scatter plot of the standard deviation of travel time (vertical axis, minutes) against the mean travel time (horizontal axis, minutes) for a congested urban road**



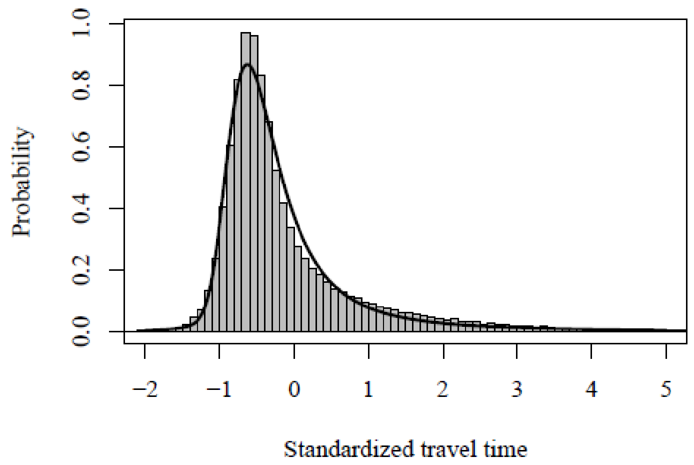
It is tempting to incorporate TTV into the GTC in a simple way by assuming TTV to be proportional to the delay caused by congestion. Then TTV, however defined, could be incorporated by assigning a larger VTT to delay. However, as the evidence in Figure 3 shows, this may not be appropriate. Figure 3 shows a scatter plot of the standard deviation of travel time against the mean travel time for a congested urban road with a distinct morning peak. Each point on the plot corresponds to a time of day, its position being determined by the mean travel time (horizontal axis) and the standard deviation of travel time (vertical axis) at that point in time. The numbers were computed using data covering a period of three months. Both the mean and the standard deviation are small in the early morning: at 7.20am the mean travel time is about 15 minutes and the standard deviation about 2 minutes.

Then the mean and standard deviation increase. At 8.14am a trip takes about 22 minutes on average with a standard deviation of just under 5 minutes. Then both measures decrease again. At 9.04am the mean travel time is about 15 minutes (same as at 7.20am) and the standard deviation is 4.5 minutes (more than double of that at 7.20am). The standard deviation peaks later than the mean, indicating that there is not a proportional relation between the mean and the standard deviation. This creates the loop that is evident on the figure. It is a characteristic pattern that has been observed many times[12], and it implies that the mean and the standard deviation of travel time ideally need to be accounted for separately in a measure of generalized travel cost.

Various measures of TTV have been employed, such as the standard deviation or the variance of travel time or a range between two quantiles (Small et al., 2005). Studies have then proceeded to estimate a value of variability based on revealed or stated preference data. This approach is not completely satisfactory without some arguments to indicate why one measure of TTV should be preferred to another. The problem is complicated since a travel time distribution is a shape rather than a number. Figure 4 shows an example of a travel time distribution. There is an infinite number of possible shapes and they cannot be described completely by a few numbers. We discuss some efforts to capture the essence of variability next.

[12] This pattern is generated by the random capacity bottleneck model for any distribution of capacity (Fosgerau, 2010).

Figure 5.2. **An empirical travel time distribution**



Intuitively, the cost associated with TTV is related to scheduling considerations. Compare two situations, one in which travel time is variable and one in which it is constant. The mean travel time is the same in both situations. Travellers have to decide when to embark on a certain trip. When travel time is constant, travellers choose an optimal time of departure which is directly associated with an optimal time of arrival at the destination; this is the logic of the simple bottleneck model discussed in Section 3. However, when faced with TTV, travellers may embark on the trip earlier than they would have under deterministic travel times. They may build in "buffer time". On average they therefore arrive earlier and sometimes they arrive later than they would have chosen with constant travel time.

To make this more formal, economic theory generally assumes that travellers have preferences that encompass scheduling considerations regarding when they depart from the origin of the trip and when they arrive at the destination. Travellers are pictured as knowing the travel time distribution and choosing the departure time optimally. A specification of scheduling preferences then leads to a definition of the GTC as the cost associated with making a trip that is optimally timed. The optimal GTC then depends only on the travel time distribution. This relationship is not in general tractable and there is generally no obvious candidate for defining a measure of TTV. There are however a few special cases, where simplifying assumptions enable a simple measure of TTV to be defined.

Fosgerau & Karlstrom (2010) consider the departure time choice of a traveller facing TTV. The traveller cares about travel time and about being early or late at the destination according to the Vickrey/Small scheduling preferences described in Section 3. The distribution of random travel time is taken to be independent of the departure time, such that a change in departure time does not affect the shape of the travel time distribution but only shifts it to earlier or later. Similarly, the monetary trip cost does not depend on the departure time.

When the traveller knows the travel time distribution and chooses the optimal departure time, it turns out that the expected GTC becomes linear in the mean and the scale of the travel time distribution[13], regardless of what the travel time distribution is.[14] More specifically, when the traveller chooses the optimal departure time, then $GTC = c + \alpha \cdot E(T) + \eta \cdot \sigma \cdot H$. In this expression, c is the monetary cost of the trip, and $\alpha \cdot E(T)$ is the VTT multiplied by the mean travel time. The last term captures the effect of TTV: η is the value associated with TTV and depends on scheduling parameters β and γ; σ is a measure of the scale of the travel time distribution; and H depends on scheduling parameters and on the shape of the travel time distribution.

At this point it can be noted that all measures of scale are essentially equivalent when the shape of the travel time distribution is constant. In this case, the standard deviation is proportional to the range between any two given quantiles. A change from one scale measure to another is then reflected in an inverse change in the value of TTV: no change results from multiplying σ by some positive number and dividing η by the same number.

Fosgerau & Karlstrom provide an example of a congested urban road in Copenhagen where the cost of TTV varies between 7% and 20% of the time cost to travellers with an average of about 15%. Including TTV in the GTC is likely also to lead to an increased estimate of the mecc. This is because TTV tends to increase with demand just as does the mean travel time

The bottom line is that there is a basis for including a measure of the scale of the distribution of travel time as a measure of TTV. Given estimates of the scheduling parameters and the shape of the travel time distribution it is possible to calculate the contribution of TTV to the GTC. It is not necessary to know the preferred arrival time of travellers since this does not appear in the GTC when the departure time is optimally chosen.

While the Fosgerau-Karlstrom result has some advantages for application, there are also some drawbacks. First, the value of TTV depends on the shape of the travel time distribution. It may therefore vary across different contexts. It still remains to gather sufficient empirical evidence to be able to judge whether this is a serious drawback in practice, given that there are many other uncertainties and approximations in play. Second, the scale of the travel time distribution can be hard to compute in networks comprising many links. This is not an issue with the mean travel time, since the mean travel time may be computed at the link level and then summed over links to obtain a trip level mean travel time. The standard deviation is not additive in this way and so the GTC cannot just be computed at the link level and summed.

[13] The scale or statistical dispersion of a distribution indicates how “spread out” it is.

[14] Previously, this had been shown for some special cases (Noland and Small, 1995; Bates et al., 2001).

In a broader perspective, it is important that the specification of scheduling preferences is consistent with empirical evidence. The Vickrey/Small specification of scheduling preferences entails the prediction for an individual traveller that an isolated increase in travel time will cause a proportional change in departure time that leaves the arrival time unchanged. An isolated increase in the standard deviation of travel time would lead to earlier departure and earlier arrival on average. This may or may not be an adequate description of actual behaviour.

There is an alternative formulation of scheduling preferences that also leads to a tractable expression for the value of TTV. It is based on a less known paper by Vickrey (1973) in which he defines scheduling preferences in terms of time-varying utility rates at the origin and at the destination of the trip. The traveller receives utility at some rate specific to the trip origin until he departs. When he arrives he begins receiving utility at a rate specific to the destination. The cost of the trip is an opportunity cost associated with the foregone utility at the origin or at the destination. When the utility rate at the origin is decreasing and the utility rate at the destination is increasing, then there is a time at which the individual would optimally travel from the origin to the destination. This view of scheduling preferences is attractive, since it treats the origin and the destination of the trip symmetrically. In general, it is hard to argue why timing at one trip-end should be more important than at the other as implied by Vickrey/Small scheduling preferences.

Using a simplified version of such scheduling preferences, Fosgerau and Engelson (2011) find an expression for the value of travel time variability that does not depend on the shape of the travel time distribution. The related measure of travel time variability is the variance of travel time. Depending on parameters, travellers may be risk averse or risk seeking and the value of travel time may increase or decrease in the mean travel time. This model has some advantages over the Fosgerau & Karlstrom model. First, the value of TTV does not depend on the shape of the travel time distribution. Second, the variance of travel time is additive over links in a network, provided travel times on links are independent.[15] Ultimately, of course, the choice between formulations of scheduling preferences and the associated measures and value of travel time variability should not be based on convenience but on conformity with observable behaviour.

Randomness is a lack of information. And so information provision is a natural policy measure in the context of TTV. Consider a situation in which travel time is variable from day to day but perfect information about tomorrow's conditions is provided to travellers. Then every day they can choose the optimal departure time and the GTC in the Fosgerau-Karlstrom model reduces to $GTC = c + \alpha E(T),$ which omits the term relating to TTV. The information does not have to be perfect in order to reduce the GTC.

[15] Travel times are not likely to be independent since delays on different links may have common causes. Still, additivity must be considered an improvement over no additivity.

In general, it can be just some signal that contains some information about tomorrow's travel time, i.e. it must have some relation with tomorrow's travel time.[16] This reduces the risk that travellers face. The value of this information may be assessed with the same models that are used to assess the cost associated with random travel time variability.

In using these results, it is important to keep in mind that no consideration has been given to equilibrium. The departure time choice of a single traveller is considered, taking the choices of all other travellers as given. The random distribution of travel time affects each traveller's choice of departure time. But there is also a causal relationship in the other direction whereby the combined departure time choices of travellers affect the distribution of travel time (Arnott et al., 1999).

Summing up, TTV matters and there are indications of how to include it in a measure of GTC that is used as a basis for determining a road toll. And just like changes in travel choices affect mean travel time, they affect its distribution. Lastly, when TTV is important, it follows that speed is not a sufficient indicator of the costs of congestion (this is in addition to its drawbacks when there is strong VTT heterogeneity).

[16] There are cases where equilibrium effects imply that imperfect information is not necessarily welfare improving (Arnott et al., 1999).

6. MEASURING AND MODELLING SUPPLY

The basis for efficient congestion pricing is the marginal external cost of congestion. It involves, essentially, the VTT and the supply relationship. So it is clearly crucial for the design of congestion pricing schemes to have an adequate understanding of the supply-side. The description of supply relationships for road travel has traditionally been considered the domain of engineering or physics and economics has tended to ignore the complexities involved, as is clear from the simple supply models embodied in the canonical model and the bottleneck model. This may have been reasonable in times when the main issues were to do with the design and capacity of road networks. But when it comes to the design of road pricing schemes or other ways of improving the efficiency of network use, it is necessary to have a deeper understanding of the supply side. In particular, it is essential to be able to estimate the effect on travel times of changing demand. Perhaps economics should get involved in this. Small & Verhoef (2007) discuss congestion from this point of view.

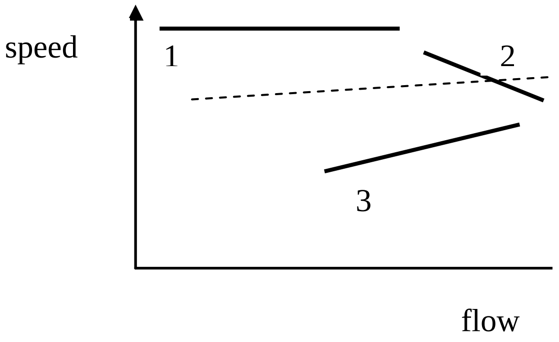
In economic models, supply is generally taken as given in the form of a supply curve for a road or a simple network. In the case of the bottleneck model, it is simply the bottleneck capacity, which is a single number. In reality, congestion is a hugely complicated phenomenon. Research into the relationship between travel demand and travel time has been ongoing for at least 75 years and we shall make no pretence of being able to summarise the state of the field. It is however useful to point out some of the issues involved. There are still very many open questions that appear when we ask about the consequences for mean travel time and variability of adding an extra traveller.

There are many causes and corresponding types of congestion. Flow congestion arises as traffic slows down due to increased density, independently of upstream or downstream links. Flow congestion is related to the microdynamics of traffic. It may arise due to small random fluctuations in flow and may be involved in the phenomenon of hypercongestion (see below). The importance of flow congestion is however debated with some arguing that congestion is more likely to be related to bottlenecks such as intersections and merge lanes and to accidents that create temporary bottlenecks. Congestion involves spill-backs such that delays on a link may be due to upstream delays. A particularly clear instance of spill-backs is when a queue behind a bottleneck blocks upstream intersections for crossing traffic. Delay for the crossing traffic is then unrelated to the demand for it.

Consider now that the objective often is to price urban networks comprising thousands of links and nodes and even more combinations of origins and destinations. It is clearly a daunting task to try to describe such systems in detail. Here rescue may come from the existence of urban-scale volume-delay relationships (Geroliminis and Daganzo, 2008), that allow the complexities of the network to be compressed into a single expression.

It is not in general sufficient to consider only the mean travel time, when travel time variability accounts for a significant share of GTC. It is necessary to be able also to predict the impact on TTV of changes in demand. As discussed in the previous section, there is no simple general relationship whereby TTV can be expressed as a function of mean travel time. The loop in Figure 5 also indicates the dynamic nature of congestion whereby even a small event at one time and place may have significant effects later and elsewhere in the system. Such dynamic phenomena create problems for the empirical measurement of speed-flow relationships.

Figure 6.1. **A stylised speed-flow relationship**



The fundamental diagram of traffic flow incorporates a relationship between speed and traffic flow on a road link, depicted as a backward bending curve as shown in Figure 5. It depicts a situation with three regimes. First, a free flow regime in which speed is about constant with flows ranging from zero to some point. Second, a congested regime in which higher flows are associated with lower speeds. Third, a hypercongested regime in which speeds are lower than in the other regimes at the same flow levels and where higher flows are associated with higher speeds.

Consider a scatter of observations on speed and flow, used in a regression of speed against flow. If there are many observations from the hypercongested regime, then it can happen that an increasing mean relationship seems to be present, as in the dotted line of Figure 5. This would imply the perverse prediction that increasing flow would lead to an increase in speed. The problem is that there is causality in two directions. The causal effect of interest is the effect of flow on speed. There is however also a causal effect in the opposite direction whereby low speed creates a blockage which causes flow to be low. The problem of reverse causality is a classical econometric problem and a range of econometric techniques exist to tackle it. In the present context it is important to realise that the measurement and modelling of supply should be taken seriously and requires sophisticated methodology.

Taking the supply side seriously sharpens insights regarding congestion tolls but also helps consider alternative policies to make more efficient use of road capacity. We give just two examples. First, a recent paper (Cassidy et al., 2010) shows that there are unexpected benefits from car pool lanes that do not have to do with pricing. The benefits arise because the carpool lanes reduce disruptive vehicle lane changing. Even a severely underused carpool lane can increase a freeway's total discharge flow. Second, in a dynamic setting there can be benefits from differentiation, even with homogenous travellers and without pricing. Fosgerau (2011) uses the bottleneck model to analyse such differentiation. The immediate cause of congestion in the bottleneck model is that travellers initially depart at a rate that is higher than capacity. Congestion is reduced by a toll that makes travellers decrease the initial departure rate. This effect may be induced in other ways. One way is to divide travellers into, say, two groups. A more than proportional share of capacity is allocated to the first group.

The second group can use the remaining capacity. It can also use the share of capacity allocated to the first group when the first group is not using it. The first group, having more capacity available per traveller, would find a new equilibrium where departures occur during a shorter interval of time and would hence experience a cost reduction relative to the situation without grouping. The cost of the second group would not increase since it is determined by the length of the interval during which departures take place and this is unchanged relative to the situation without grouping. The overall result is that costs are lower to serve the same total demand.

7. SECOND-BEST ISSUES

The key idea of congestion pricing is that it reduces waste by alleviating the misalignment of private and social costs of travel that is caused by the congestion externality. The previous sections have pointed out how hard it is to establish the magnitude of this misalignment, making abstraction of the wider context in which pricing might be introduced and assuming that sophisticated instruments for charging are available. Second-best analysis asks the broad question of how the basic analysis is modified when these simplifications are abandoned. The literature on the subject is vast and we make no attempt at providing an overview, instead referring the interested reader to a concise discussion in, for example, Small and Verhoef (2007). We limit ourselves to discussing some examples of second-best reasoning, and try to draw conclusions on how second-best analysis can help improve the practical implementation of congestion charging systems. Section 7.1 investigates the consequences of the fact that practical systems are approximations to the ideal charging system, and section 7.2 asks what are the consequences of the fact that congestion charging – even if potentially ideal – is implemented in an economy characterized by other inefficiencies than just the congestion externality. Section 7.3 infers some guidelines for practical analysis.

7.1. Imperfect implementation

An ideal congestion charging system charges the marginal external congestion cost[17] at each time and place in the road network.[18] A glance at existing and planned systems shows this ideal is not reached. More complex charging systems are more costly and complex price structures are hard to communicate to travellers. So there are good reasons for actual systems being less complex than the theoretical ideal. System costs apart, simpler systems yield lower welfare gains than can be attained with an ideal system. Cases can even be envisaged where less than ideal charging leads to a welfare loss. It is therefore of interest to investigate how systems should optimally be designed when there are practical constraints on tolls.

The problem of which links to charge and what tolls to set when only some links in a network can be charged has a conceptual solution (for a static network) that is difficult to translate into a practicable one (e.g. Verhoef, 2002).

[17] Or whatever turns out to be the optimal charge.

[18] Perhaps with allowance for imperfections in the wider economy.

Simulations using detailed network models suggest that reasonably performing pricing schemes can be designed even with a small number of tolled links or cordons, e.g. by choosing high volume and high-speed links with poor substitutes (Safirova et al., 2004). While this is in line with common sense, no links may fit the bill perfectly, so that choices can be hard in practice. Furthermore, the question of how much to charge remains unresolved. Also, when the choice is where to place a cordon instead of what link to charge, deciding where to place one or several cordons and what charge to levy appears to be particularly challenging, with the results from simulation work sometimes differing from common sense judgment (Sumalee et al., 2005). Systematic search algorithms are helpful for making sound decisions about where to charge and how much.

It seems reasonable to conclude that imperfect implementation is unavoidable but that nevertheless good results can be obtained. Systematic assessments of where and how much to charge can improve considerably on common sense judgment or at least help avoid big mistakes. Detailed analysis using traffic models is likely to have considerable payoffs. Also, as discussed earlier in the paper, the relative performance of second-best charges tends to be underestimated when heterogeneity in travellers' value of time is ignored. The intuition is that second-best charges often involve a degree of price differentiation over the road network, e.g. when only a few links in a congested network are subject to a charge, so that users can choose between alternatives representing trade-offs between tolls and congestion. Such differentiation is appealing when values of time differ between users, as a variety of options will match heterogeneous users better than when a single type of service is offered to all. This is true even when the differentiation differs from what would be first-best but, as long as it is in line with second-best guidance.

7.2 Interactions with tax distortions and distributional concerns

- Constraints on what types of congestion charges are feasible are only one source of second-best. The first-best analysis, as well as the type of second-best discussed in Section 7.1, focuses on the congestion externality in transport. It implicitly assumes that there are no distortions (deviations from efficiency) in the rest of the economy, or at least none that should be taken into account when thinking about charging for the external cost of congestion, so that – if possible – tolls ought to equal marginal external congestion costs (Pigouvian tolls). In reality, the economy is rife with distortions that potentially do matter, so that there may be optimal deviations from the Pigouvian toll.
- The transportation economics literature that has studied the relevance of interactions with tax distortions and of distributional concerns for setting congestion charges is embedded in theory on the optimal provision of public goods and on optimal taxation in economies where distortionary taxes must be used and equity is relevant to social welfare. More specifically, by far most contributions rely on "the standard approach" to modelling such economies. The insights from this standard approach are challenged by those from "the new approach", which relies on the benefit principle. Here, we summarize the intuition of both approaches, relying on the work of Kreiner and Verdelin (2012), and discuss what this means for the characterization of optimal congestion charges.

- The basic justification for a congestion charge is that it improves economic efficiency, i.e. it increases the total economic surplus generated in the economy. This is a good thing in itself, but there are two potential problems. First, apart from efficiency (the total surplus), the distribution of the surplus is of interest as well. Should distributional considerations be taken into account when setting the congestion charge? If some people are better off and some are worse off, is a higher total surplus still necessarily a good thing? And second, when there are other inefficiencies in the economy, it is no longer straightforward that removing or reducing one of them increases the total surplus, as the policy intervention might exacerbate these pre-existing distortions. More generally, interactions with other distortions could affect the optimal (second-best) level of the charge, so that it could differ strongly from the first-best level. Should pre-existing tax distortions be taken into account when setting congestion charges?

- The “standard approach” to answering these questions concludes that pre-existing distortions and distributional concerns indeed do matter. In general, it finds that the social cost of public goods is higher when distortionary taxes are used to fund them, so that less of the public good should be provided compared to the first-best case of distortion-free funding. This is the logic of assigning a marginal cost of public funds to arrive at an accurate estimate of the social opportunity cost of providing public goods in cost-benefit analysis.

- The “new approach” to analysing the optimal provision of public goods points out that the result of the standard approach is driven by essentially arbitrary restrictions on what kind of reform of the income tax system is associated with the change in the provision of the public good. These restrictions often are the consequence of assuming that income taxes are linear, and have the effect of introducing distributional considerations into the decision of how much of the public good to provide. If, however, it is assumed that the income tax system is sufficiently flexible that it can be adapted to make sure that each individual ends up paying an amount equal to their benefit from the increase in the public good (this is the benefit principle), then decisions on how much of a public good to provide are independent of distributional considerations as long as an individual’s ability to earn income and home produce is independent of its ability to derive benefits from the public good (at given income).

- Applying the “new approach” to the principle of congestion charging is simple. In our framework the charge is the instrument used to provide the public good of less congested roads. If individuals’ benefits from less congestion do not relate to their income-earning and home production capacity (at given income), then the optimal congestion charge is the Pigouvian charge, since the new approach makes clear that distributional concerns and other tax distortions are irrelevant to how much reduction in congestion should be provided.

- Most of the literature on second-best congestion charges up to now has adopted the “standard approach”, assuming that the compensating income tax change does not exist. Specifically, in line with much of the literature on second-best taxation, it has assumed linear income tax schedules, and has not considered the possibility of a compensating change to the income tax schedule. The consequence is that in these models pre-existing tax distortions and equity concerns affect the second-best congestion charge (see e.g. the cost-benefit rules in Calthrop et al., 2010), often in a drastic way, echoing results from the broader environmental charges literature. If the assumptions on the availability of compensating income tax schedule changes in the new approach (the benefit principle) are superior, the relevance of the results from the standard approach is limited. And that the assumptions of the new approach in general are more natural seems clear, if only because the ones of the standard approach are arbitrary and perhaps more driven by concerns on analytical tractability than accuracy. The "new approach" says that restrictions to income tax changes can affect optimal congestion charges, but this would depend on specific circumstances and the general recommendation is to set Pigouvian charges.

- We note that the argumentation of the "new approach" requires the existence of a compensating income tax schedule, but not necessarily its implementation. This is similar to the idea of a potential Pareto-improvement that underlies cost-benefit analysis (the Kaldor-Hicks criterion): a project is deemed desirable when it increases total economic surplus, so that those who stand to gain from the project could in principle compensate those that incur losses, even if that compensation is not paid in practice.[19] While at first sight it seems strange to rely on a potential that may never be realised, the opposite (requiring that compensations are carried out for each policy reform) is equally strange and ultimately futile as it is the impact of a multitude of ongoing policy changes that matter from a distributional point of view, not the impact of a single project. In that sense, it may be better to let an isolated project or policy pass if it has at least the potential to improve everyone's surplus. And it certainly makes no sense to go through with a project that does not pass the test, unless there is a strong argument that the special distributional objectives served by it could not be produced otherwise and are distributionally appealing. Such projects can exist as long as the income tax is not sufficiently flexible to account for all characteristics relevant to individuals' welfare generating capacity, a possibility recognized in the "new approach".
- We have up to now discussed how insights from the "new approach" to characterizing the optimal provision of public goods apply to congestion charges, concluding that they broadly favour adoption of the Pigouvian principle. But two related issues need further clarification. First, whereas public goods usually require spending public funds, congestion charges raise revenue. What to do with that revenue? The general answer is to use them to keep the distortionary cost of taxation as low as possible, which may mean reducing marginal labour taxes. Second, might this involve not charging tolls for commuters, a result suggested by the "standard approach" in which not charging tolls to commuters and higher tolls to non-commuters can outperform a uniform congestion charge (where in both cases net toll revenues are used to reduce the marginal labor tax in a linear tax schedule); see e.g. Van Dender (2003)?

[19] In contrast to the Kaplow approach, the Kaldor-Hicks argument does not require a social welfare function. But as no social welfare function would reject a Pareto-improvement, the views are consistent.

- The desirability of differentiating taxes between commuting and non-commuting trips in the standard approach results because this differentiation helps shift the tax burden from market activities (work) to non-marketed activities (household production), which reduces distortionary costs (Corlett and Hague, 1953; Munk, 2006). Kreiner and Verdelin (2012) show that this argument holds when preferences do not differ among individuals and only labour – leisure choices are considered, which is quite restrictive. In the more general "new approach", there is no clear general justification for differentiating charges according to trip types, either directly or indirectly. Otherwise said, using part of the revenue from a Pigouvian charge to reduce labor taxes specifically for car commuters is not a particularly appealing policy option.

- This section has focussed on the interaction of labour tax distortions and congestion charges, but of course there are several other important market imperfections that may matter. We mention just two, and relegate detailed discussions to future work. First, there are interactions between congestion charges and search unemployment. The latter occurs because it takes time for separated workers to match with new employment and the duration of which likely increases when transport costs rise (see Pilegaard and Fosgerau, 2008; Zhu et al., 2009). Second, agglomeration effects, i.e. external economies of scale due to spatial proximity of producers, may affect congestion charges as well. If all workers contribute equally to congestion and agglomeration, this suggests congestion tolls should be reduced by the value of the agglomeration externality unless a separate instrument is available to stimulate agglomeration. If workers differ in how they contribute to agglomeration but have equal impacts on congestion, this could be a reason to differentiate congestion charges (Graham and Van Dender, 2009). While much remains to be done on the understanding of interactions between congestion and agglomeration, it is increasingly clear that congestion and congestion management policies should not be considered in isolation from the productivity of urban economies, although it does not follow that strong deviations from external cost charging are in order.

7.3 Implications for implementation

One response to the possibility that several potentially important market imperfections interact with the congestion externality is to construct a model that encompasses the main imperfections (as judged by the model builder) and derive a rule for the assessment of charges from it. For example, Calthrop et al. (2010) propose a rule for transport infrastructure investments (which could be modified for transport pricing) that includes some of the interactions discussed above (within the "standard approach").[20]

Their framework emphasizes the role of distortions, and this comes at the cost of a strongly simplified representation of the transport markets. Fosgerau and Pilegaard (2007) take the opposite route, showing how some general equilibrium interactions related to tax

[20] We criticized the same paper in the previous section for its reliance on income tax restrictions that strongly affect results. That issue is separate from the consideration of the general modelling strategy considered here. Whether the complex rule is correct or not is obviously important, but here the question is whether the complexity itself is worthwhile.

distortions can be integrated into cost-benefit analyses based on traffic models. This has the advantage of allowing a detailed model of the transport market (relying on traffic models that are often used in the practice of transport project appraisal), but the range of general equilibrium interactions is more narrow.

The use of sophisticated rules is sometimes thought to be superior to the simple first-best rule on the grounds that the latter is shown to imply large mistakes in some cases, because the underlying model is not a very good approximation to real economic conditions. But even if the model underlying the complex rule is reasonable, it remains that given our imperfect understanding of the broader context of transport policy reform in a conceptual sense and even more in an empirical sense [21], it is not obvious that the approximation of the sophisticated rule is necessarily better.

Furthermore, even the most comprehensive models on general equilibrium interactions are (by definition) highly stylized representations of reality. They highlight distortions thought to be of particular importance (with judgment ideally based on evidence) while ignoring others, and they miss features that matter in the applied analysis of proposals for congestion charging. For example, few models contain a sufficiently detailed representation of the capacity and usage of multimodal transport systems that would allow a comparison of policies that use revenue to reduce labour taxes or to improve the supply of public transport. Clearly, such comparison would be relevant to the design of charging proposals, both from the point of view of economic optimality and political feasibility. Parry and Small (2009) present a detailed analysis of rationales for subsidizing public transport. They do not focus on the interaction with broader distortions, but do suggest that their impact on the optimal subsidy is limited.

Our view, then, is that the models generate insights that ought to be part of debates on the implementation of congestion charging, with a clear understanding of their limits and what drives the results. Applications of simple models can help clarify the importance of second-best concerns ("model-assisted reasoning"[22]) and suggest rules of thumb, but we should not expect a full-fledged general equilibrium analysis to be carried out that can directly prescribe the details of specific reforms. Instead, we think that concrete policy guidance on the design of congestion tolls is best served by the deployment of state-of-the-art traffic models that account for the major network usage issues identified in this paper: heterogeneity of travel times, travel time reliability, and bottleneck congestion.

This is a tall order in itself, but with considerable potential payoffs. General equilibrium concerns can be accounted for in ad-hoc ways, and the emerging insight is that in important cases they are less important than previously thought.

[21] The evolving insight on the relevance of other distortions, discussed in Section 7.2, and the conceptual and empirical uncertainty on the size and nature of agglomeration economies can serve as examples.

[22] Richard Arnott (1998) distinguishes between model-based and model-assisted reasoning, where the latter refers to the use of models to illuminate a broader argument, and the former is where the model is the argument.

8. CONCLUSION

Unregulated congestion entails an efficiency loss and a corresponding possibility for obtaining a welfare gain. This gain can be achieved through road pricing, decentralising the decision about who should travel when and where. In the first step of analysis, the toll should equal the value of the delay that a marginal car imposes on other travellers. The size of the total delay associated with the marginal car is determined from traffic models, ranging from simple supply curves to complex traffic models. The transformation to monetary value accomplished through the value of travel time (VTT), which can be measured in various ways.

Congestion arises because people tend to travel at the same times. With an even distribution of traffic over time, there would be no congestion. As the example of the bottleneck model shows, there are potentially large benefits that can be achieved if the trip timing aspect of demand is taken into account by varying tolls over peaks, inducing travellers to distribute departure times more evenly.

The simplifications involved in the textbook analysis of congestion pricing allow the central insights to be easily communicated. There are however a number of complications that must be taken into account when this theory is taken to practice. First, people are different. The stylised facts state that many people have low VTT but some have very large VTT. The mean VTT is larger than the median. There is much variation between people but also between seemingly identical individuals and even within the same individual in different contexts. Recognition of this heterogeneity will tend to lead to higher suggested tolls and will tend to reveal larger benefits from price differentiation between roads.

Travel times are inherently random. As congestion increases, travel times become not only longer but also increasingly variable and unpredictable. This travel time variability contributes significantly to travel costs. Taking travel time variability into account will generally lead to higher suggested tolls. Theory exists whereby the value of travel time variability can be expressed in terms of the standard deviation or the variance of travel time in a simple and readily applicable way.

Traffic systems are hugely complicated and the complexities of measuring and modelling supply should not be underestimated. It is difficult to establish the size of the delay associated with a marginal vehicle and even more difficult to establish the consequences for travel time variability. The simple descriptions of the supply side often employed in economic papers are not adequate for real world assessment of road pricing systems. More realistic and tractable representations are available, however, that may be adequate for guiding practice.

Constraints on charging instruments lead practical systems to fall short of ideal (in a first-best or second-best sense) congestion charging, and this implies lower benefits than would be obtained in that ideal. Charging can of course still produce net benefits (after subtraction of investment and operational costs). Careful assessment of charging systems becomes crucial however, because of the multitude of design options available and the large differences among them in terms of benefits produced. Such assessment should tackle the second-best aspects explicitly, and not evaluate a system as if it were first-best. Changes in travel speeds are not a sufficient indicator of the benefits of charging.

Congestion charging should not be considered in isolation, as there are economic interactions that have large potential effects on the benefits of charging. As indicated, we are sceptical about the possibility of capturing these interactions in one elaborate model, and indeed about the methodological validity of doing so. In summary, it seems reasonable to use the basic analysis of congestion charging as a first approximation when considering its implementation, but to check whether important interactions with other market imperfections can be expected and revise the analysis when there are concerns about large indirect effects. Good practical preparation of congestion charging mechanisms is first and foremost a matter of using the best available traffic models.

This paper has reviewed some important considerations for the determination of congestion charges. In particular, there are important implications from congestion dynamics and the endogeneity of trip timing, from the heterogeneity of travellers and from the presence of travel time variability. It is then also highly desirable that traffic models are able to handle these dimensions. Models that incorporate these aspects do exist, although they are still rare and need to be developed further.[23]

[23] Engelson et al (2012) compare the models METROPOLIS (de Palma et al., 1997) and SILVESTER (Kristoffersson and Engelson, 2009) in an ex post study of the Stockholm congestion charge. METROPOLIS handles dynamics, trip timing and heterogeneity, while SILVESTER also goes some way to take travel time variability into account. Both models provide significant improvement in realismover static models.

REFERENCES

Arnott, R., 1998. William Vickrey: Contributions to Public Policy. International Tax and Public Finance 5, 93–113.

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

Arnott, R.A., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. American Economic Review 83, 161–179.

Arnott, R.A., de Palma, A., Lindsey, R., 1994. The Welfare Effects of Congestion Tolls with Heterogeneous Commuters. Journal of Transport Economics and Policy 28, 139–161.

Arnott, R.A., de Palma, A., Lindsey, R., 1999. Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand. European Economic Review 43, 525–548.

Bates, J., Polak, J., Jones, P., Cook, A., 2001. The valuation of reliability for personal travel. Transportation Research Part E 37, 191–229.

Calthrop, E., De Borger, B., Proost, S., 2010. Cost-Benefit analysis of transport investments in distorted economies. Transportation Research Part B 44, 850–869.

Cassidy, M.J., Jang, K., Daganzo, C.F., 2010. The smoothing effect of carpool lanes on freeway bottlenecks. Transportation Research Part A 44, 65–75.

Corlett, W.J., Hague, D.C., 1953. Complementarity and the Excess Burden of Taxation. The Review of Economic Studies 21, 21–30.

De Borger, B., Fosgerau, M., 2008. The trade-off between money and time: a test of the theory of reference-dependent preferences. Journal of Urban Economics 64, 101–115.

De Borger, B., Proost, S., 2001. Reforming Transport Pricing in the European Union - A Modelling Approach, Transport Economics, Management and Policy Series. Edward Elgar, Cheltenham, UK.

de Palma, A., Fosgerau, M., 2011. Dynamic Traffic Modeling, in: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), A Handbook of Transport Economics. Edward Elgar.

de Palma, A., Marchal, F., Nesterov, Y., 1997. METROPOLIS - Modular System for Dynamic Traffic Simulation. Transportation Research Record 1607, 178–184.

Engelson, L., Kristoffersson, I., de Palma, A., Motamedi, K., Saifuzzaman, M., 2012. Comparison of two dynamic transportation models: The case of Stockholm congestion charging. Presented at the 4th TRB conference on Innovations in Travel Modeling, Tampa, Florida.

Fosgerau, M., 2006. Investigating the distribution of the value of travel time savings. Transportation Research Part B: Methodological 40, 688–707.

Fosgerau, M., 2010. On the relation between the mean and variance of delay in dynamic queues with random capacity and demand. Journal of Economic Dynamics and Control 34, 598–603.

Fosgerau, M., 2011. How a fast lane may replace a congestion toll. Transportation Research Part B 45, 845–851.

Fosgerau, M., Engelson, L., 2011. The value of travel time variance. Transportation Research Part B: Methodological 45, 1–8.

Fosgerau, M., Karlstrom, A., 2010. The value of reliability. Transportation Research Part B 44, 38–49.

Fosgerau, M., Pilegaard, N., 2007. Cost-benefit rules for transport projects when labor supply is endogenous and taxes are distortionary. Munich Personal RePEc Archive 3902.

Fosgerau, M., Van Dender, K., 2010. Road pricing with complications. OECD/ITF Joint Research Centre Discussion Papers 2010-2.

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation Research Part B: Methodological 42, 759–770.

Graham, D., Van Dender, K., 2009. Pricing congestion with heterogeneous agglomeration externalities and workers. mimeo.

Kreiner, C.T., Verdelin, N., 2012. Optimal Provision of Public Goods: A Synthesis*. The Scandinavian Journal of Economics 114, 384–408.

Kristoffersson, I., Engelson, L., 2009. A Dynamic Transportation Model for the Stockholm Area: Implementation Issues Regarding Departure Time Choice and OD-pair Reduction. Networks and Spatial Economics 9, 551–573.

Mas-Colell, A., Whinston, M., Green, J., 1995. Microeconomic Theory. Oxford Press, Oxford.

Menendez, M., Daganzo, C.F., 2007. Effects of HOV lanes on freeway bottlenecks. Transportation Research Part B 41, 809–822.

Munk, K.J., 2006. Rules of normalisation and their importance for interpretation of systems of optimal taxation (No. Working Paper 2006-13). School of Economics and Management, University of Aarhus.

Noland, R.B., Small, K.A., 1995. Travel-Time Uncertainty, Departure Time Choice, and the Cost of Morning Commutes. Transportation Research Record 1493, 150–158.

OECD, 2010. Implementing Congestion Charging: Summary and Conclusions. OECD/ITF Joint Transport Research Centre Discussion Papers.

Parry, I., Small, K.A., 2009. Should urban transit subsidies be reduced? American Economic Review 99, 700–724.

Parry, I.W.H., Small, K.A., 2005. Does Britain or the United States Have the Right Gasoline Tax? The American Economic Review 95, 1276–1289.

Pilegaard, N., Fosgerau, M., 2008. Cost-benefit analysis of a transport improvement in the case of search unemployment. Journal of Transport Economics and Policy 42, 23–42.

Raux, C., 2005. Comments on "The London congestion charge: a tentative economic appraisal" (). Transport Policy 12, 368–371.

Safirova, E., Gillingham, K., Parry, I., Nelson, P., Harrington, W., Mason, D., 2004. Welfare and distributional effects of road pricing schemes for metropolitan Washington DC, in: Road Pricing: Theory and Evidence, Research in Transportation Economics. Elsevier JAI, Amsterdam, pp. 179–208.

Small, K., 1982. The scheduling of Consumer Activities: Work Trips. American Economic Review 72, 467–479.

Small, K.A., Van Dender, K., 2007. Long run trends in transport demand, fuel price elasticities and implications of the oil outlook for transport policy. ITF Discussion Paper 16.

Small, K.A., Verhoef, E.T., 2007. Urban transportation economics. Routledge, London and New York.

Small, K.A., Winston, C., Yan, J., 2005. Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability. Econometrica 73, 1367–1382.

Small, K.A., Yan, J., 2001. The Value of Value Pricing" of Roads: Second-Best Pricing and Product Differentiation. Journal of Urban Economics 49, 310–336.

Sumalee, A., May, T., Shepherd, S., 2005. Comparison of judgmental and optimal road pricing cordons. Transport Policy 12, 384–390.

Van Dender, K., 2003. Transport taxes with multiple trip purposes. Scandinavian Journal of Economics 105, 295–310.

Verhoef, E.T., 2002. Second-best congestion pricing in general networks: heuristic algorithms for finding second-best optimal toll levels and toll points. Transportation Research Part B 36, 707–729.

Verhoef, E.T., Small, K.A., 2004. Product differentiation on roads: constrained congestion pricing with heterogeneous users. Journal of Transport Economics and Policy 38, 127–156.

Vickrey, W.S., 1969. Congestion theory and transport investment. American Economic Review 59, 251–261.

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

Zhu, X., Van Ommeren, J., Rietveld, P., 2009. Indirect benefits of infrastructure improvement in the case of an imperfect labor market. Transportation Research Part B 43, 57–72.